Objective 1. Continue to develop and enhance genomics-assisted breeding resources and tools to facilitate routine application in public breeding programs.

Objective 1 can be broken down into five sub-objectives for which we can report specific and significant progress during these past six months.

Sub-obj 1: Continue to genotype with genome-wide and trait-targeted markers all new breeding lines entered in the Northern Uniform Soybean Tests

Progress: As reported in the last semi-annual report, 511 new breeding lines submitted to regional trials were grown. Tissue was collected. This past reporting period DNA was extracted from the tissue. We sent DNA of the 511 breeding lines to Dr. David Hyten at UNL for genotyping via skim sequencing. Previously we sent all samples to a private service vendor. However, because of dramatic prices increases, we decided to work with our co-PI Dr. Hyten who could provide us very similar data for 70% of the cost. This delayed our data delivery, but we are confident we will have the data in hand by May. We now have genotyped over 4000 advanced breeding lines entered into the public regional trials, creating an impressive resource helping current and future soybean breeders and geneticists connect genotype to phenotype, and developing genomics-assisted breeding resources.

Data from the 2024 NUST trials was collected this past fall. We formatted the data and sent it to Rex Nelson at Soybase, where it will be uploaded very soon.

A manuscript on this work we have pursued for the last several years has been submitted to the scientific journal Crop Science. It was recently accepted pending revision. We are currently editing the manuscript for final acceptance

1) **Wartha, C.A., B. Campbell, V. Ramasubramanian, L. Nice, ….19 authors….A.J. Lorenz*. 2025. Genomic analysis and predictive modeling in the Northern Uniform Soybean Tests. Crop Science (*Accepted pending revision*).**

Sub-obj 2: Enable individual public breeding programs to test and use genomic prediction

Originally, this project explicitly funded the integration of genomic prediction into the public soybean breeding pipelines to expedite yield improvement. Because of budget cuts, the funding for this part of the project was removed. Nevertheless, this project has continued to instigate and enable several public programs to start using genomic prediction routinely. Below are some highlights from reports from individual programs that are part of the SOYGEN initiative.

University of Nebraska

Aiming to generate new recombinant populations with high yield and resistance to biotic stress, the UNL soybean breeding program conducted a genomic selection analysis following the 2024 field trials, utilizing phenotypic datasets from multiple years and locations to train the prediction model. This dataset was formed by UNL lines that belonged to elite populations designed to carry resistance alleles to the Soybean Cyst Nematode (*Heterodera glycines*) for the rhg-1a//rhg-1a, Rhg-2//Rhg-2, and Rhg-4//Rhg-4 genes. These lines have been evaluated in field trials since 2022. In addition, the Northern Uniform Soybean Trials yield datasets from 2012, 2018, 2019, and 2020 were also added to train the model. Lines from both

datasets have been extensively tested in maturity 2 and maturity 3 locations in Nebraska and other surrounding states. UNL lines were genotyped with micro-inversion probes (MIP), and the NUST lines were genotyped with 6K SNP chip. Genotypes were imputed and filtered accordingly. For the analyses, yield values were adjusted for the experimental design model and the best linear unbiased estimations (BLUEs) were used as input for the genomic selection analyses. The genomic selection analyses accounted for the genotype-by-environment (GEI) interaction, considering that complex, non-linear interactions between lines and environments regularly occur in the soybean breeding context. The GS4PB R Shiny App (previously known as SOYGEN2 R Shiny App) and its codes were used to run the analyses.

Seven lines were selected by the breeder following the analyses. These lines were highly ranked based on their GEBV from the genomic prediction analyses. In addition, these lines contain at least one allele of interest for the three mentioned genes, and some of these lines are homozygous for two loci of significant interest (rhg-1a//rhg-1a, Rhg-4//Rhg-4).

These seven lines were selected as parents for eight new combinations in the UNL Winter Nursery Crossing Block project, conducted in Puerto Rico between January 2025 and April 2025. It is expected that F1 plants of these seven populations will be planted and genotyped in Lincoln, NE, in June 2025. As future directions, these seven new populations will be advanced, and their selected progenies will be tested in multi-environment field trials to identify superior lines for yield and resistance to Soybean Cyst Nematode. Additionally, using SOYGEN2 yield datasets in both regular and sparse genomic selection designs will enable the UNL Soybean Breeding Program to efficiently select superior lines.

