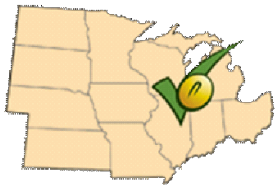
Project report- McHale

March 30, 2018



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

The SOYGEN group had two project meetings this Spring, the first was at the Soybean Breeders' Workshop on February 13, 2018. During this meeting we discussed updates and action points for each main project objective. Specific attention was paid to potential modifications for FY19. It was suggested that additional, high impact phenotyping of the PIs grown as part of objective 3 would be extremely valuable and maximize the use of the multi-location field trials that are unlikely to be grown again outside of this project. Present in St. Louis for this meeting: Aaron Lorenz, George Graef, Danny Singh, Katy Martin Rainey, Ed Anderson, Brian Diers, Dechun Wang, Pengyin Chen, David Hyten, Steven Clough. Present on the conference line: Leah McHale, Matthew Hudson, Jianxin Ma, Andrew Scaboo, Bill Beavis.

The second meeting was web-based Zoom meeting on March 30, 2018. The primary objective for this meeting was to ensure that plans were in place and understood for the collaborative field trials (Objective 1 and 3). Important points of discussion included the addition of in-season selections made by Rainey's group in 2018 in order to ease harvest of the ~5000 progeny rows and reduce the number of lines actually needed for harvest. In addition, future potential changes to the project for FY19 were discussed, these included additional phenotyping for Objective 3 (R1, R3, R5 dates, V4 image data, canopy temperatures, and weather data). Participating breeders will attempt to gather this data in FY18 as well, though it was not budgeted for this granting period. For FY19, we also proposed to collect functional genotype data (maturity genes and branching genes) for the germplasm panel. Work with tomentella-derived lines has been re-focused to the identification of introgression of tomentella DNA into soybean. Attending zoom meeting: Leah McHale, Bill Beavis, Steve Clough, George Graef, Matthew Hudson, Jianxin Ma, Aaron Lorenz, Katy Martin Rainey, Andrew Scaboo, Bill Schapaugh, Dechun Wang.

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.

Breeders submitted the following phenotype data from approximately 40,000 progeny rows grown in 2017, which is summarized below. Breeders Selections refers to the binary keep/discard information provided for each lines by the individual breeders using their current methods. The Rainey Lab quantified canopy cover from aerial and ground imagery using two photogrammetric workflows and shared these data with the breeders.

Breeder	# of lines	Breeder Selections	Yield	R8	Canopy Imagery	Lodging	Maturity Group	R1-R8
Diers	2800	✓	✓	✓	✓	✓	-	-
Lorenz	7200	✓	-	-	✓	-	-	-
Schapaugh	6111	✓	✓	-	✓	-	✓	-
Rainey	5231	✓	✓	✓	✓	-	-	-
Graef	3240	✓	-	✓	-	-	-	✓
Wang	3120	✓	✓	-	-	-	-	-

Chen	4253	✓	-	-	-	✓	✓	-
Singh	2335	✓	✓	-	-	-	✓	-
Scaboo	954	-	✓	-	-	-	✓	-
McHale	4746	✓	✓	✓	-	-	-	-

Due to a variety of issues, the following programs had difficulties fully implementing the protocols in 2017:

- Chen, Diers, and Lorenz were not able to save all lines as specified.
- Chen did not collect additional data as specified.
- Scaboo had very poor germination and emergence and harvested all of just 954 lines, and did not collect and additional phenotype as specified.
- Cianzio's health problems have prevented data submission.

Task 2: Selections from progeny rows.

A framework for submission of data and pedigrees for statistical analyses was developed by the post-doc, Meng Huang, and summary statistics were calculated. Multiple statistical genetics softwares were tested using cross validation to identify the most stable workflows, and selection models were developed and evaluated. Breeder-submitted phenotypes were adjusted using row and column planting information in a spatial analysis to remove micro-environmental variation. Pedigree information was summarized to assess extent of pedigree sharing, shown in the table below. After consultation with breeders at the Soybean Breeders' Workshop in St. Louis in February 2018, and given consideration of pedigree overlap, multiple selection models were tailored to individual programs, and included variations in spatial and pedigree adjustments of phenotypes, and use of pedigree within breeding program vs. across breeding programs. For each program, two to three selection models were used to select the top 8% of lines within each model. In combination with random selections as a control and breeder selections, 8% of lines across 4-5 categories were tabulated. The top ranked lines within selection categories are summarized along with reports by breeders.

