Project report, Final

February 5, 2018

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

The group had a productive meeting at the Soybean Breeders' Workshop in St. Louis in February. During this meeting, personnel changes were discussed as well as project updates and any concerns were presented to the group and are summarized in the report below.

In lieu of our planned project meeting at the World Soybean Research Conference in Savannah, we had a conference call on September 21, 2017. The call was productive and served to update project members on progress, remind those participating in multi-state field trials of expectations, and to vet details of experimental design.

The group's next meeting is planned for February 13, 2018 at the Soybean Breeders' Workshop in St. Louis.

Last year, the group had a brain-storming session over email in which it was decided to "brand" our project with a name that can envelope this group, but also serve as an umbrella that could potentially include future projects (hopefully from other funding sources, including federal grants). The name that was chosen was SOYGEN: Science-Optimized Yield Gains across ENvironments. A project website and a logo are under development.

Specific updates for each objective follow.

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.

General Update: Postdoctoral researcher Meng Huang started in October 2018, who is responsible for data management from 11 cooperators, and statistical genetic analyses.

Task 1: Collection of additional data in all progeny rows.
In 2017, each breeder phenotyped thousands of progeny row lines for new or additional traits, summarized in the list below. All new traits required significant effort. Analyses of all submitted phenotype data indicate that we have generally high-quality phenotypes for analyses. Ground-based canopy images from Brian Diers were converted to canopy coverage values by a Rainey lab student using a Matlab workflow used in previous research. George Graef could not harvest lines for yield as intended, so visual selections, R1-R8, pedigree data and spatial analyses were used to develop selection categories prior to harvest. Specifications for aerial imagery were communicated to breeders collecting aerial data (Lorenz and Schapaugh), and the aerial image analysis pipeline was modified to accommodate cooperator's data.

1. Chen: 4253; NDVI
2. Cianzio: pending
3. Diers:2800, yield and canopy cover
4. Graef : 3240, R1-R8
5. Lorenz: 7200, canopy cover

6. Mchale: 4746, yield
7. Rainey: 5231, canopy cover
8. Scaboo: 954, yield
9. Schapaugh: 6111, canopy cover
10. Singh: 2335, yield
11. Wang: 3120, yield

Task 2: Selections from progeny rows.
A framework for submission of data and pedigrees for statistical analyses was developed by the post-doc, 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. 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.

Task 3: Preliminary yield trials to evaluate the increase in the rate of genetic gains.
This task will be implemented in 2018. The approach was modified via group decision in the Fall so that preliminary yield trials testing selection categories will be organized within breeding programs, rather than coordinated across breeding programs. We will test thousands of lines per selection category, so this detail does not impact fulfilment of the objective.

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.

Yield data from four locations grown in 2016 was collected and compiled on all 1011 progenies chosen by the four breeding programs. This data was combined with the 2015 data on the same progenies. The quality of the data was good, with progeny-mean heritabilities of around 0.77 for seed yield. Originally this project involved the selection of lines by several different methods: phenotypic selection, genomic selection, random selection, and genomic selection on yield+protein. However, little difference between these methods due to premature genomic prediction model development led us to treat all lines from all methods as a single validation population. Now that we have two years of data for four locations, we have a superior validation set for testing genomic prediction using the SoyNAM population.

Progress was also made in refining the genotyping by sequencing of the validation progenies. An improved SNP calling pipeline developed by M. Hudson and B. Diers provided us with approximately 19,411 SNPs (<80% missing data, > 0.05 MAF). We are more confident in these SNP calls compared to the original SNP calls. 449 lines have been re-genotyped, with the remaining 562 in progress. The NAM parents have also been re-genotyped using GBS, and the GBS SNPs will be projected onto the NAM progenies using the 5K SNP data. This will allow us to connect the NAM progeny training data to the validation progenies using the improved SNP calling pipeline.

While this has been in process, we took the time to use the old SNP data to begin to explore genomic prediction accuracy using the NAM population. We compared selection accuracy between genomic prediction and plant row phenotypic selection using the 2015-16 validation data as a measure of success. We found the predictive ability of the NAM-based genomic predictions to be 0.40, while the predictive ability of the plant row phenotypes was only 0.095. This indicates that genomic prediction holds potential to make selections at the plant row stage.

