May 14, 2021 Annual project report.

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

Investigator Contact Information:

Leah McHale (PI), Department of Horticulture and Crop Science, The Ohio State University, Columbus, OH 43206, 614-292-9003, mchale.21@osu.edu

Pengyin Chen, Division of Plant Sciences, University of Missouri, Columbia, MO 65211, 573-379-5431, chenpe@missouri.edu

Brian Diers, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, 217-265-4062, bdiers@illinois.edu

George Graef, Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588, 402-472-1537, ggraef1@unl.edu

Matthew Hudson, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, 217-244-8096, mhudson@illinois.edu

David Hyten, Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588, 402-472-3255, david.hyten@unl.edu

Aaron Lorenz, Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, 612-625-6754, lore0149@umn.edu

Katy Martin Rainey, Department of Agronomy, Purdue University, West Lafayette, IN 47907, 765-494-1212, krainey@purdue.edu

Nicolas Frederico Martin, Department of Crop Sciences, University of Illinois, Urbana, IL 61801, 217-300-3016, nfmartin@illinois.edu

Andrew Scaboo, Division of Plant Sciences, University of Missouri, Columbia, MO 65211, 573-882-3462, ScabooA@missouri.edu

William Schapaugh, Department of Agronomy, Kansas State University, Manhattan, KS 66506, 785-532-7242, wts@ksu.edu

Asheesh Singh, Department of Agronomy, Iowa State University, Ames, IA 50011, 515-294-7920, singhak@iastate.edu

Dechun Wang, Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI 48824, 517-353-0219, wangdech@msu.edu

Collaborator Contact Information:

Rex Nelson, USDA-ARS, Ames, IA 50011, 515-294-1297, rex.nelson@usda.gov

Objective 1: Elevating collaborative field trials

This project aims to develop a database to collect/compile, store/manage, query, and publicly distribute data from collaborative field trials (Northern Uniform Soybean Trials and SCN Trials). We aim to further add value to these collaborative trials by genotyping the breeding lines tested.

1c. Key performance indicators

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

We have communicated with the Northern Uniform Soybean Trials' collaborators to make an agreement to access GPS coordinates of test locations.

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

We have mapped out the structure of and databased the UT data set. We are currently testing the database.

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

We collected 6K genotype data on all 2020 UT lines. The 2020 SCN UT lines will be planted in the field along with all 2021 UT and SCN UT lines for tissue collection and genotyping.

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

Weather datasets were collected in the site years corresponding to NUST field trials from using the geographic coordinates of the field trials linked with the DAYMET weather data. This information along with field trial phenotypic information will be used to compare the year to year site trialing similarity.

(5) The data from the NUST will be analyzed to determine the usefulness of test locations in predicting the performance of the experimental lines.

Weather datasets along with field trial phenotypic information is be used to compare the year to year site trialing similarity.

1d. Deliverables

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

Database tables and draft query user interfaces have been created. Beta testing of the interface by project participants continues.

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

Phenotypic data from collaborative trials from 1989 to the present have been loaded into the data tables and are accessible to project participants. Environmental data will be available through an interface to the DayMet meterological API.

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

GPS locations of field trials and genotypic data of breeding lines are available.

Objective 2: Development of a genomic breeding facilitation suite

Genomics-assisted breeding entails the use of genome-wide molecular marker data to aid in breeding decisions that make breeding programs faster, more efficient, and more effective. To better adopt these methods, we aim to make high-quality, inexpensive genetyping methods, decision support tools, and user friendly analysis pipelines available.

2c. Key performance indicators

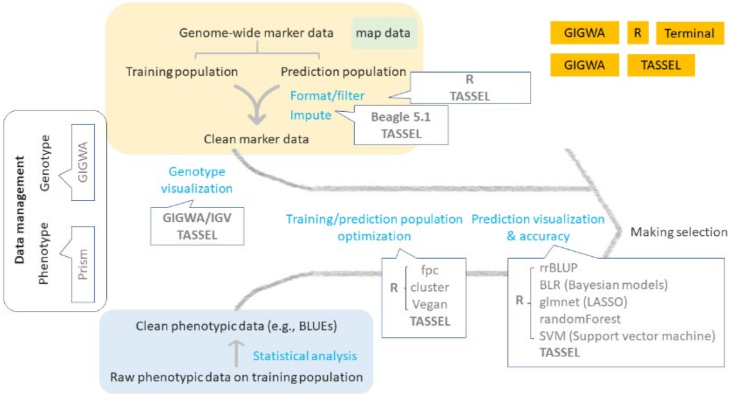
(1) Genotyping of 10,000 breeding lines using targeted GBS approach on 1k SNPs during first year of project.

We have received 7,730 DNA samples to run with the 1k SNP set. We are currently processing these samples. The first 2,592 are in the process of being sequenced.

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

A haplotype map of soybean germplasm is being developed, which will aid in transferring genotypic data across platforms. 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 such as 1k. These analysis tools are complete, and we are developing tools for use by other labs with specific needs to convert between genotyping platforms. So far, we have focused on 6k – 50k conversion for this purpose..