Meng Huang intends to develop an R package for selection of soybean progeny rows with canopy coverage, pedigree, and spatial adjustments; outputs could support a database in SoyBase.

	Diers	Lorenz	Schapaugh	Wang	Rainey	Singh	Scaboo	Graef	Chen	McHale
Diers	2800 (39)	386	330	266	47	225	492	493	-	-
Lorenz	6	7200 (212)	-	-	-	-	-	14	-	-
Schapaugh	2	-	6111 (50)	-	-	-	1937	-	445	340
Wang	2	-	-	3120 (49)	330	306	-	-	-	116
Rainey	2	-	-	3	5231 (74)	968	977	-	-	340

Singh	2	-	-	2	9	2335 (29)	897	373	-	100
Scaboo	6	-	10	-	11	12	954 (112)	-	-	34
Graef	4	1	-	-	-	1	-	3240 (12)	-	-
Chen	-	-	1	-	-	-	-	-	4253 (28)	-
Mchale	-	-	2	3	2	1	5	-	-	4746 (57)

*The diagonals show the number of lines considered for each breeder, and in parentheses the total number of parents used for each breeder. Above the diagonal is the number of liens with at least one parent in common between breeders. Below the diagonal is the number of parents that overlap between the breeders.

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

As a group, the breeders have discussed how to organize the 2018 trials, but tests have not been finalized as of 04.02.2018.

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

- Additional data are collected on progeny rows.
  - This KPI was met for about 35,000 lines.
- A list of all breeders' lines ranked simultaneously for yield breeding value, maturity prediction and a metric of diversity.
  - This has not yet been fulfilled because we are still working with model selection and organization of 2018 trials.
- Cooperative preliminary yield trials are organized to test selection accuracy.
  - To be implemented in the 2018 season.

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.

Data collection and analysis for this objective is complete and manuscript preparation is still underway as mentioned in the last report.

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.

We are continuing to work on developing a publicly available relational database to house this data and make it easily downloadable and searchable (see http://cfans-lore0149-prd-web-01.oit.umn.edu/). Such a database will facilitate the use of this pedigree and phenotypic data for development of genomic prediction models. The historical phenotypic and pedigree data from 1989 to 2016 has been compiled in text files and cleaned. The data from the 2017 season was just received and will be processed shortly. Entry information for 7895 lines from 1989 to 2016 has been complied and

cleaned. Pedigree data of 2925 lines has been parsed and cleaned, with the remaining in progress. Data from years 1989-2015 has been uploaded as a starting point, and will be uploading 2016 and 2017 data in the near future.

For the 1989 to 2015 data, we have a total of 8,798 strains in this dataset from 3,280 entries with parents. Line names and pedigrees were curated to 1) identify and eliminate typos and duplicate names created by typos; 2) generate an alias database of lines with duplicate names; 3) generate machine readable and traceable pedigree structures for complex crosses. Metadata and trait ontology has been refined and has been uploaded.

Regarding the database development, we have used the T3 database platform and have completed the migration of this database to a University of Minnesota College of Food, Agriculture and Natural Resource Sciences server. We have also discussed this with Soybase.org curator David Grant, and he is enthusiastic about pointing Soybase.org to our new database.

In order to make our pedigree files more robust, we have begun mining pedigree data from the USDA-ARS Germplasm Resource Information Network (GRIN) database. Our goal is to try to better connect the breeding lines from different breeding programs to each other and all lines back to the original North American founder lines. From this database mining work, we now have 12,048 strains in the pedigree database and 5,700 strains with parents.

Last year we uploaded the phenotype dataset to Benson Hill Biosystems database, and we have provided only the University of Minnesota genotype data set to Benson Hill Biosystems. Benson Hill has a unique system where they take available genotype-phenotype data and use it to help enable predictive plant breeding in the private sector. We feel this is one direct way in which the data we generate from this project can have impact on soybean improvement.

Next steps for this part of the project include: 1) Uploading 2016 phenotype data to the database, 2) cleaning and uploading 2017 data to the database, 3) uploading marker data to the database 4) uploading pedigree data to the database.

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

Currently 1,336 lines have 6K genotype data. An additional 324 lines are about to be submitted for genotyping. Some historic lines no longer have seed available or the seed is longer viable. We believe we have collected all available lines, and we have started genotyping the new entries to the 2018 Preliminary Regional Trial tests.

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

We have completed our work on GBS and have developed approaches using the R programming language and environment that allow us to cross-analyze GBS data and array data.

