SOYGEN 2: Increasing soybean genetic gain for yield by developing tools, know-how and community among public breeders in the north central US

Objective 1: Elevating collaborative field trials

Task 1: Development of a database to store, query, and distribute data from collaborative field trials

Drs. Nelson and Lorenz have mapped out an overall structure of this database, with appropriate dropdown menus, etc.

Weather data will be accessed from Daymet (or a similar resource) automatically utilizing web services and displayed to the user in real time.

Task 2: Updating the Uniform Soybean Trials

(1) Continued collection of genotypic data for UST as well as the SCN regional trials.

Another 288 breeding lines from the 2019 and 2018 NUST trials were genotyped with the 6K array this past winter. A total of 2238 lines have now been genotyped. Some 2018 lines were missed the first time and we will request those seeds from breeders as soon as campuses re-open. The 2020 NUST lines were collected and delivered to UMN for genotyping, which will occur once shelter in place orders have been lifted.

(2) Incorporation of weather station data or use of interpolated weather data from commercial sources

At this time, we anticipate accessing weather data from Daymet automatically utilizing web services and displayed to the user in real time.

(3) Evaluation of suitability/representation of environments

A graduate student will be starting in the fall on this work. In the meantime, a postdoc at Univ of Illinois will be advancing these goals over the summer.

1c. Key performance indicators

(1) Standardized data input methods will be developed and will include data quality control methods.

Data for NUST will be extracted from data sheets provided by NUST coordinators. Parsing scripts will be created to automatically extract data from the Excel data sheets.

(2) Existing data from collaborative trials will be quality checked.

Data for NUST 1989-2017 have been incorporated into the database.

(3) Collection of genotypic data from the Soy6KSNP chip for UT and SCN regional trial entries.

As stated above, we have genotyped another 288 breeding lines and are making plans to genotype the 2020 NUST and SCN NUST lines once campus re-opens.

(4) Weather data will be collected for the majority of the future Uniform Test field environments.

Data will be accessed programmatically from Daymet (or a similar resource) and incorporated into report pages.

1d. Deliverables

(1) Database framework for agronomic, environmental, genotypic, meta and other trait data for collaborative trials.

A preliminary database schema has been created to accommodate NUST data 1989-2017.

We all helped Rex develop a basic design. He is implementing this framework right now.

(2) Database populated with historical and current data from collaborative trials, including agronomic, environmental, genotypic, meta and other trait data.

Data for NUST 1989-2017 have been incorporated into the database.

(3) Data from the uniform tests will become more useful as it will be connected to environmental and genotypic data.

Incorporation of environmental data into reports will provide a way to start addressing issues of genotype by environment interactions in phenotypic values.

Objective 2: Development of a genomic breeding facilitation suite

Task 1. Genotyping methods

We plan to include optimization experiments when we receive more samples from breeders during the summer. We have done a load test to identify potential problems with running a large number of samples and are working on improving our sample tracking, the genotyping workflow, and improving the analysis pipeline.

Task 2. Imputation methods

The Hudson group is developing a haplotype map of soybean germplasm, and using this to detect signatures of selection in the U.S. Soybean Germplasm. We are using haplotype imputation in R to be able to convert haplotype location and information between the whole genome sequences, 50k array, 6k and smaller genotyping arrays. These scripts are complete, and we are developing imputation scripts for use by other labs with specific needs to convert between genotyping platforms. Since the scripts are platform-specific, this must be done in collaboration with breeders using specific platforms. So far, we have focused on 6k – 50k conversion for this purpose.

Using these methods, have applied a combination of different methodologies to discover the biological basis of the history of genetic breeding in US.