University of Missouri

Andrew Scaboo's lab is diving into the data we collected as part of SOYGEN2 in the genomic selection experiment. This experiment tested genomic prediction versus phenotypic selection versus random selection at four universities: University of Minnesota, North Dakota State University, University of Illinois, and University of Missouri. The selection treatments we applied in the originally designed experiment were not as successful as we had hoped. Currently, we are thoroughly analyzing the data to figure out why the genomic selection treatment was not as successful as we had anticipated, and how we can better understand and utilize it in the future. Because this multi-institutional dataset is very large and complex, we are first starting to develop the analysis framework and treatments using the Missouri data only.

Several initial analyses were described in the last progress report. During this last reporting period, we extensively evaluated the effect of genotype imputation. In the below figure, it can be seen that methods of genotype imputation implemented in the "GS4PB" application developed with NCSRP funding improves prediction accuracy overall. This indicates that these methods will be powerful approaches towards improving the cost effectiveness of genomic prediction for driving genetic gain in yield.

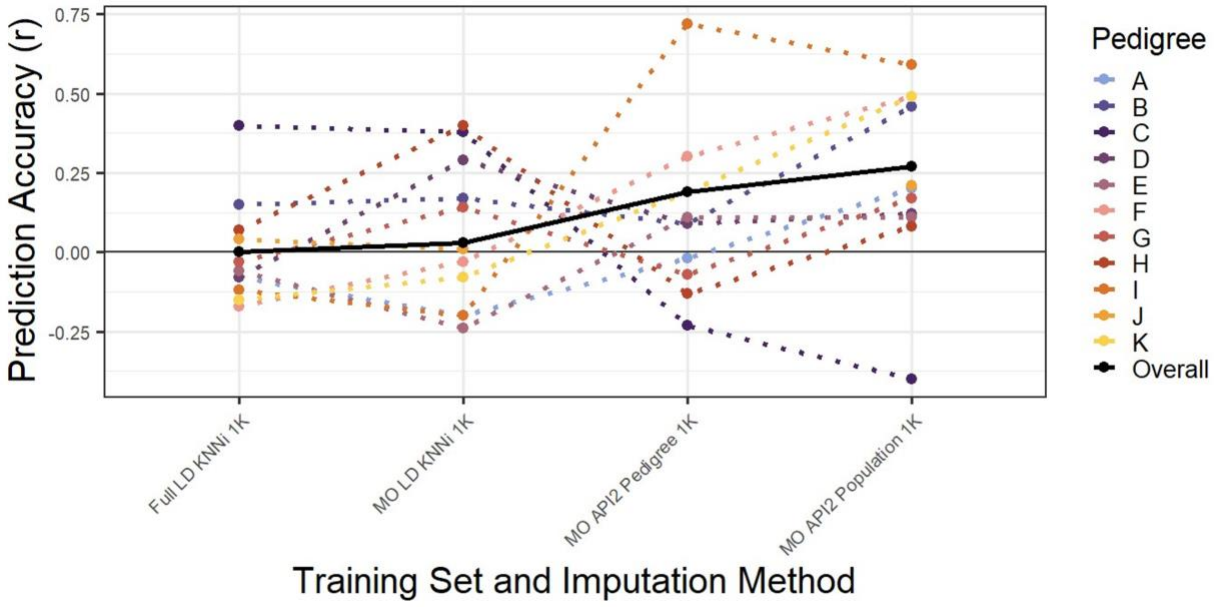

Figure 1. Various methods of marker imputation to increase prediction accuracy. Marker imputation uses low density molecular markers, combined with a training genotyped with high density markers, to project in silico markers onto breeding progenies with no additional cost. The "API2" methods displayed on the right side of the graph are the most computationally challenging, which are now implemented in the GS4PB application developed as part of the SOYGEN project. This improves their accessibility to practicing plant breeders.

University of Minnesota

As described in the last report, the UMN soybean breeding program has refined its GS pipeline and tested it extensively on the UMN Preliminary Yield Trials (PYT) data. PYT 2023 progeny population lines were assayed using 1K low-density (LD) genotyping assay and parents of PYT23 lines from the crossing block were assayed using a low-pass sequencing platform to generate high density (HD) variant data. The 50K SoySNP Chip subset from the HD data set as the parental reference panel to impute 1K LD set to 50K HD set (~30K SNPs after QC). We used this imputed data to make genomic predictions using genomic prediction models that include GxE interaction effects. In the summer of 2024 we planted a trial including lines selected using genomic prediction and phenotypic selection. The trial was successfully planted at six locations in Minnesota. Every location yielded good data recently collected during harvest.