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

We have been working with David Hyten to select a reduced panel of SNPS that capture as large a percentage of the genotypic diversity as possible. Our initial approach was to consider polymorphic information content of the markers in combination with code to optimize the equidistant spacing of th markers across the genome. However, we have opted for a haplotype based method (with haplotypes identified by the software Haploview) that considers the recombination frequencies across the

genome rather than physical distance across the genome. So this improved approach tries to optimize the identification of haplotype blocks across the genome with respect to the natural recombination rates.

Since the 288 probes worked well and are individually synthesized we tested whether increasing or decreasing their individual concentrations in the MIPs protocol would influence the number of reads produced by each probe. If a probe produced too many reads than we diluted it and if it did not produce enough reads we increased its concentration. We term this "balancing the probe set" since we are trying to balance the number of reads a probe produces by putting different concentration of probes in the final MIP reaction. We tested a balanced probe set based on our initial sequencing results and compared it to an unbalanced probe set where all the probes are in equal concentration. Balancing probes did tend to reduce the amount of reads coming from high performers and increase the number of reads coming from the low performers. While balancing further could produce a better read coverage for each SNP, we will wait until we have the final set of 1000 probes before performing an additional probe balancing experiment.

We have performed a series of enzyme titrations to try to reduce the units of key enzymes in the MIPs protocol. These experiments determined that the protocol is very amendable to reducing the amount of enzyme used in the extension and ligation reaction when using our specific good DNA extraction method. Our mapped read percentage is between 69%-86% even with using the lowest amounts of titrated ligase and DNA polymerase.

We have continued to test different DNA extraction methods. Out of five different methods, we have obtained alignment percentages ranging from 11% up to 42%. The method that produced 42% alignment is an increase over previous dirty DNA methods. One issue is that we also tested some samples from "clean" DNA extraction methods and they produced alignments of 8% up to only 54%. This was surprising since all of our previous result with the clean DNA method we had been using has been consistently over 69% alignment and most of the time they are 80-86%. This suggests that the MIPs protocol is being inhibited by something that is present in some of these other clean DNA extraction methods.

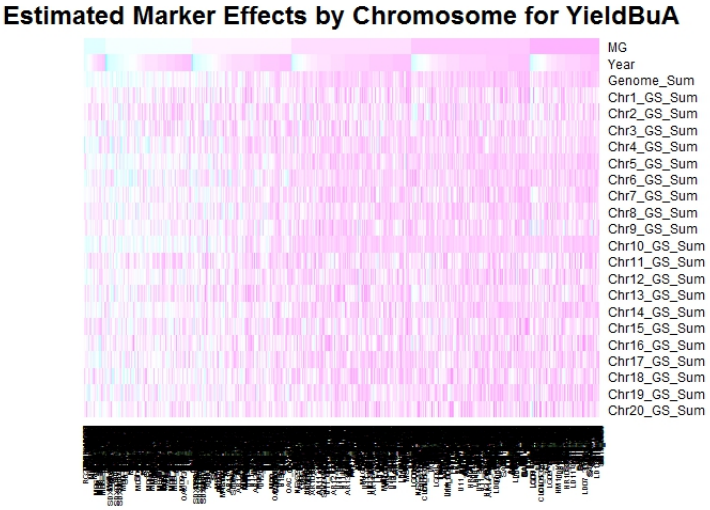
In the next quarter, we plan to adapt our protocol from 96 well plates to 384-well plates and begin automating our protocol. This will be essential for gaining enough throughput to handle thousands of samples per year. A major focus during this next quarter will be trying to make the protocol more robust to DNA input. We will perform a series of dilutions along with testing different PCR enhancers to try to make the protocol more robust to different DNA inputs, which includes clean and dirty DNA extraction methods. We will also order the rest of the probes to complete the 1,000 SNP set for genomic selection.

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

Scripts have been written to calculate best linear unbiased estimates of URT lines by factoring in trial and maturity group effects. Overlapping checks were leveraged to combine data across trials. These estimates have been combined with the 6K SNP data collected to date and algorithms to perform and validate genomic prediction have been designed and implemented. A ten-fold cross validation has been performed both within maturity groups and across maturity groups. The traits yield, maturity, seed protein, seed oil, plant height, lodging, and seed size were assessed.

Genomic predictive ability based on a preliminary analysis was estimated to be 0.72, which is quite high. We will continue to review these models to ensure that all confounding effects of maturity have been removed. We are in the process of designing a leave-one-trial out cross validation which we feel will be the most indicative in terms of estimating genomic prediction accuracy.