All data analysis is complete. Preparation of the manuscript is underway. We hope to have a manuscript submitted by 2018. In summary, genomic prediction appeared to work very well within the NAM population, but a substantial amount of accuracy was lost when going from the NAM population to the new breeding populations. While this is a disappointing result, we feel it illuminates our approach to incorporating genomic selection into practical breeding programs and gives us direction on how to build the phenotype-genotype databases described below.

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.

Progress on this task has been slower than desired because the postdoc working on this project resigned and took a permanent job with Nature Source Genetics last October. Lorenz has recently found a replacement and who started working on this project in March. Nevertheless, we have made considerable progress in compiling and cleaning historical regional trial data. We are working on developing a publicly available relational database to house this data and make it easily downloadable and searchable (see https://plpa201657036.cfans.umn.edu/soyurt/). 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 going back to 1989 has been compiled in text files and pedigree cleaning is underway. Entry information for 3570 lines going back to 2004 has been uploaded to the database. Pedigree data of 2925 lines has been parsed and cleaned, with the remaining in progress. While the phenotypic data has been compiled into single files, we are still working on getting it formatted, cleaned, and appropriate metadata assigned and trait ontology developed so we can upload to the database. Data from years 2004-2005 has been uploaded as a starting point, and we are using this to discover many issues to address.

Dr. Ben Campbell has settled into his role as the new postdoc helping to manage this project since March 2017. Since then, he has leveraged data and ideas from this project to obtain a NSF Plant Genome Research Program Postdoctoral Fellowship (NSF Award #39207273).

Historical phenotypic data (yield, maturity, lodging, protein, oil, disease resistance, etc.) going back to 1989 has been completely cleaned and compiled and is ready for upload to a relational, publicly available database (described below). Pedigree data going back to 1989 has been cleaned and compiled and is ready for upload. We have a total of 8120 strains in this dataset and a thorough pedigree cleaning was a lengthy task. 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 is ready for upload.

Regarding the database development, we initiated a database using the T3 platform on a local machine at the time of the last report, and have since migrated that to a University of Minnesota College of Food, Agriculture and Natural Resource Sciences server. This will serve as a long-term home for the database as well as make it publicly available. The platform construction is underway and is expected to be completed shortly. Once the platform is constructed, we will immediately upload all URT data. We have discussed this with Soybase curator David Grant, and he is enthusiastic about pointing Soybase to our new database. Finally, we have hired a part-time data curator to help with the task of cleaning and curating data.

Another activity we have been involved with is working with BensonHill Biosystems in uploading our data into their database. We are one step away from uploading this data to them. BensonHill 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) Upload all available URT data to database, 2) Begin compiling and cleaning all other publicly available genotype-phenotype datasets in soybean.

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

There were 3784 lines tested in the URTs going back to 2004. We have successfully genotyped over 930 lines tested in the URTs with the 6K SNP chip. About 440 lines have had their DNA extracted and these are currently in process at the UMN Genomics Center. Some breeding programs will not allow us to genotype their lines, totaling 288 lines. 237 lines have not been found. 49 lines were proprietary company lines. We are still waiting on seeds, or are trying to locate 1716 lines. In summary, we can account for and genotype about 50% of the lines going back to 2004 currently. We expect this number to get better as we remind more breeders and more thoroughly seek out seeds. We requested the 2017 newly entered lines be sent to UMN for genotyping with the 6K SNP chip. We do have nearly complete data genotype data from 2014-2016.

DNA from an additional 376 lines were extracted and submitted for genotyping to the UMN Genomics Center, including all newly entered lines from the 2017 URTs. We inquired with breeders about additional seed available and we received an additional ~200 lines from previous trials. We have cleaned and organized this seed and are in the process of extracting DNA.

Next steps: 1) Continue to identify additional missing lines and genotype them (for example, we are traveling to SDSU next week to obtain seed from their defunct breeding program); 2) QC and upload new genotype data in relational database and link to phenotypic data.