By analyzing the results that we obtained so far, we have observed a partial parallelism between the conventional public elite lines and the alternative gene pool lines previously developed by Dr. Randall Nelson of USDA-ARS. Our results from haplotype-based analyses (Rsb, XP-EHH, and hapFLK) support the hypothesis that in different gene pools, unique haplotypes have been fixed or are being fixed in different rates. Among the candidate haplotypes under selection, we identified 138 haplotypes underlying QTL regions, including 46 candidate haplotypes underlying QTLs related to yield-related traits. These promising results may contribute to the development of varieties with greater combinations of alleles conferring advantageous characteristics, by combining conventional and alternative gene pool germplasm.

Task 3. Genomic Prediction Facilitation Suite

A postdoc, Sushan Ru, was just hired (April) at UMN to take charge of this task. She has already fully surveyed genomic data management systems and tools available (or lack thereof). We are currently laying plans to put together an easy-to-use workflow that will take data from the Soybase database and provide predictions on target populations using genomic selection best practices.

2c. Key performance indicators

(1) Genotyping of 10,000 breeding lines using targeted GBS approach on 1k SNPs during first year of project.

I think this is for David. We have genotyped three plates of PYTs through David, but technical problems persist in the bioinformatics, and the COVID19 crisis has basically halted further data collection.

(2) Beta version of R script to impute underlying whole-genome haplotypes developed.

Several imputation scripts have been developed for use in the Hudson lab. A beta testing version of a conversion script to convert 50k data into haplotype information for the 6k and smaller arrays has been supplied to the Lorenz lab for testing.

(3) Workshop or webinar given on application of genomic selection to soybean breeding.

I gave a special workshop at the 2020 Soybean Breeders' Workshop on March 2, 2020. Over 45 people attended this session covering an entire morning. Here is a link to the workshop materials: https://drive.google.com/open?id=16ZNJZYYaPisusKowUAX10F1KPLTEEf7n

A survey was sent out to attendees following the workshop, and all survey responses were favorable or highly favorable. This workshop will be offered again including tools and data built on the progress from the SOYGEN2 project.

(4) Genomic data management system and allied analysis tools for adoption by soybean breeding community identified.

Postdoc hired one month ago to make progress on this KPI.

2e. Deliverables

1) Streamlined public genotyping service for the public soybean breeding sector at a low enough cost to afford genomic selection on a wide scale.

We received two batches of samples from KSU for the rapid cycling project. We genotyped the 210 samples and returned the data to KSU.

2) Workshop on genomic selection delivered to public soybean breeding community.

Completed March 2, 2020.

Objective 3: Evaluation of soybean breeding methods that increase gain

Breeder participation in implementation and evaluation of selection models

3b. Brief description of proposed research

Task 1. Advanced spatial analysis

To set up protocols, in 2019, soil sampling was done on progeny rows and early stage yield trials near the Ames locations, and soil maps have been made (collaboration with Dr. B. Miller, ISU). Weather stations were also deployed in the field to have more precise measurements across the field. Drone RGB imagery was also taken in these field tests. These will be repeated to standardize the data collection and analysis pipeline in 2020.

Task 2. Development of breeding program specific genomic prediction models

Plants are being grown to produce tissue that will be used in the marker analysis needed for developing the prediction models.

To date, we have genotyped ~350 experimental lines and varieties with the Illumina 6KSNP chip technology from within our breeding program. These data represent 3 years of genotyping only the advanced lines in our breeding program, each year. Starting in 2020, we will also be genotyping experimental lines (~1500) that are advanced from the 2019 plant row stage to the first preliminary yield trails during the summer of 2020. At the end of 2020, we will have a large enough population size to start developing robust prediction models for cross validation and evaluating selection strategies based on genomic estimated breeding values in 2021.

Task 3. Genomic plus secondary trait selection at the progeny row stage

Progress: We have mapped out various scenarios for implementation of this goal and discussed them in the Rainey Lab and among the project PIs. I have created detailed schematics projecting out three years. We have designed an experiment for 2020 and packaged seed to compare heritability of secondary trait phenotypes and yield in small format progeny row plots vs. replicated yield trial plots.