Another thing we have used the 6K SNP on the URT lines for is characterizing the "germplasm architecture" in the URTs. We estimated marker effects using all data on all lines, and calculated the "genomic breeding values" of each chromosome of each URT line (see figure below). Preliminary results suggest that there are many lines that could complement each other in terms of chromosomes with different genomic breeding values, resulting in breeding populations with high amounts of variance for yield. Also, there appears to lines of early maturity groups carrying chromosomes with high predicted breeding value for yield. These could be isolated and crossed to attempt to combine all high yielding chromosomes. More work needs to be performed to ensure the optimal model has been used.



2.f. 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.
    - The current GBS method does 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 – see Table 4 and Figure 3C). The 10% missing data fraction is a function of the population size. For all studied populations, this metric was achieved by the modified TASSEL parameters used in the cited study and accuracy measured by whole-genome sequencing. We also developed a new software approach that greatly outperforms TASSEL if 1) a greater missing data fraction is allowed and / or 2) our recommended association population experimental design is followed. The new 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

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

Task 1: Exploration of a retrospective analysis

A two-part survey has been distributed to determine the ability and willingness of public soybean breeders to provide this information. Part one of the survey was administered to gather the level of interest in participation from 19 breeders participating in the URT and requested identification of genotyped URT lines (see objective 2) which have been used in bi-parental crosses. Part two of the survey aimed to collect data on the relative success of each cross combination. Relative success is estimated by the ratio of number of advanced lines produced from a cross to the number of progeny rows initially grown from an individual cross combination.

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

The 2017 year was used for increase of the 250 lines selected from the quintiles of the distribution of the 9,000+ untested accessions from the USDA germplasm collection. Test locations for 2018 have been finalized among collaborators. Seed has been prepared and will be shipped in the coming days.

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

With the GBS data generated from 416 lines derived from 23 different crosses involving XX elite soybean varieties and XX G. *soja* accessions, we profiled genome-wide distribution of *G. soja*-introgressed segments in each line using 10-kb continuous windows along each chromosome. In general, progeny lines derived from a same cross share more introgressed *G. soja* fragments than those from different crosses, but exceptions were also observed. For example, of the 111 lines from the Williams 82 × PI 479752 cross, 61 and 35 lines were clustered into two major clusters established based on shared introgressed segments, but the rest were shattered into four clusters with lines from four different crosses. This pattern was also observed in lines from other crosses. These observations suggest that the selection for desirable phenotypes may have resulted in biased retention of *G. soja* segments, although which phenotypes may be associated with such biased retention remains to be elucidated. On the other hand, all of the 416 lines appears to have lost *G. soja* segments in the ~200 selective sweep regions as well as domestication QTLs that reflect the outcome of domestication for desirable agronomic traits in cultivated soybeans.

Instead of individual single nucleotide polymorphisms (SNPs) detected in the 416 lines and their parents, *G. max*-*G soja* segment polymorphisms (Gm-GsSPs) were used as molecular markers to determine genomic regions from either the *G. max* or *G. soja* parents associated with yield performance of these progeny lines by association analysis. A total of nine regions/segments (located on chromosomes 2, 7, 8, 11, 13, 15, and 19) associated with high yield were identified. Of these regions, two located on chromosomes 8 and 11 from *G. soja* appears to have positive contributions to high yield. In this summer, a set of lines with multiple high yield-related segments will be crossed to combine all these segments into a same lines assisted with the Gm-GsSP markers. We will further test the effects of the two *G. soja*-segments on yield potential in newly developed progeny lines with and without these segments.

With the support from this NCSRP project, we have fine mapped several important QTLs that affect soybean yield performance, such as leave shapes and branching angels using two large RIL populations derived from crosses between Williams 82 and PI 468916 or PI 479752, and a large F2:3

