Final project report (FY22)

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

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

Carrie Miranda, Department of Plant Sciences, North Dakota State University, Fargo, ND 58102, 701-231-8136, carrie.miranda@ndsu.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 Federico 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

This SOYGEN (Science Optimized Yield Gains across Environments) project leverages and builds upon ongoing and previously funded work to increase soybean genetic gain for yield and seed composition by developing tools, know-how and community among public breeders in the north central US. Specifically, we have created and tested (or are testing) breeding resources and methods that can be applied to our own breeding programs, or more broadly to soybean breeding programs in general.

Objective 1: Elevating collaborative field trials

Key performance indicators

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

Forms provided to Northern Uniform Regional Trial collaborators were updated to include GPS coordinates. These are now consistently reported for all trial locations. Additionally, data from forms are uploaded to a database. Future plans entail direct uploading by collaborators of data to a SoybeanBase database, however this is still in progress.

https://soybase.org/ncsrp/queryportal/

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

Uploading data (past and present) to a database requires quality checking, which has been done https://soybase.org/ncsrp/queryportal/

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

We have genotyped a total of 3813 UT and SCN UT lines since 2019. We now have a database of 2510 NUST lines genotyped with the 6K SNP chip uploaded to soybeanbase.breedinginsight.net. In 2020 we switched to genotyping via low pass sequencing and imputation provided by Gencove. Imputation accuracy was determined to be extremely high, >99%. This provided us with many many more SNPs (~3 million versus 6000), allowing much more powerful analyses to be performed on this germplasm in future years.

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

Regular reporting of GPS coordinates of field trials allows us to connect to weather databases. We can do this through Soybeanbase and have loaded templates for the 2022 trials into Soybeanbase. Weather data from GPS coordinates was used to determine the independence of trial sites.

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

We leveraged the newly acquired genotype data and combined with our new database holding the UT phenotypic data to test how well genomic prediction might work with the UT trials. We used a leave-one-trial out cross-validation scheme, which basically means we dropped all data from one complete trial out of the dataset, and used the remaining data to develop a genomic prediction model. We then used this genomic prediction trial to predict the trial left out, and correlated observed yield performance with predicted yield performance. We only designated those trials containing more than 20 genotyped lines as validation trials for better estimates of correlations coefficients. This left 17 validation trials. Prediction accuracies were all quite good, ranging from 0.46 to 0.95 (see table below). This indicates that the genotype-phenotype data resource we began building as part of this project has value in terms of assisting future efforts towards genomics-assisted breeding.

Table 1. The NUST historical data was used as a training population to predict the performance of experimental strains for seed yield in specific large tests with at least 20 genotyped experimental strains from 2018 to 2020. Estimates of prediction accuracy (r_{MG}) values were obtained by dividing the predictive ability (r_{MP}) by the square root of the phenotypic reliability (*i*). A 95% confidence interval (in parentheses) for the predictive ability (r_{MP}) was estimated from 10,000 bootstrapping samples.

Test	$r_{MP}(\hat{y}, y)$	$r_{MG}(\hat{y}, y)$	**N***
PTIV2019	0.64 (0.33, 0.95)	0.95	24
UTIV2020	0.71 (0.45, 0.96)	0.89	20
PTII2020	0.61 (0.47, 0.75)	0.85	67
PTIII2018	0.76 (0.60, 0.92)	0.84	24
UTIII2018	0.79 (0.60, 0.98)	0.83	29
UTIII2020	0.63 (0.36, 0.90)	0.77	39
PTII2018	0.63 (0.43, 0.84)	0.75	24
PTIV2020	0.54 (0.24, 0.84)	0.74	26
UTII2020	0.41 (0.12, 0.69)	0.72	39
PTIII2020	0.49 (0.30, 0.68)	0.70	63
PTI2019	0.58 (0.38, 0.78)	0.63	34
PTIII2019	0.44 (0.24, 0.64)	0.62	55
UTIII2019	0.5 (0.09, 0.90)	0.60	27
UTII2018	0.49 (0.15, 0.84)	0.59	26
PTI2020	0.47 (0.27, 0.68)	0.59	30
PTII2019	0.42 (0.17, 0.67)	0.51	60
UTII2019	0.41 (0.13, 0.69)	0.46	35

* Number of genotyped experimental strains available in each test.