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



During this past reporting reporting period we were able to install a genome-wide marker database called GIGWA (https://gigwa.southgreen.fr/gigwa/). We have deposited our current genome-wide marker data into this, including all the genotype data collected on the UT as part of this project. A workflow of software tools and scripts was initiated to seamlessly combine data held in this database with phenotypic data and genomic prediction models to ease the use of genomic selection in a practical breeding context. There are a few steps that need to be developed, such as low-to-high marker density imputation and training population optimization. The current postdoc left for a permanent position, and we are currently seeking another postdoc to continue this work.

On a related front, co-PI Nelson, with input from Lorenz, is research the adoption of a platform called BreedBase (breedbase.org). We are hoping this can be installed at Soybase and be available to public breeders for depositing the phenotypic and genotypic data and facilitate the use of genome-wide marker data for breeding. This is in the early stages of development right now.

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.

This first batch of 7,000 lines is helping us to streamline our submission process and determine what parts of the genotyping process need to be improved for this summer.

Objective 3: Evaluation of soybean breeding methods that increase gain

Many alternative strategies and tools exist to improve genetic gain. With expertise in genomics, characterization of genetic diversity, analyses of spatial variation and breeding methodology, this team will be able to simultaneously attempt to improve the genetic gains in our breeding programs while also asking and answering questions to identify best breeding practices.

3c. Key performance indicators

(1) Preliminary single-site validation of spatial statistics are selection of added growth stage and/or drone based phenotyping and soil parameter factors (Task 1).

Preliminary yield prediction models have been run on single location progeny rows from 2019 using elastic net, ridge regression and lasso. Preliminary results show RMSE of 7 bu/acre and R2 of 0.69. Models have shown relative maturity and pedigree information to have the largest effect on yield. Soil parameters and canopy area have also shown some significance. Soil data is extracted using fine scale soil maps generated in collaboration with soil scientist Dr. Miller and his postdoc Dr. Khaledian. With these soil maps we get soil nutrient data (N,P,K, CA,MG, CEC, NO3, OM) as well as soil texture data on a 3m x 3m scale. Further machine learning and model development and selection criteria are being developed with Dr. Sarkar and his graduate student Luis Riera.

(3) Validation and selection of spatial statistics and added factors based on multi-location data (Task 1).

In collaboration with statistican Dr. Dutta and his graduate student, Dongjin Li, we have prepared a tutorial using the statgenSTA R package. This tutorial includes videos, and an html notebook showing the steps from data preparation, fitting and running models, as well as outlier analysis. The statgenSTA package allows users to fit traditional non-spatial models, as well as spatial models, by including row and column information as well as replications. Users can use the lme4, SPATs or ASREML packages for fitting the data. This tutorial will be shared with the breeding community prior to the fall season. We used the SPATs engine, which uses a penalized spline for spatial correction. This allows for a more dynamic spatial correction compared to the traditional moving means corrections. We also used this tool in our spatial adjustments for 2020 yield trials, and compared it with the traditional moving means method that we have used in the past. We have not validated results yet, on which method used for selection gives more accurate results, and this is an on-going work.

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

7000 advanced lines have been submitted and in the process of being genotyped (see Objective 2).

(7) 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).

We used genomic prediction to predict the mean, variance, yld-pro correlation, and superior progeny mean of all possible crosses among 2019 and 2020 UT lines. We made this information available to all SOYGEN2 breeders for the planning of crossing blocks for each breeding program.

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

Crosses were made in Nebraska summer 2020 and sent F1 seeds to Puerto Rico to grow F1 plants from October ’20 to January ’21. Intermating among F1 plants were attempted, but virus issues in Puerto Rico caused issues and we were not able to obtain all of the F1 x F1 crosses. Instead, F2 seeds were harvested from all of the confirmed F1 plants and are now crossing among F1:2 lines for the second intermating.

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

Due to inability to MTA from the USDA for many of the cultivars used in the pedigrees of these lines, we were only able to complete a single cross combination: LG09-8165 x LG11-5120. F2 seed has been generated and is being advanced.

(12) Develop markers for selection of putative yield alleles (Task 6).

Markers have been developed based on key SNPs at four potential yield loci to be tested in families derived from the LG09-8165 x LG11-5120 cross.

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

Our previous NCSRP project included an evaluation of 750 soybean accessions from the USDA Soybean Germplasm Collection in yield tests taking place in 30 environments to obtain high-quality phenotype data on this diverse panel of accessions representing the genetic variation of the entire Collection. This work has allowed us to identify best-practices for sampling of large populations. Genotype-phenotype associations have been made, identifying some key loci for selection in breeding programs. Additionally, genomic predictions for yield have been completed and provide a means to select predicted higher yielding lines from the 20,000 germplasm accessions within the full Collection.

4c. Key performance indicators

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

Predictions for crosses are now currently being obtained. This work has been leveraged in other projects to identify lines with predicted high yield.