population derived from a cross between LD00-3309 and a (Williams 82 and PI 468916) RIL. The leave shape QTL was mapped to <200-kb region and the branching angle QTL was mapped to a ~20kb region harboring only three genes according the soybean reference genome. Further comparison of the gene sequence and expression between parental lines have already pinpointed the candidate gene (tentatively designated *GmBa1*) underlying narrow branching angle (compact plant architecture). Although these QTLs were identified using mapping populations derived from crosses involving *G. max* and *G. soja*, these traits vary among cultivated soybean varieties including elite cultivars. Thus the results from this task can facilitate efforts defined in Objectives 1 and 2 in this project. For example, we found that *GmBa1* was located in the QTL region associated with canopy coverage that was identified using SoyNAM population (Xavier et al., 2017) . Further, the functional *GmBa1* mutation appear to be able to distinguish the SoyNAM parental lines showing distinct canopy coverage, suggesting that *GmBa1* may have pleiotropic effects on both branching angle and canopy coverage, or in nature, canopy coverage is largely determined by branching angle. Given that canopy coverage is one of the most important traits to investigate in Objectives 1 and 2 as described in this project, the functional markers for branching angle would be very useful for validation of the prediction models for selection. As such, we plan to further validate the functional role of *GmBa1* in modulating canopy coverage and design functional markers for branching angle/canopy coverage and soybean maturity to facilitate genomic selection in the 3rd year of the this NCSRP project.

In addition to the above work with QTL analysis, we have now succeeded in identifying haplotypes associated with yield derived from the alternative and conventional ancestors that are both wild and domesticated soybean:

Methods:

Using the SoySNP50k data available online from SoyBase, plus the data previously generated in this project, we derived a phased and imputed SNP data set using our newly genotyped lines from Dr. Randall Nelson's alternative gene pool breeding program (Alternative elite (Ae), Alternative intermediate (Ai), and Alternative ancestors (Aa)) and the conventional lines genotyped in the public SoySNP50K project (Conventional elite (Ce), Conventional intermediate (actually old variety releases) (Ci), and Conventional ancestors (Ca)).

The entire data set was processed using the Haploview algorithm (maintained by PLINK) to define haplotype blocks, groups of LD-linked SNPs consistently inherited together.

For each of these blocks we conducted a Tajima's D test in each individual group (Ae, Ai, Aa, Ce, Ci, Ca) and identified haplotype blocks that are under selection in the elite lines (Ae, Ce) but not the ancestral lines (Aa, Ca). These haplotype blocks are thus targets of selection in conventional and alternative breeding programs.

The full data set was also used the following analyses:

1) Discriminant analysis of principal components (DAPC) was used to identify those linear discriminant components capturing the variance associated with selection in the elite lines (Ce, Ae), and those SNPs that most strongly contributed to that variance (using eigen deconvolution of the loadings).
2) A per-population sliding window scan of the calculated H1 (Garud et al., 2015) metric for each haplotype, which isolated ranges of SNPs with lower haplotype diversity in the elite lines compared to the ancestral lines, likely targets of artificial selection.
3) A Fixation Index (Fst) analysis comparing elite to ancestral lines to identify any SNPs that may have become fixed by the artificial selection effort.

These lines of evidence were joined by first placing candidate SNPs within respective haplotype blocks, then separately for the conventional and elite selection efforts isolating those haplotype blocks

that overlapped across lines of evidence. Once we had a complete set for each selection effort we assessed the overlap between them.

Findings:

Our analyses identified two linear discriminant dimensions likely correlated with yield across the two independent selection efforts. This finding implies that largely the same set of haplotypes are under selection in the two populations, allowing us to infer that these haplotypes are indeed under selection for yield and are not the result of drift during the selection process.

Our tests further identified 15 haplotype blocks across 7 Glycine max chromosomes shared by both selection lines. These block spans contain a total of 273 potential candidate genes for yield improvement.

In addition, Ae lines had 9 haplotype blocks across 3 chromosomes unique to that selection effort, while the Ce line had 42 haplotype blocks across 13 chromosomes. This result is indicative of variation and pedigree history, as Ce lines have experienced selection other than that associated with yield over a much longer period. While the haplotypes under selection in Ae could be the result of drift, they could also be haplotypes unique to the Alternative gene pool that could potentially increase the yield of conventional varieties through directed breeding approaches.

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

The two *G. tomentella* RIL populations were evaluated at nine locations in 2017. All RILs have been genotyped-by-sequencing. Phenotypic and genotypic data will be analyzed to find QTL related to yield.