Deliverables

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

Soybeanbase (https://soybeanbase.breedinginsight.net/) and a SQL database hosted at Soybase (https://soybase.org/ncsrp/queryportal/) have been created and are available for researchers to deposit their data. The long-term plan is to host both genotype and phenotype data at Soybeanbase, which is part of SOYGEN3 objectives. Currently, Soybeanbase hosts genotype data on 2510 UT and SCN UT breeding lines. We are still determining how to host data for the 1303 lines genotyped with low-pass sequencing. Soybeanbase also holds data for SoyNAM and internal breeding lines, amounting to genotype data being stored for over 10,000 genotypes. The Soybase SQL database holds phenotypic

data on more than 8000 advanced breeding lines dating back to 1993. The phenotypic data represents over 1650 unique environments, from years ranging from 1993 to 2021. Data from 2022 is currently being imported. The total dataset consists of over 128,000 yield datapoints, as well as data on 18 other traits including maturity date, seed composition, and disease resistance.

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

We performed several analyses on the genotype and phenotype data that make up this dataset, and we have prepared a manuscript that is very close for submission to a peer-reviewed scientific journal. In this manuscript, we have made the UT genotype and phenotype data public, characterized the data available to researchers all over the world, determined the genetic relationships among all lines submitted to the UT, made genotype-phenotype associations, and tested genomic prediction models. Some interesting findings included the lack of strong population differentiation among breeding programs. This finding highlights the role of the cooperative Uniform Trials in facilitation of germplasm sharing among breeding programs, which helps all programs achieve greater sustained genetic gain. Secondly, we found the biggest driver of population differentiation among maturity groups was the E2 locus, with a few other loci showing effects. We also identified some genomic regions lacking genetic diversity in one maturity group, but for which there was genetic variation in other maturity groups. This could help future breeding efforts identify such regions for targeted incorporation of diversity into key genomic regions lacking diversity, perhaps caused by genetic drift. We performed genome-wide association mapping, and found a total of 30 marker-trait associations representing 30 independent QTL. These results helps researchers determine which loci are driving phenotypic variation in the UT germplasm, and tells us that this dataset contains good genetic signal for performing future analyses perhaps on specific questions. Finally, as mentioned above, we were able to train accurate genomic prediction models using these data.

(3) Breeders will better understand how to weigh data from different environments of the NUST understand where new cultivars be more likely to be adapted and tested successfully.

Ranking of entries is dependent on Uniform trial locations, and though site redundancy (in terms of cultivar ranking was correlated to physical distance between trial site locations, it had a slightly higher correlation to environmental variables. The most influential trial sites (those which other sites clustered around in terms of cultivar ranking) were Ames Iowa, Urbana Illinois, Manhattan Kansas, West Lafayette Indiana. The most influential variables grouping sites together were coldest quarter precipitation, driest month/quarter precipitation, annual mean temperature, and annual precipitation.

Objective 2: Development of a genomic breeding facilitation suite

Key performance indicators

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

Participating breeding programs each genotyped ~2500 lines as part of the selection experiment, and are continuing to genotype new and advanced breeding lines to develop breeding program specific training sets and use these training sets.

Development of the 1K marker set has been published: Wang, H., B. Campbell, M. Happ, S. McConaughy, A. Lorenz, K. Amundsen, Q. Song, V. Pantalone, D. Hyten. 2022. Development of molecular inversion probes for soybean progeny genomic selection genotyping. Plant Genome doi.org/10.1002/tpg2.20270.

(2) Annual workshop or webinar given on application of genomic selection to soybean breeding.