During the past reporting period, we analyzed the data to compare the phenotypic selection versus the genomic selections. For each selection treatment, we selected both high yielding lines and low yielding lines. Selecting low yielding lines is actually important as it tests our ability to identify the poorly yielding lines. This ensures that we are not devoting precious field phenotyping resources towards lines that are predicted to have low yield.

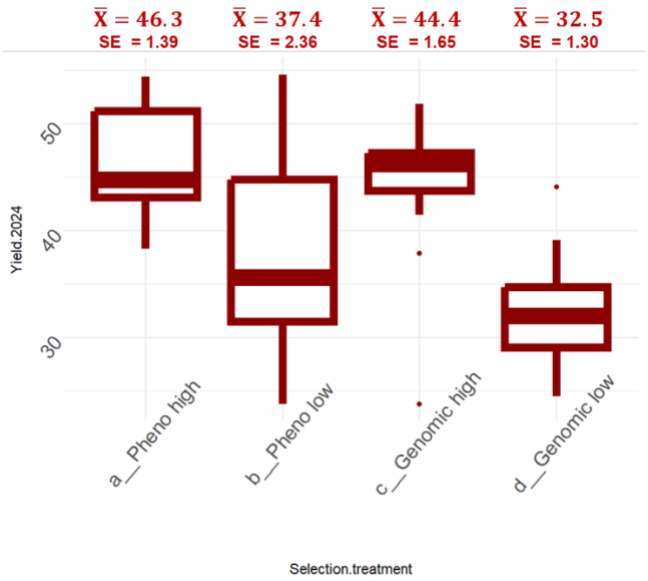

Figure 2. Yield performance of groups of breeding lines selected using either phenotypic selection or genomic selection. Lines with both high yield and low yield were selected. The groups of lines were tested the following year for their yield performance at three locations. The lines selected by genomic selection performed just as well as those selected by phenotypic selection. Moreover, the lines selected for low yield were lower yielding in the genomic selection treatment. These results suggest that genomic selection can replace phenotypic selection in the preliminary yield testing phases of a breeding program.

Sub-obj 3: Development of a genomic prediction R-Shiny app for easy implementation of GS for breeders.

The first version of this application is now complete. We recently submitted a peer-reviewed journal article to the journal Plant Genome.

1) **Ramasubramanian, V., C. Wartha, L. Singh, P. Vitale, S. Ru, and A.J. Lorenz*. 2025. GS4PB: An R/Shiny Application to Facilitate a Genomic Selection Pipeline for Plant Breeding. Plant Genome (*Submitted*).**

This manuscript describes the development of the application, the components, how it can be used to execute genomic selection for plant breeding, and how it can be accessed. This application is freely available to the public, and will enable plant breeders to implement genomic selection.

