Project report

April 1, 2017

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.

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.

A postdoctoral researcher position was advertised via a quantitative genetics listserv. A qualified candidate has been identified and has verbally agreed to accept the position, and to start in June. The position description has been finalized, including funding and immigration details. The request for a waiver of posting has been sent to Talent Acquisition for review.

A protocol for image-based canopy phenotyping is being prepared by a graduate student in the Martin-Rainey research group. A framework for submission of data and pedigrees for statistical analyses will be developed and communicated by the post-doc as top priority in the coming months.

Progeny row phenotyping plans were discussed with cooperators. This was accomplished at the project meeting held during the 2017 Soybean Breeders' Workshop in St. Louis, and via follow- up email.

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.

We are currently exploring alternative ways to construct the genomic prediction models and will apply these once we have all the final SNP genotype data in hand. A manuscript is being prepared that will report the accuracy and optimization of genomic prediction within the NAM data, as well as the NAM-based predictions applied to the validation progenies grown in 2015-16.

Next steps include collecting the remaining "new" SNP data on the validation progenies, project the GBS SNPs onto the NAM progenies, perform the predictions, compare different strategies of genomic prediction model construction to phenotypic selection on plant row data, and make progress on the manuscript.

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. A. 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.

Next steps for this part of the project include: 1) Finish parsing and cleaning the pedigree data as best we can; 2) Upload all URT phenotypic data going back to 2004, starting with the most recent years; 3) Begin the process of compiling the available environmental data and linking it to the phenotypic data through the relational database.

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.

Next steps: 1) Genotype the 2017 newly entered URT lines; 2) Identify additional missing lines and genotype them; 3) QC and upload genotype data in relational database and link to phenotypic data.

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.

OBJECTIVE 3: Increasing additive genetic variance

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.

The use the phenotype and genotype information developed from the models to go back into the soybean germplasm collection and identify ~200 high-yield genotypes based on the information from the models is currently underway. We will share the results of the predictions with cooperators to make final decisions on sampling protocol for new accession from the collection for the validation study.

Based on this prediction, seeds will be obtained from the USDA Soybean Germplasm Collection for selected accessions and increased in Nebraska during 2017. This will be completed by April 15 to obtain seeds for planting the increases in 2017. The seed increase will provide seeds for 2018 and 2019 multi-environment yield tests to test and validate the predictions.

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 being 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 wild soybean parent.

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 lines 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.

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.

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

Source data from Syngenta are in a useful format and have been through a QA process. The mechanisms for transfer of the very large data sets from Syngenta field trials (including genotypic and meta-data for environments) to ISU are currently being discussed.

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

The source for data from public uniform regional trials have been identified, but are not in a format that is useful for aggregation and combined analyses. The next step is to convert the data into a format that will be useful for analyses.