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

This task has been in limbo with the exit of Henry Nguyen from this project; however, it is being picked up Matt Hudson's group. Thus far, Hudson's group has been focused on genotyping using sequence data from "genotyping-by-sequencing" (GBS). The group compared the performance of five GBS software pipelines using low-coverage Illumina sequence data (typical for breeding projects) from three soybean populations. To address issues identified with these existing methods, a new solution, GB-eaSy, was developed. GB-eaSy is a GBS bioinformatics workflow that incorporates widely used genomics tools, parallelization and automation to increase the accuracy and accessibility of GBS data analysis. Compared to other GBS pipelines, GB-eaSy rapidly and accurately identified the greatest number of SNPs, with SNP calls closely consistent with whole-genome sequencing of selected lines.

Tests showed that GB-eaSy is approximately as good as, or better than, other leading software solutions in the accuracy, yield and missing data fraction of variant calling, as tested on low-coverage genomic data from soybean. It also performs well relative to other solutions in terms of the run time and disk space required. In addition, GB-eaSy is built from open-source modular software packages that are regularly updated and commonly used, making it straightforward to install and maintain. A paper describing GB-easy was just published in Dec 2017: Wickland, D. *et al.*, BMC Bioinformatics 18:586.

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

Implemented a CTAB DNA extraction suitable for GBS sequencing. This protocol can be used to extract DNA from 1000-1500 samples a week at a cost of approximately $0.40 per sample. This protocol is now being used as our standard control as we test potentially cheaper DNA extraction methods.

Our targeted GBS protocol has been successful for enriching for the targeted sequence. A total of 3,397 probes were enriched from a 4k probe set. The enrichment of the targeted probes is currently between 10-17% of the total sequence. Our current goal is to have between 50-60% of the total sequence be enriched for the probe's targeted sequence. In the next quarter, we will be running experiments to test different parameters to increase the on-target sequence of the probe set. We have also ordered a new probe set with changes in the design which should increase the on-target percentage of the sequence data. This new probe set will be used to test the low cost, high throughput DNA extractions that we will be testing over the next quarter.

In the development of the genotyping methods, we have screened 14 different methods of DNA isolation that could have the potential for producing DNA suitable for sequencing. From these methods we selected the most promising one that met all the criteria of being cheap, automatable and would likely produce data with the GBS protocol. We compared the new DNA extraction method to a standard high quality DNA extraction method using a 4k-plex genotyping-by-sequencing method. For the high quality DNA extraction, our on-target sequence was 60-70%. This matches the expected on-target rate for this GBS protocol. The new method produced an on-target sequence rate of 30%. While this is lower than the expected 60% this is a good indication that this new DNA extraction method has potential for GBS. We will continue to work on optimizing this method and test methods that could help purify the DNA further to increase on-target sequence without adding a significant amount of cost.

We have also tested a 288 SNP probe set created from individual oligo synthesis. This is different from the 4k-plex which was created by a pool oligo synthesis method. The individual oligo synthesis provides the ability to adjust oligo concentration based on the efficiency of each oligo within a reaction. Since each oligo is individually synthesized we will also be able to create low-density GBS marker sets specific to germplasm that would need to be genotyped. The individual oligo synthesis method has an initial large up-front cost for oligo purchase but the amount of oligo received is enough to do 1 million reactions. With the initial cost spread out over the lifetime of using the oligo in GBS reactions, the individual synthesis method will not add a significant cost per reaction for genotyping. We tested the 288 probes on 96 soybean samples. From the 288 probes 278 successfully averaged over 10 reads per probe. On-target sequence was above 90% which is extremely good. There was only 4% missing data (<10 reads for a SNP datapoint) and genotyping accuracy was over 98% accurate for most samples. The individual synthesis probe method appears to produce better data than the probe pooled synthesis method.

OBJECTIVE 3: Increasing additive genetic variance

Task 1: Exploration of a retrospective analysis

A part time post-doc, Dr. Mao Huang, has recently been hired to determine to the feasibility of a retrospective analysis using existing public data from the URT (compiled as part of objective 2). The retrospective analysis is designed to predict the success of parental combinations and evaluate their success based on the relative performance of their progeny (total progeny tested/total progeny advanced).