The Federal hiring freeze continues to frustrate our efforts to develop new perennial *Glycine*-derived lines for Task 7. We selected a very qualified candidate to fill a new position to work on developing new techniques for wide hybridizations, but unfortunately, we could not finalize her appointment before the 2017 Federal hiring freeze was implemented and she has yet to join our group. However, we continue to test 2n=40 lines developed from previous crosses between Dwight and PI 441001 (*G. tomentella*) and crosses using these lines as parents. The attached file (tomentella summary) presents data collected in 2017. All of these lines have 20 pairs of chromosomes. The worksheet labeled “first generation lines” has data on recently developed lines. Many lines exceed the yield of the soybean parent Dwight and the best lines are competitive with the highest yielding checks. The worksheet labeled “second generation F3 lines” has data for F3 lines from crosses between two first generation lines and with other soybean lines. These are very preliminary data but provide some evidence that there are genetic differences affecting yield among these *tomentella*-derived lines. Most of the crosses that are producing the highest yielding lines have parents that were derived from different BC2 plants. These are lines that are most likely to be genetically different because each BC2 line should have initially contained different subsets of *G. tomentella* chromosomes. This is best illustrated by the data from the GB II test. In the GF III test there are high yielding lines from parents derived from either different or the same BC2 plant. Many of these lines are competitive with the best checks.

Additional evidence for genetic differences affecting yield is presented in the table at the bottom of the worksheet. In this table, we summarize the results of crosses using 10 first generation *tomentella*-derived lines. These parents were used in 2 to 6 crosses each. Some

crosses were between *tomentella*-derived lines and some were between *tomentella*-derived lines and conventional soybean lines. These crosses were classified based on whether they produced any lines that were not significantly different in yield from the best check. The first three parents had 0 to 20% of the crosses that produced lines that were not significantly different than the best check. The next three parents had 30% of the crosses that produced such lines. All of the crosses for the last four parents produced lines that were not significantly different from the best check. There is still much to learn about the genetics of these high yielding lines that we have produced from crossing PI 441001 with Dwight.

The third worksheet (second generation F6 lines) presents data on our most advanced second generation lines and some of these pedigrees combine both *G. soja* and *G. tomentella.* One such line was nearly 5 bushels higher yielding than the best check (SE III test). Other lines developed by crossing *G. tomentella*-derived lines with other soybean lines are competitive with the best checks.

In this project to increase the rate of genetic gain for yield, the research coming from Urbana has the only projects explicitly using exotic germplasm although perhaps some breeders may have lines derived from exotic germplasm in their 5000 plant rows. There is much evidence that exotic germplasm can be very beneficial. The SoyNAM project demonstrated that experimental lines developed from exotic germplasm have yield QTL that were not found in any of the conventional lines. In this report we have data on lines derived from crossing two wild soybean accessions to Williams 82 (yield genetics that are nearly 50 years old) that are equivalent in yield to lines currently used as checks in the uniform test. We have third generation lines that could be significantly higher yielding than the checks with more testing. We have *G. tomentella*-derived lines that are significantly higher yielding than the soybean parent and, like the *G. soja*-derived lines, could be significantly higher yielding than the checks with more testing.

Task 6: Development of breeding lines from perennial Glycine

As highlighted above, we have been seeing real benefits from incorporating exotic germplasm into our soybean breeding program, reinforcing the decision of NCSRP to invest in this work. However, the future of our ARS breeding program at Urbana is uncertain, as the past breeder, Randy Nelson, retired in Feb 2017, and due to the Federal hiring freeze, we have been unable to hire a replacement him. If we cannot replace Randy soon, or establish him as an official collaborator in the interim (something we are actively pursuing with ARS administrators), then there might be difficulty in fully completing the long-term goals of this grant. Other breeders who are interested in improving the rate of genetic gain by using wide crosses with *Glycine* perennials, might also wish to incorporate this approach to expand the genetic base for US soybean breeding as a part of their strategies.

3.f. Key Performance Indicators or performance measures met thus far (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).
- 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

Task 1. Recruit an appropriately skilled PhD graduate student.

The first graduate student, John Cameron graduated. While recruiting a replacement for John, a cohort of graduate students consisting of Vishnu Ramasubramanian, Danielle Dykema and Haley Trumpy have been filling in project gaps until the replacement arrives. A second graduate student, Matheus Dalsente Krause from Brazil, has been recruited and will join the group in May 2018.

Task 2. Communication and outreach.

Danielle Dykema and Haley Trumpy have been interviewing farmers about their understanding of the terms genetic gain. Many of these interviews have been video-taped. They also developed a video presentation describing soybean breeders understanding of genetic gain and how it related to the concepts expressed by soybean farmers. Last, they obtained video from NCSP to frame their interviews. Danielle and Haley are now negotiating with a group to provide technical expertise in cutting, splicing and override audio portions of the recordings.

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

Syngenta has signed a MOU granting us access to data from their variety development projects.

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

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.

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

We have communicated the data content and data formats we need to Syngenta Information Systems Management group, and as far as we know the data are being aggregated and formatted for our use by the time.