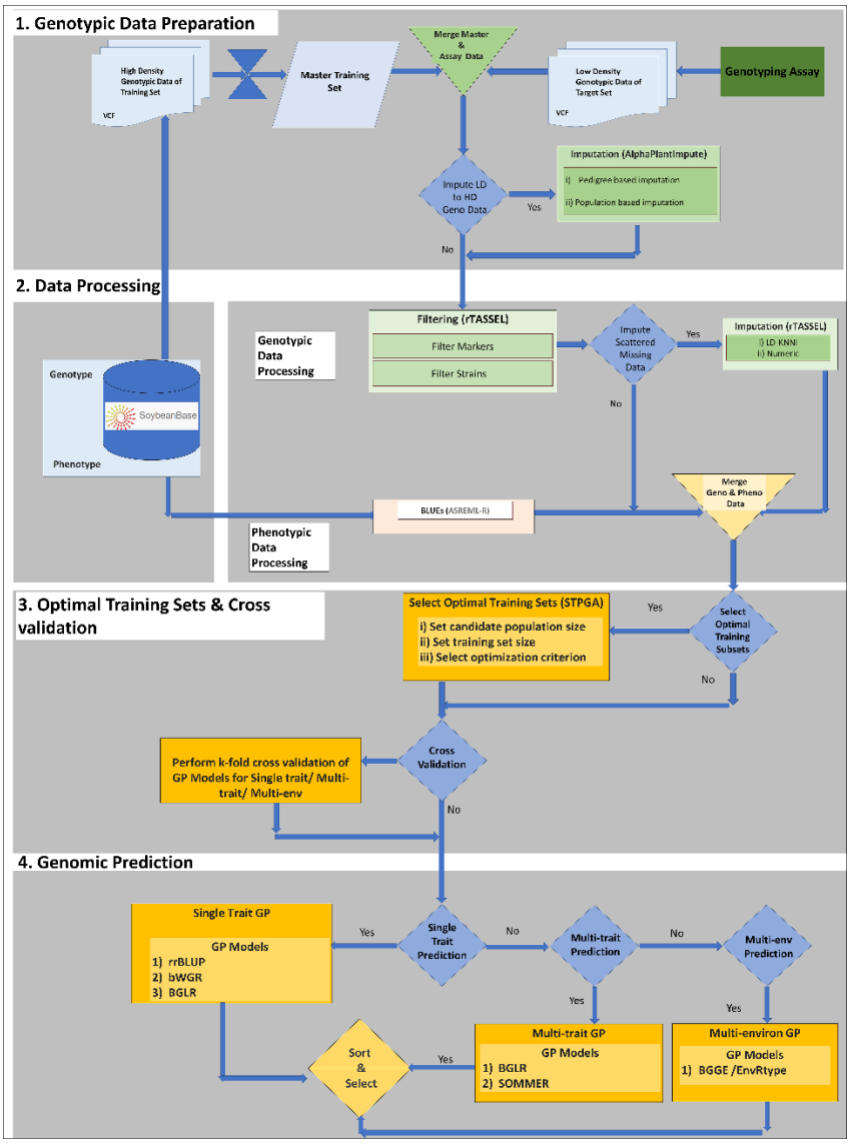

Figure 3. Schematic of a genomic selection pipeline implemented in the GS4PB shiny app. The pipeline is split into 4 steps. Genotypic data preparation (Panel 1): High-density genotypic data from one or several programs or historical populations is retrieved from databases and compiled into a master training set and filtered to create a high-quality training dataset. The target population assayed using low-density genotypic data is combined with high-density genotypic data from parental lines or a reference population for imputation from a low-density to a high-density marker set using the AlphaPlantImpute program. The imputed data is then filtered to create high-quality target data, which is then merged with the training data. Data processing (Panel 2): The data processing step is split into two tracks: a) genotypic data processing, where the merged data from step 1 is filtered for minimum missing frequency for markers, strains and minor-allele frequency. b) phenotypic data processing where best linear unbiased estimators (BLUEs) are calculated externally. The processed genotype and phenotypic data are then merged to create a combined data table that is used further. Optimal training set selection and cross validation (Panel 3): An optional training population optimization step is implemented using the STPGA package, followed by a cross-validation routine whether the training population optimization step

was elected or not. Genomic prediction (Panel 4): Genomic predictions are calculated according to a variety of single trait, multi-trait, and multi-environment models selected by the user. Predictions and reliability of predictions are output.