We determined that the number of lines tested in the URT which were derived from parents which were tested in the URT and are part of the set of lines to be genotyped as part of objective 2 is small (~80) and insufficient for a valid statistical analysis. Thus, any retrospective analysis would require further data acquisition from breeders on the number of progeny tested from specific cross combinations. A two-part survey is being distributed to determine the ability and willingness of public soybean breeders to provide this information. Part one of the survey has been 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. Thus far, 9 of 19 breeders expressed an interest in participating (no response, yet, from 10) and four have supplied the requested information.

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

We have completed analysis of the 2016 field data for yield, maturity, height, and lodging recorded for 500 accessions from the USDA Soybean Germplasm Collection. This information has been added to the 2015 data, so we now have a set of 14 environments of data, 28 replications, on 500 PI accessions in MG I to IV from the germplasm collection.

We have completed the NIR analyses for protein, oil, and fiber for all plots grown in 2016. The seed weight and seed composition data from both years, 2015 and 2016, will be used for the final analyses.

We have compared models for predictive ability for yield for each year and over years. We are comparing models based on different training sets: models based on each sampling method, cluster (CLU), supersaturated design (SSD), and random (RAN), and the complete set of 500 lines. In addition, predictive models from each year and over years are being evaluated. We also are conduction genome wide association analyses of the data to potentially identify specific loci that show a significant effect on yield in these accessions.

We compared models with and without genotype data, and with and without GxE effects, to predict yield performance of soybean germplasm accessions. We performed predictions using each sampling group as a training set (SSD, CLU, and RAN) as well as using all the data for all 500 PI over all sampling groups as the training set. The SSD set performed as well as the complete set, and addition of GxE effects in the model did not improve prediction accuracy.

From the set of ~9,400 new accessions whose yield was predicted based on our 2015-2016 2-year yield evaluations in 14 environments and 28 reps, we selected 250 accessions to use in the validation set. The distribution of yield predictions was divided into quintiles, and we sampled 50 lines from each of the five quintiles of the distribution so we would end up with a sample of lines spanning the entire range of yield predictions. To select the 50 lines within each quintile, we used the supersaturated design analysis to identify 50 lines with maximum diversity within each group. Those 250 lines were then obtained from the collection and are currently growing in Nebraska for seed increase. We collected descriptive data on all lines, as well as maturity date and lodging and shattering scores. We just completed harvest of the MG1 accessions, and will complete harvest of all accessions by November 1. This is the seed source that will be prepared and distributed to cooperators for the 2018 and 2019 multi-location yield evaluations for these 250 PIs to obtain actual yield and agronomic data to validate the predictions.

Task 4: Identify signatures of selection in G. max derived lines selected for high yield

We have assessed population structure and signatures of selection in founder lines and high yielding elites of both conventional US cornbelt varieties and the alternative gene pool ancestors and elite lines developed in Randy Nelson's breeding program. When assessed by selection pedigree, the lines generally cluster with prominent overlap between ancestor lines suggesting that those genetic groupings are not completely genetically distinct, but share a pool of most common alleles.

Analysis of Fst agrees with the cluster analysis where ancestor and elite lines from each selection effort differ more from each other than ancestor lines differ between each other. A novel finding from the Fst analysis was that elite lines also do not seem to differ from each other across the genome except at specific markers (e.g., no selective peaks, but individual SNPs). Linkage-based assessments (Rsb) mirror Fst findings, and further suggest that elite lines from both selection efforts differ from respective ancestors at similar genetic regions. Another linkage-based analysis (H1) showed few regions under selection shared between elite lines, more regions unique to conventional elites while a few regions unique to alternative elites. We are now grouping the genome into haplotype blocks to identify specific haplotypes under selection for elite traits.