An in-person workshop was held at the SBW in 2020, a second training was done specifically for the SOYGEN team via Zoom in 2021, more recently the Breeding Insight team has made themselves available for training of the soybean community in database management through our implementation BreedBase.

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

Excitingly, we have identified and adopted the breeding database and genomic data management system, BreedBase. The soybean implementation of this is called Soybeanbase (https://soybeanbase.breedinginsight.net/). This has been adopted and implemented through the help of Rex Nelson and the Breeding Insight team.

We have also created a streamlined analysis pipeline that can be run as an R shiny app, greatly increasing the ease with which these data can be analyzed. The app takes data in a standardized format exported from Soybeanbase and executes many of the major steps in a genomic prediction pipeline, including marker data filtering, imputation, training population optimization, model selection, cross validation, and prediction of genetic values for defined target population.

https://github.com/UMN-Lorenz-Group/SoyGen2App

To run app, click on “launch binder” icon under SoyGen2App heading → Go to App folder under files → open app.R → Run all code to launch application.

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.

Previously we had developed a set of markers, genotyping methods, and DNA isolation methods to cost effectively provide a genotyping service to the SOYGEN breeders (David Hyten, UNL). Yet, recognizing the throughput limitations of an academic lab with an active research program, we have more recently been working with Agriplex.

(https://www.agriplexgenomics.com/1k-soy?utm_medium=email&_hsmi=244361042&_hsenc=p2ANqtz-8clalasz9m8NLRed190_rTJ1F-kI3CIngv8nPCRRzbDvDj8dMc5CXAces3K1CjF9vnooKH8gjrpzQ2cCAiAKRgM0dafQ&utm_content=244361042&utm_source=hs_email)

The Soybean Community panel, which consists of 1326 SNPs, was developed in collaboration with members of the SOYGEN team and AgriPlex. Using their AgriPlex Connect program, public soybean breeders and geneticists can take advantage of discounted pricing and expedited turn-around times during critical times of the year for selection.

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

As above, an in-person workshop was held at the SBW in 2020, a second training was done specifically for the SOYGEN team via Zoom in 2021, more recently the Breeding Insight team has made themselves available for training of the soybean community in database management through our implementation BreedBase.

Objective 3: Evaluation of soybean breeding methods that increase gain

Key performance indicators

(1) Genotyping of 2500 F4 lines in two years for each participating breeding program.

This was completed for four participating breeding programs using either the 1K set of SNPS from UNL or the 1.3K set of SNPs from AgriPlex.

(2) Application of 4 different selection schemes.

Thus far, only a single year of validation has been summarized, thus no significant results have been reported. Yet, in an application of random selection, genomic selection, and a combination of genomic selection plus selection on canopy coverage from University of Minnesota, genomic selection did increase our chances of selecting superior lines, however we did not achieve statistical significance in most cases. A second year of validation data has been collected in 2022. The student managing this project has since graduated, and the data is being analyzed right now by a student at the University of Missouri. In the first year, we did see genomic prediction work quite well, especially in the southern region of Minnesota. This could be due to the fact that the MG I germplasm is better connected to the UT training set used, and these populations were larger than the ones tested in central and northern Minnesota. In the table below, the top ten lines by validation yield data were defined for each location. The tabled valued indicate whether genomic prediction (GP) canopy coverage selection combined with genomic prediction (CC+GP) or random selection selected that line (Table 2). It can be seen that, especially in the south on average and the two individual south locations (Lamberton and Waseca) that the best 10 lines were much more likely to have been chosen by either GP or CC+GP than random selection.

Table 2. Performance of genomic prediction (GP) canopy coverage selection combined with genomic prediction (CC+GP) or random selection selected that line from UMN in test locations going from North to South.