Dr. Martin Rainey also was the recipient of a $500,000 AFRI Plant Breeding grant (ranked 'Outstanding') for three years that leverages this objective, and provides additional funding to develop non-destructive drone-based biomass estimation protocols applicable to the collaborators on this project.

Task 4. Exploration of genomic prediction to reduce unfavorable correlations between seed yield and protein

The collected genotype and phenotype data on the NUST trials has been used to predict the genetic correlation between protein and yield among all possible breeding populations that could be created. As expected, the mean of all these correlations is highly negative, but few populations are predicted to have a less severe correlation. Crosses are being designed that are predicted to produce progeny superior for both yield and protein. (Aaron)

Task 5. Rapid cycling

Create a random mating population using 13 parental lines to generate 100 to 150 F1s each intermating cycle. Genotype and use genomic selection to identify the top 30 to 40% of the F1 plants before flowering and random mate the selected F1s each season. Selection criteria will include yield, maturity, protein and oil concentrations, and genetic distance. Attempt to perform this selection and intermating process three seasons per year, one season in the field, and two seasons in the greenhouse during the fall and winter. Continue this intermating and selection process to complete three cycles of genomic

selection. Following each cycle (C0, C1, C2, and C3) inbreeding of the selected and unselected F1s will be implemented for multiple generations to produce F4-derived lines of unselected (random) and selected (based on genomic selection) for each cycle of selection. These lines will be evaluated in replicated field trials to characterize the effectiveness of the genomic selection and rapid cycling methodology. (Bill)

Task 6. Evaluation of putative "yield" alleles

F1s have been generated in order to develop NIL varying for one pair of putative yield alleles. However, material transfer agreements (MTAs) from the USDA were unable to be obtained for other selected parental lines. This was due to a lack of proper MTAs in place for generating these parental lines. A new set of parental lines have been selected and we are initiating crosses this summer.

3c. Key performance indicators

(4) Genotyping of advanced lines, development, and cross-validation of breeding program specific models (Task 2).

At Univ of Illinois, we started growing in a greenhouse approximately 1800 lines that were evaluated over the past two years in the University of Illinois preliminary and advanced yield tests. Tissue had been collected from these plants that will be used in DNA extractions. Tissue was collected on about one-half of the lines until these activities were halted because university lab activities were shut down do to Covid-19. This will be completed once the university allows us to restart lab work.

At Univ of Missouri, to date, we have genotyped ~350 experimental lines and varieties with the Illumina 6KSNP chip technology from within our breeding program.

At the Ohio State University, we started growing in a greenhouse approximately 1000 lines that were evaluated in preliminary and/or advanced yield tests in 2018 and/or 2019, isolated DNA was being quantified and normalized when all lab activities were shut-down due to COVID-19. These activities will be resumed when we are able to begin lab work again.

(8) Generate crosses for 5 cross combinations based on breeder selections and 5 cross combinations based on genomic mating selections for protein and yield (Task 4).

This is being done this spring, right now.

(10) Perform crosses, genotyping, and line advancement according to rapid cycling breeding scheme (FY20-22) (Task 5).

In Nebraska, crossing and genotyping to be conducted during 2020 season.In Kansas, during the summer of 2109, 13 elite parents ranging in maturity from mid-group III to early-group IV were intermated using a diallel design in 138 parental combinations that produced a total of 627 CO F1's. The parents were selected phenotypically for seed yield, maturity, protein and oil concentrations, and for genetic diversity based on pedigree information. In the Fall of 2019, about 140 CO F1's were planted in the greenhouse and genotyped. Based on genomic predictions for seed yield, maturity, protein, oil and genetic distance, about 43 F1's were selected and intermated at random. Over 300 emasculations were completed and 221 C1 F1 seeds were produced. In the Winter of 2020, about 140 C1 F1's were planted in the greenhouse and genotyped. Based on genomic predictions for seed yield, maturity, protein, oil and genetic distance, about 43 F1's were selected and intermated at random. Over 400 emasculations were completed with seeds to be harvested in May 2020 for planting in the field in June. To eventually produce F4 derived lines, a random sample of F2 seed from all CO F1's were advanced for inbreeding in

the F2 and F3 generations during the winter of 2019/20. Random samples of F2 seed from C1 and C2F1s will be advanced in the field in 2020.