Task 5: Identify introgressed regions from wild soybean in regions of domestication genes and regions associated with high yield

Genotyping-by-sequencing (GBS), a method to identify genetic variants and quickly genotype samples, reduces genome complexity by using restriction enzymes to subset the genome into fragments whose ends are in both accuracy and number of SNPs identified sequenced on next-generation sequencing platforms. GBS uses a relatively simple protocol for library preparation and reduces costs by multiplexing samples. However, incomplete genomic data and complex bioinformatics analysis have hindered the widespread adoption of GBS in gene mapping studies. Moreover, in polyploids such as soybean with homeologous regions resulting from genome duplication, GBS SNP-calling tools may align reads to the wrong homeolog or fail to distinguish between-sample SNPs from within-sample SNPs. These "homeoSNPs" generate background noise that encumbers gene mapping. We have addressed these concerns by developing SBSBV, a streamlined GBS analysis pipeline that outperforms widely used pipelines, especially on soybean data. It is much faster, more automated, and uses less disk space than other methods. Compared to IGST and TASSEL, it is simpler, easier to use, and more accurate. Large GBS datasets totaling over 6000 lines have been analyzed in a single batch, which would not have been possible (in any reasonable timeframe) with other pipelines.

A set of 416 *G. soja* derived lines from 23 different crosses were characterized by GBS to produce 80,000 SNP markers. The Williams 82 x PI 479752 cross contributed the largest number of lines (111) and was used to first investigate the distribution of alleles from the *G. soja* parent. Overall, the lines contain a reduced proportion of SNPs derived from *G. soja* (25%) compared to a random biparental population (50%), which was caused by our selection for desirable phenotypes. When looking across chromosomes, regions can be identified with both substantially lower and higher frequencies of *G. soja* alleles. To examine the effect of selection on regions around domestication QTL, a shattering locus and 100-seed weight locus were selected. Both regions had low frequencies of *G. soja* alleles in the Williams 82 x PI 479752 lines. Stronger selection was seen against the *G. soja* shattering QTL due to its stronger effect on the trait compared to the 100-seed weight QTL. Eight lines were identified with *G. soja* segments adjacent to the shattering locus without displaying the shattering trait and averaging 47 bu/ac. Nearly 50 lines were identified with *G. soja* segments within the selective sweep at the 100-seed weight QTL, although all lines also had smaller seeds than Williams 82. Additional domestication QTL are being surveyed to identify lines with *G. soja* introgressions within selective sweeps. This approach is also being applied to the entire set of *G. soja* derived lines, which should allow us to improve are ability to detect introgressions within regions of low genetic diversity.

We identified lines derived from Williams 82 x PI 479752 that yielded 5% more than Williams 82 but that difference was not statistically significant. We are begining the analysis to compare high and low yielding lines within the same pedigree to determine if there are any consistent differences in the chromosomal regions introgressed from the various wild soybean parents.

We completed the identification of QTL controlling key domestication-related traits (DRT) using 151 RILs from Williams 82 x PI 468916 and 510 RILs from Williams 82 x PI 479752. We measured 11 phenotypic traits: flowering date, maturity date, main stem length, lodging, stem diameter, growth habit, leaflet length, leaflet shape, shattering, pubescence form and 100-seed weight. A total of 97 QTL where identified with 2 to 11 QTL per trait. The majority of the domestication related traits examined in this study were controlled by minor QTL, with QTL explaining over 50% of the variation only detected for flowering date, maturity date, shattering, and pubescence type. The 97 QTL were distributed across all 20 chromosomes within 36 genomic regions. These findings identified additional QTL not detected in previous studies using smaller populations while also confirming the quantitative nature for several of the important domestication related traits in soybeans.

Additionally, we have isolated a gene controlling soybean seed coat bloom and seed oil content. We have also fine-mapped a gene associated with leaf shape, which appears to be an important yield component.

The domestication QTL regions have also been compared with the previously identified over 200 large selective sweeps (i.e., genomic regions showing dramatic reduction in genetic diversity in G. max compared with overall genetic diversity along individual chromosomes). Some of these selective sweeps overlaps with domestication QTLs and some of the selective sweeps do not. Next, we will coordinate with the UIUC team to integrate these comparative genomics data with the yield data from the 416 lines to identify genomic regions positively or negatively associated with yields, and to understand how the domestication-related QTLs may affect yields, and which of the selective sweep regions may be related to yields. In an attempt to identify/validate genetic variation responsible high yields, the genomic regions associated yields as defined by analysis of the 416 *G. soja* derived lines will be examined in the 500 highly diverse PI accessions described in Objective 3 – Task 2.