Rank	North	Moorhead	Shelly	Central	Morris	Rosemount	South	Lamberton	Waseca
1	CC+GP	CC+GP	GP	RAND	RAND	GP	GP, CC+GP	GP	CC+GP
2	RAND	RAND	GP	CC+GP	RAND	GP, CC+GP	CC+GP	GP, CC+GP	GP, CC+GP
3	GP	RAND	CC+GP, RAND	GP	RAND	GP	GP	CC+GP	GP, CC+GP
4	CC+GP	GP, CC+GP	CC+GP	RAND	RAND	RAND	GP, CC+GP	GP, CC+GP	GP
5	RAND	CC+GP	RAND	GP, CC+GP	RAND	CC+GP	GP, CC+GP	GP, RAND	RAND
6	GP	GP	RAND	GP, CC+GP	GP	RAND	RAND	GP, CC+GP	GP
7	RAND	CC+GP	RAND	RAND	RAND	GP, CC+GP	GP, CC+GP	GP	GP
8	CC+GP	RAND	RAND	GP	CC+GP	CC+GP	CC+GP	GP, CC+GP	GP
9	GP	GP	GP	GP	RAND	CC+GP	CC+GP	RAND	GP, CC+GP
10	GP	RAND	GP	CC+GP	CC+GP	CC+GP	RAND	GP, CC+GP	GP, CC+GP

Selection category: GP: genomic prediction; CC+GP: canopy coverage plus genomic prediction; Random: random sampling.

Similarly, data from University of Missouri showed no significant difference among selection methods based only on a single year of data; yet, random selections had lower yield estimates (Figure 1 and 2).

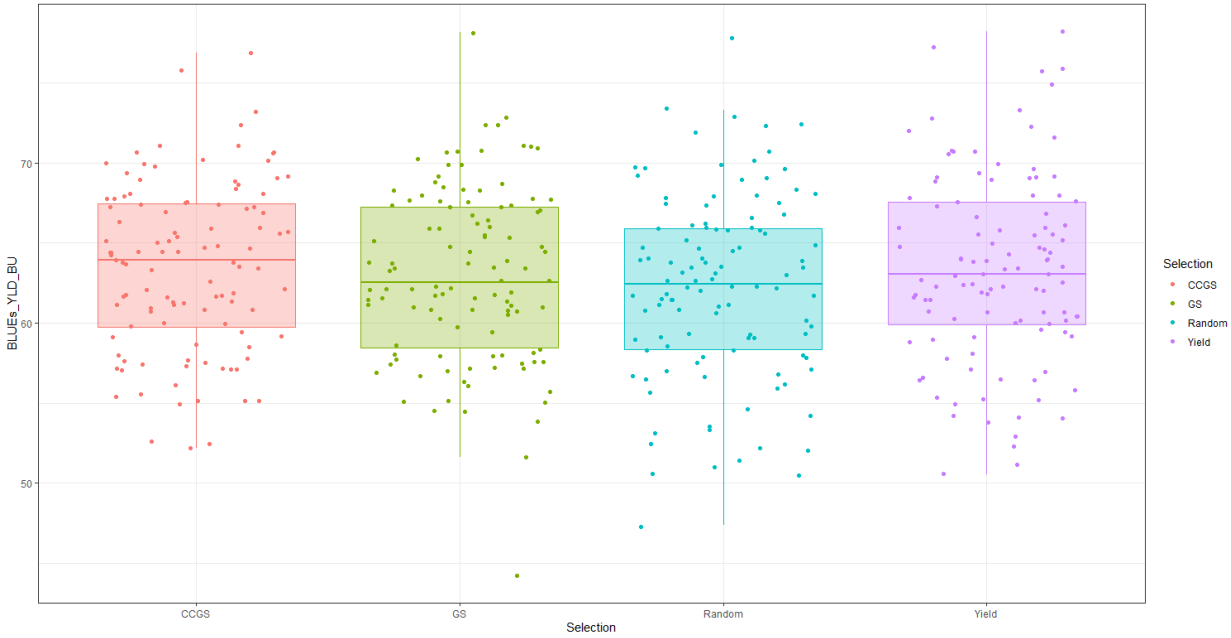

Figure 1. Mean separation of yield for progeny within groups selected by four different methodologies including 6% selection intensity based on GEBV (GS), 12% selection intensity on GEBV + 50% selection intensity based on canopy coverage (CCGS), 6% selection intensity on yield (Yield), and 6% random sampling (Random) from the University of Missouri breeding program grown at one location in Missouri during 2022.