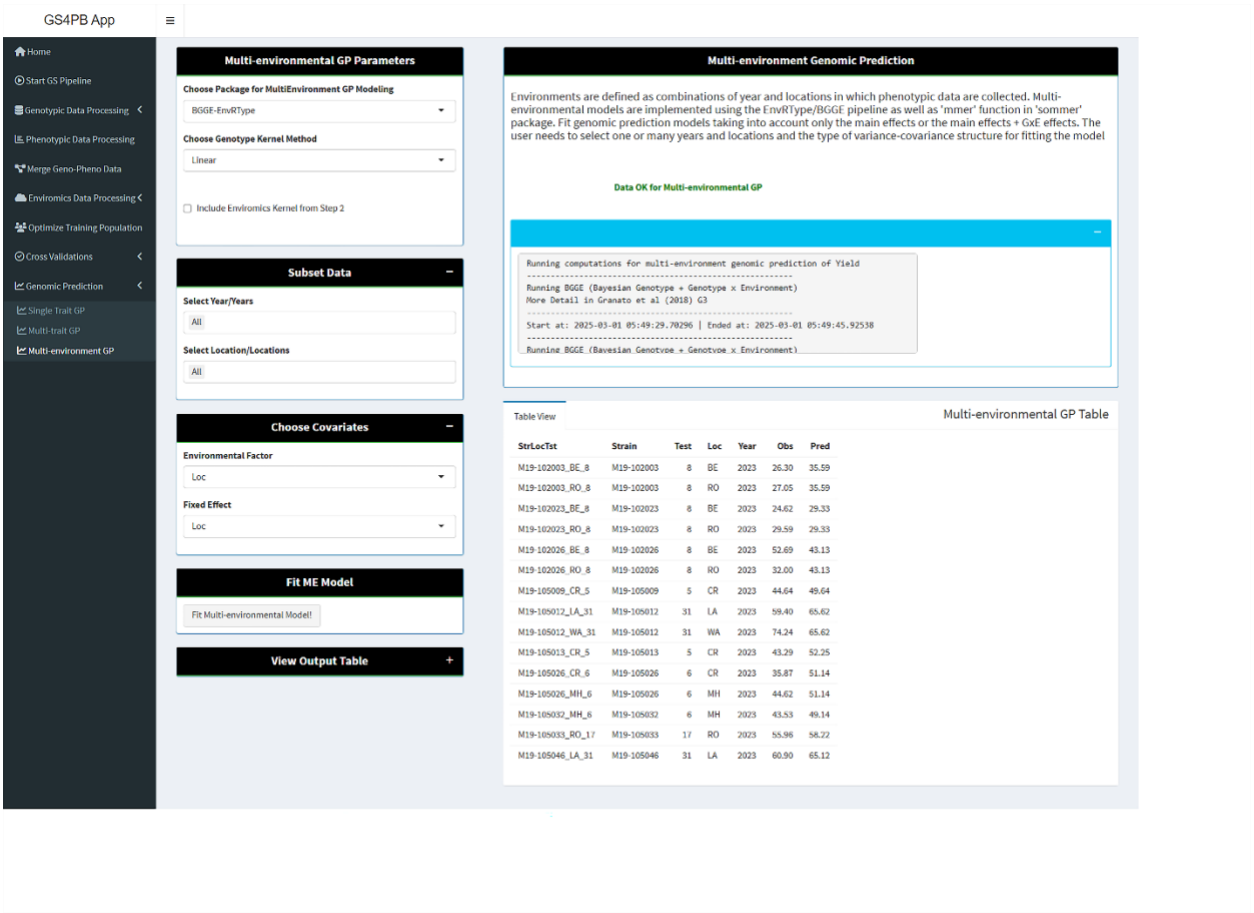

Figure 4. Screenshots of GS4PB multi-environmental genomic prediction (bottom panel). In the ME-GP panel, the user can select years, locations and set covariates such as the data column corresponding to the environmental factor, and fixed effect. In the current implementation, only BGGE-EnvRtype package is selected. The user has the option to select a linear or Gaussian kernel method for estimating relationship between strains. Once the GP model is fit, a table containing the observed and predicted values for the selected ME model is displayed and the table can also be downloaded from the app.

Sub-obj 4: Adopting and advancing BreedBase for storage of information for soybean genomic prediction.

This database is currently working for this project, so there is nothing new to report here. The next step to make on this database is to ensure all collected data are there and available.

Sub-obj 5: Connect target and training populations using imputation that leverages pedigree relationships and enhance this capacity by inclusion of this method in the software application.

As mentioned in the last report, this sub-objective has been completed. We have implemented these methods in our software application GS4PB. For example, using this software, we have used these methods in the analysis of the Missouri data described under sub-objective 2, as well as the University of Minnesota genomic selection pipeline described under sub-objective 2.

Objective 2. Develop and test methods for predicting cultivar performance in future target environments through genomics-assisted breeding models, phenomics, and environmental characterization.

Three main tasks are associated with this objective.

1. Conducting a large, multi-institutional multi-environment trial to create a "genotype-by-environment interaction" dataset for soybean that current and future soybean researchers can use identify methods to improve genomic prediction methodology for yield.
2. Define environmental variables driving genotype-by-environment interactions that can be used to enhance prediction models.
3. Quantify biomass non-destructively to that crop growth can be evaluated in relation to certain environmental stressors.

Progress on each task is reported in order below:

1. *Conducting a large, multi-institutional multi-environment trial to create a "genotype-by-environment interaction" dataset for soybean that current and future soybean researchers can use identify methods to improve genomic prediction methodology for yield.*

For this objective, we are conducting a multi-environment, multi-institutional coordinated performance trial of 1200 diverse breeding lines. Each breeding line will be phenotyped for several agronomic and phenological traits, and each will be genotyped using low pass re-sequencing technologies. Detailed environmental for each growing location in each year will be collected and analyzed. The ultimate goal is to better predict the interactions between the environment and genotype. If we are successful, we will leverage genomic data, phenotype data, and environmental data to predict how new breeding lines may perform in future environments that a producer is most likely to encounter.