We found three QTL that were related to both lodging and stem diameter, which would be expected; but there were three lodging QTL that were not related to either main stem length or stem diameter. These may be related to specific characteristics that affect stem strength.

Introgressions from PI 479752 were not randomly distributed across the genome, but were detected in regions of low and high frequency. Only a few regions contained *G. soja* alleles at frequencies near or above 50%, predominantly on chr 3, 4, 7, 8, 15, and 18. Areas of higher *G. soja* frequency most often occurred in regions without domestication-related QTL. This would indicate that these regions contain neutral or potentially positive alleles. With a few exceptions, all 32 regions containing DRT QTL were located at or adjacent to regions of low *G. soja* allele frequency. Regions with DRT QTL with relatively high frequency of *G. soja* alleles were QTL that had small effects.

We have begun the process of identifying experimental lines with *G. soja* introgressions close to the regions of DRT QTL in lines that do not have the *G. soja* phenotype for the DRT. These lines are likely to continue introgressions from *G. soja* that were lost during domestication. For example, qSH-16 is a major QTL that explains 53% variance in shattering. Only two lines contained *G. soja* alleles at this locus, and they were also the most shattering susceptible. A total of 20 lines contained *G. soja* introgressions within 500 kb of the qSH-16 locus, and all displayed low levels of shattering. Two lines with introgressions adjacent to qSH-16 yielded the same as Williams 82. These lines could be useful for increasing diversity within the selective sweep at qSH-16 without introducing undesirable traits.

Selection for 100-seed weight occurred only indirectly as agronomically good phenotypes would generally have larger seeds so recovering the 100-seed weight of the soybean parent was more difficult than reducing shattering. Lines from the Williams 82 x PI 479752 pedigree averaged 9 g/100 seeds, considerably less than the 16.1g 100-seed weight of Williams 82. Unlike shattering, 100-seed weight is controlled by many QTL with small to moderate effects. This caused weaker selection against negative *G. soja* alleles at the seed weight QTL. While qSW-17-1 accounted for the largest phenotypic variance in the Williams 82 x PI 479752 mapping population, the *G. soja* allele at that locus still persisted in about 25% of the lines. From Williams 82 x PI 479752, 31 lines had 100-seed weights of 10 g or more. Of these, eight contained *G. soja* introgressions within one or more of the major seed weight QTL, qSW-12, qSW-17-1, or qSW-19. Two lines carried *G. soja* alleles at two of the three seed weight QTL while still maintaining 100-seed weights of 10 g.

Task 6: Map yield QTL in G. tomentella*-derived lines*

We grew 225 *G. tomentella*-derived lines for each of two populations in replicated tests at one location in 2016. The parents of the population are Dwight x LG11-12313 (T Map II) and Dwight x LG11-3187 (T Map III). LG11-12313 was developed by backcrossing PI 441001 (*G. tomentella*) four times to Dwight and LG11-3187 was developed by backcrossing PI 441001 three times to Dwight. Both *G. tomentella*–derived parents yield significantly more than Dwight. One line was dropped from the T Map III populations for 2017 because of very low yield. Both tests will be grown at 9 locations in 2017. We are currently preparing the seeds and they will be shipped during the first week in April.

DNA has been sampled from each line and genotyping by sequencing has been completed. For marker analysis we will use SNPs that were identified from the sequences that align to the Williams 82 reference genome as well as sequence tags that were identified as from *G. tomentella* and not existing in either soybean or wild soybean.

This summer we grew two tests. A maturity group II test had 225 entries derived from Dwight x LG11-12313. LG11-12313 is from Dwight (5) x PI 441001 and in tests at 12 locations yielded 5.9 bu/a more than Dwight and was only one day later in maturity. A maturity group III test had 225 entries derived from Dwight x LG11-3187. LG11-3187 is from Dwight (4) x PI 441001 and in tests at 15 locations yielded 7.6 bu/a more than Dwight and was six day later in maturity. Both tests were being at grown at 9 locations. These tests are designed to identify the specific introgressions from *G. tomentella* that are associated with these increasing in yield.