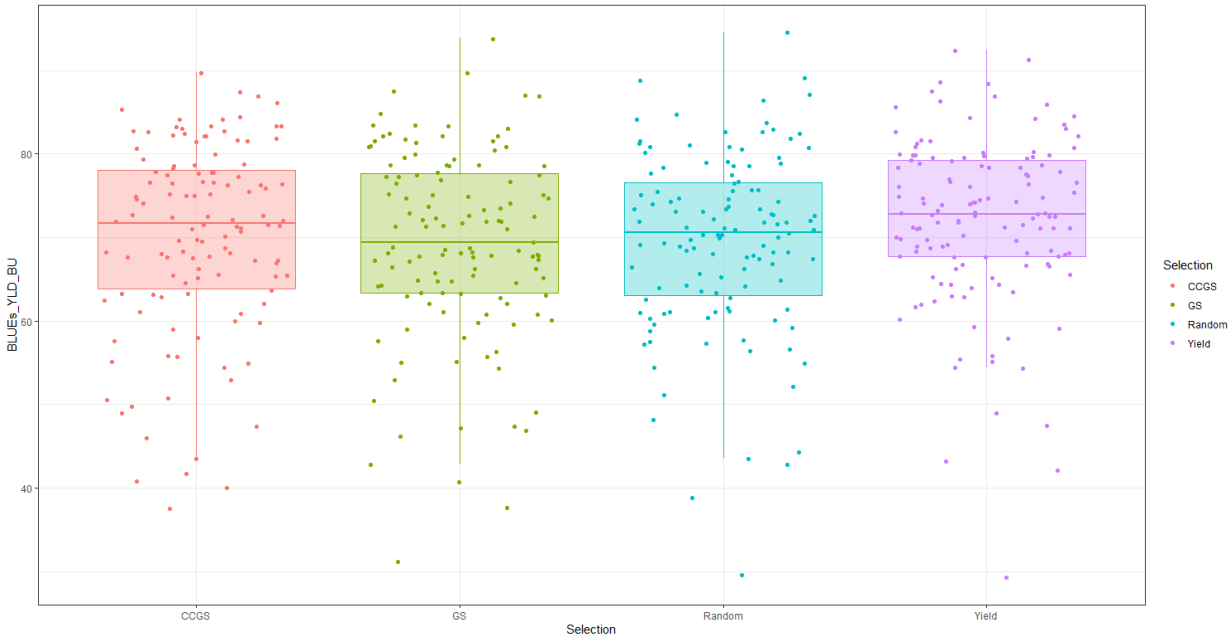

Figure 2. Mean separation of yield for progeny within groups selected by four different methodologies including 6% selection intensity based on GEBV (GS), 12% selection intensity on GEBV + 50% selection intensity based on canopy coverage (CCGS), 6% selection intensity on yield (Yield), and 6% random sampling (Random) from the University of Illinois breeding program grown at one location in Illinois during 2022

(3) 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).

Each year, UMN predicted the mean, population variance, and genetic correlations among traits for all possible crosses among UT breeding lines and provided these predictions to breeders. A total of eight participating breeding programs used this information to select crosses. UMN also selected 5 cross combinations that strived to create breeding populations with high protein, high yield and a reduced genetic correlation between these traits. These crosses were made in 2020, and populations were advanced to the progeny row stages in 2023. This fall we will be able to measure protein on these populations.

Additionally, we also leveraged the existing SoyNAM dataset to validate these models for predicting genetic variances and genetic correlations among traits. Using these 39 biparental populations, models to predict genetic variances and genetic correlations between traits for all possible crosses, we validated models by correlating observed parameter values with predicted parameter values. We found that in 17 out of 21 cases, there was a positive correlation between predicted genetic correlations and observed genetic correlations, indicating this methodology holds promise for identifying breeding crosses that could have less detrimental correlations between traits. This manuscript is written and will be submitted in the coming months.