(12) Develop and increase seed for NILs varying for alleles at putative yield loci (Task 6).

F1 have been generated and are currently growing for production of F2 seed to study one putative yield locus.

3d. Deliverables

(3) Application and limitations established for rapid cycling genomic selection in soybean. For the Nebraska rapid cycling GS, we are focusing on water productivity (WP) and genetic diversity in the rapid cycling set of material. The training set for water productivity comes from results of a dissertation study on soybean response to water. This is important across the north central region in rainfed production systems, as well as for 2.5 million acres of irrigated soybean production in Nebraska. (George)

Objective 4: Characterization and use of the USDA Soybean Germplasm Collection, a foundation for future success

4c. Key performance indicators

(1) Soybean breeding programs choose soybean accessions for use in their breeding programs based on results of this work.

The predictions for all accessions in the collection were provided to participants in early 2016. Some programs have used the original set of genomic predictions from our first stage of evaluation of the 500 germplasm accessions to select new soybean PI accessions for use in their breeding programs to increase genetic diversity and yield. We selected several PIs from the list that represented new diversity in the US soybean germplasm pool for use in our Nebraska breeding program. Other programs may provide information on their specific use of selections from this work.

4d. Deliverables

(1) Comparison of sampling methods and effective ways to efficiently sample the genotype collection, particularly for improvement of quantitative traits like yield.

The SSD method is more effective at sampling the range of genetic diversity in the germplasm collection, and so leads to more efficient sampling and evaluation of a subset of lines for quantitative traits like yield. For example, our group of about 145 SSD selections in that sample represented the range of genetic diversity in the 19,000+ accessions in the USDA Soybean Germplasm Collection as well as the entire set of 500 entries.

(2) Means and variances for traits in the different sampling groups (so, effect of sampling method on those estimates).

Means for Yield, Seed Weight, Plant Height, Lodging, Shattering and Maturity were similar for each of the sampling groups for yield and most other traits. Interestingly, the genotypic variance for lines in the SSD group was more than twice that for the CLU and RAN samples, and overall (ALL), while the environment variance and line x environment variance estimates were similar across groups. Variance component estimates for traits in each group, and specifically for yield, show distinct difference in the SSD sample group with major effects of the main effect for genotype and environment, but relatively

small genotype x environment interaction variance. Details and implications of this finding for modeling, prediction, and parent selection for traits like yield will be discussed in the publication that is being developed.

(3) Identify loci associated with yield and other traits in this diverse panel of accessions that represents the genetic diversity in the collection, so we may ID new loci and alleles that will be useful for commercial and public breeding programs.

The genome wide association analyses for each trait and sample group show different results among sample groups, with some overlapping loci detected. It seems like the SSD group may provide a more robust analysis of the genotype-phenotype association and so shows fewer loci, whereas some of the associations in the other sample groups may be spurious. These results are being investigated further as we develop the manuscript for publication.

(4) Provide genomic predictions for yield (done), maturity, seed protein and oil %, and other traits as appropriate, for all untested accessions in the USDA Soybean Germplasm Collection.

This was done previously and provided the inputs for the current discussion.

(5) Investigate genotype-environment interaction effects on traits, and evaluate stability of yield and composition traits across environments.

This was done previously and provided the inputs for the current discussion.

(6) Use data/results in implementation in Objectives 1, 2, and 3 of this project (FY20-22).

With these results, we can start to use this information for implementation in other parts of this project.

(7) Preliminary analysis of data from the validation set of 250 entries.

Only basic yield and field results done. This will be a focus later this year after submission of the manuscript that documents results from the first training set of 500 entries.