Task 7: Development of breeding lines from perennial Glycine

We tested 67 advanced *G. tomentella*-derived lines in maturity group II, III and IV at 5 locations in 2016. We identified 12 lines that were significantly (p = 0.05) higher yielding than Dwight, the soybean parent. Six lines had Dwight as the female parent and thus had soybean cytoplasm and six lines had PI 441001 as the female parent and *G. tomentella* (PI 441001) cytoplasm. We are creating an MTA that will allow the use of these lines in bi-parental crosses in 2017. We identified approximately 100 new lines for yield testing that are derived from PI 441001.

We have callus tissue in culture from the following crosses: PI 505151 (*G. argyrea*) x Dwight; PI 440932 (*G. canescens*) x Dwight; Dwight x PI 505151 (*G. argyrea*); Dwight x PI 440932 (*G. canescens*); PI 559298 (*G. latifolia*) x Dwight; and Dwight x PI 559298 (*G. latifolia*).
We have been unable to get calli from the crosses with *G. latifolia* as either male or female to produce shoots despite several changes in medium. Small shoots are beginning to form from calli from the other crosses. Shoots from PI 505151 x Dwight were transferred to rooting medium but the shoots have failed to produced roots.

This summer we yield tested 186 new lines derived from crossing with *G. tomentella* PI 441001. Approximately 1/3 of the lines have *G. tomentella* as the female parent and thus *G. tomentella* cytoplasm.

Because of a hiring freeze within ARS, we were unable to make a new hire to continue the research on making successful crosses between soybean and other perennial *Glycine* species. Some cultures are still being maintained but most of the research is on hold. The new person has been selected and will be brought into the project as soon as the hiring freeze is lifted.

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

Task 1. Recruit an appropriately skilled PhD graduate student.

This student has been recruited.

Task 2: Communication and outreach.

Student learned history and methods for predicting and estimating realized genetic gains and developed a power point presentation that was used in a summer short course. Three students began work on development of videos by interviewing Ed Anderson and about 35 soybean farmers. From the interviews it is clear that the genetic gain presentation is not adapted to the perspective of soybean farmers. Soybean farmer's perspectives includes concepts such as yield potential and yield protection for the crop on their landscapes. The presentations are not ready for video development until they are reframed for the farmer's perspectives.

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

A one page research proposal to use Syngenta data to develop, validate and determine limitations of Realized Genetic Gain methods was added to a new umbrella agreement with Syngenta Soybean Product Development Program in April 2017. Syngenta was in the process of merging with Chem China and has not signed a new umbrella agreement. It is not clear whether Syngenta's data, previously used to compete for the Wagner Prize, will be available for this project. Alternative sources of data are being investigated.

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

Relevant data from the Uniform Regional Trial is being assembled under Objective 2, Task 2, so we are not repeating this time-consuming task here. The graduate student has not yet met with individual PI’s responsible for the URT’s. Relevant data from Syngenta have been identified and the Program Manager, Dr. Craig Davis, has been asked to meet with the graduate student. Based on these interviews student will determine QA protocols for the types of data. Gary Stacey has asked us to collaborate in writing a manuscript on QA protocols for these types of data.

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

QuGene is available through the Breeding Management System (BMS) of the Integrated Breeding Platform (see https://www.integratedbreeding.net/384/breeding-services/more-software-tools/qugene). The BMS code was unstable throughout the first year of our project, and it is still not clear whether the BMS will be available for free to academics in the US. So we developed our own simulation code. Our simulations have been used to investigate the impact of multiple cycles of Genomic Selection for traits consisting of additive genetic architectures with 40, 400 and 4000 QTL in genomes of the SoyNAM founders. Genotypic and recombination data for the markers in the founders can be imported to create the foundations for simulations.

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

URT data are being aggregated and placed in a database by Aaron Lorenz’ group (objective 2, task 2). The contract for transfer of SSPDP data has not been signed, so no activity to report for this data set this year.

Task 7: Develop four methods for calculating RGG.

A single method for calculating realized genetic gain was developed and published in a special issue of Interfaces journal dedicated to the Wagner Prize competition.