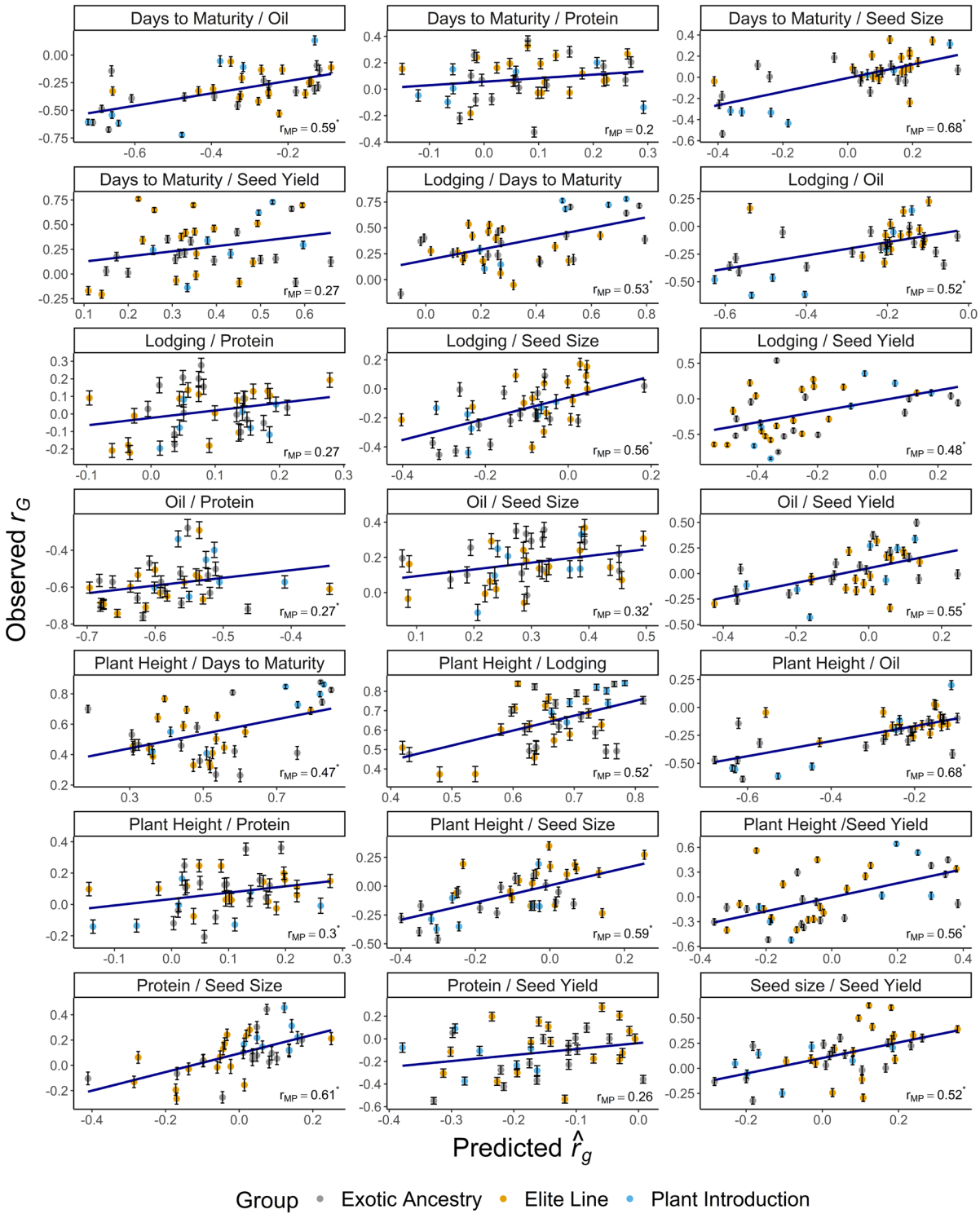

Figure 3. Correlations between predicted genetic correlations and observed genetic correlations in SoyNAM populations.

(4) Advance generation by single seed descent for generated crosses in (5) and perform preliminary yield trials with protein data collected by NIRS on F3 or F4 derived lines in FY23.

These are in progress and will be tested and compared by multiple breeding programs in-field in 2023, with yield data available in 2024.

(6) Perform crosses, genotyping, and line advancement according to rapid cycling breeding scheme.

During the summer of 2020, a Cycle 0 (base population) population of F1 plants was created by random mating 13 parents. The parents were selected for yield potential, genetic diversity and seed composition. Genomic predictions for seed yield, genetic variation, and seed composition were used to select superior F1 plants, and intermate the selected plants to produce a new cycle of F1 plants. This rapid cycling process of selection and intermating was repeated three times to produce Cycle 1, Cycle 2 and Cycle 3 generations. No phenotypic data was collected on these progeny during this process. Creating of Cycle 0 through Cycle 3 was completed in less than two years by growing two generations in the greenhouse in the winter, and two in the field in the summer. Each generation between 100 and 250 new F1 progeny were created and between 20 to 30% of the F1s were selected for intermating. In each cycle, the F1 plants were allowed to self over multiple generations, and inbred populations of random F3 or F4-derived lines were created from each cycle of selection. The inbreeding and seed increase process to complete lines for evaluation was completed in the winter of 2022. In the summer of 2022, 150 to 160 random F3 or F4-derived lines from Cycle 0 through Cycle 3 were evaluated in the field at three locations in KS. In additional to obtaining information on seed yield, maturity, and seed composition (seed protein and oil), each of the 633 genotypes in the trial were genotyped using the Agriplex 1000K SNP array and monitored using remote sensing from about the V2 growth stage until maturity. In 2023, these field evaluations are being repeated.

(7) Generate near isogenic lines varying for putative “yield alleles” previously identified from landscape genomics analyses.

Our previous NCSRP project identified 26 putative yield-related alleles based on a population genetic evaluation of haplotypes under selection in an alternative (from Randy Nelson’s breeding program) and a conventional gene pool. To test the value of these alleles, the OSU breeding program focused on four loci which differ between the breeding line LG09-8165 and LG11-5120. Material transfer agreements were obtained for these lines with complex pedigrees. In FY20, crosses were made. Since then, reselections from heterozygous individuals were carried out to develop F4 derived families which primarily differ only for the yield allele of the targeted loci. Currently, these F4 derived lines are being grown in progeny rows and will be available for preliminary yield tests by collaborators in FY24.

Deliverables

(1) Methods to improve selection of progeny rows based on genomic selection with secondary traits and/or improved spatial statistics.

Although the multi-year data necessary to obtain a definite conclusion on best methods has not yeat been obtained, the participating breeding programs have implemented the procedures and methods necessary to apply these methods and have shared this knowledge with the SOYGEN group. The feasibility of genotyping and making selections on large numbers of early generation materials during the field season can be logistically difficult, requiring the implementation of new field protocols; which have been and will continue to be shared among breeding programs.

(2) Application and limitations established for rapid cycling genomic selection in soybean.

The KSU and UMN worked together to implement and establish methods for a rapid cycling genomic selection program. Results of these upcoming field trials will be used to characterize the effectiveness of rapid cycling to increase genetic gain, and understand the impact of rapid cycling on the phenotype of the progeny and genetic makeup of each cycle of selection. Ultimately, providing data to support a specific number of rounds of rapid cycling based off a given model, for a given population diversity.

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

Key performance indicators

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

Data summaries from the tests were shared with cooperators. Breeding programs have used accessions from this study as parents in their breeding programs. For example, the McHale breeding program has sub-selected lines predicted to have good agronomic traits and yield from a selection of exotic germplasm screened for disease resistance traits.