Last summer these breeding lines were successfully grown and phenotyped for yield at 21 locations. Data has been delivered to the University of Minnesota were it has been deposited in a centralized database. Plans were laid for scanning of all samples (12,400 samples) with NIR to measure protein and oil. We worked with Perkin-Elmer to develop a protocol to standardize instruments across universities. Standardization samples were collected across universities to represent diversity in germplasm and growing conditions. From UMN, samples were split and sent out to collaborators for scanning on their instruments to create the standardization file. We anticipate the samples to be scanned for protein and oil content during the months after planting when time allows.

Another major activity regarding this objective was the design and packaging of the 2025 yield trials. All seeds have been delivered to 2025 field locations, and 90% of the packaging is complete. Fields have and designed and sent to cooperators. All is on schedule for a successful 2025 planting once weather allows.

Genotype data from the Hyten lab was collected and returned on April 11. The genotype data includes 15 million molecular markers (SNPs) imputed from skim sequencing data, with a 50K subset available. The SNPs were mapped to the latest Williams82 genome version, Wm82a6v1. Analysis of these data will proceed over the coming months.

2. *Define environmental variables driving genotype-by-environment interactions that can be used to enhance prediction models.*

We completed the development of the Seasonal Characterization Engine (SCE). The SCE is a specialized tool integrated within the R environment, designed to streamline the analysis of agricultural trial data using the Agricultural Production Systems Simulator (APSIM) model. Users begin by uploading trial data formatted according to guidelines specifying key parameters such as location, latitude, longitude, and maturity. Predefined input files, such as the soybean seed composition test, facilitate accurate data setup.

After data upload, users select a crop model template (e.g., soybean or maize) and specify maturity handling methods. Users then choose appropriate weather and soil databases, considering geographic coverage to avoid analysis errors.

The APSIM model generates environmental variables specific to different crop growing stages, providing a detailed characterization of conditions throughout the growing season. Upon initiating the analysis, the SCE provides real-time process updates via the R console. Analytical results become available through multiple visualization tools:

1. **Results Viewer:** Offers box plot visualizations of selected variables, downloadable individually or collectively.
2. **Heat Maps:** Enables detailed inspection of seasonal and environmental covariates by growth stage, comparing specific parameters within consistent genetic maturity groups.
3. **Trial Similarities:** Presents correlation heat maps illustrating seasonal profile similarities among trial sites, complemented by dendrograms to clarify site relationships. Outputs are readily exportable for further analysis.
4. **Thermal Time and Precipitation:** Summarizes accumulated thermal time and precipitation over growing seasons, offering comparative analyses between and within sites across multiple years. These insights support understanding climate variability and trends.

3. *Quantify biomass non-destructively to that crop growth can be evaluated in relation to certain environmental stressors.*

During the 2024 season, Rainey Lab at Purdue University has focused on phenotyping soybean crops using high-resolution UAS imagery from 10 SOYGEN sites to extract canopy traits. A data management system has been set up at Purdue to help SOYGEN collaborators with secure storage of UAS image data, efficient data sharing, and access to processed plot-level outputs. To date, key canopy traits including spectral, textural, and structural features, have been extracted from 6,320 two-row plots covering about 1,200 genotypes across V3 to R5 growth stages. The processed plot-level data includes labeled and curated RGB image clips, binary masks, and 3D point cloud data. This growing dataset provides valuable resources for soybean breeders, supporting the development of advanced AI/ML algorithms and trait derivation.

For biomass prediction, we established 'Calibration Plots' — 8-row plots planted adjacent to the SOYGEN plots to enable non-destructive biomass sampling. A representative subset of 36 lines from the SOYGEN3 GEI panel, with three replications, was used in this experiment. Ground biomass sampling was performed on rows 2, 3, and 4, while rows 6 and 7 were used for UAS-based data collection and harvest yield measurements. UAS flights were conducted within one day before or after each ground sampling event, and from these flights, we extracted several key drone-derived metrics, including canopy coverage, canopy height, green leaf index (GLI), and an array of structural and textural features.

In addition to the calibration dataset, we incorporated our previous 'Public Biomass' experiments (2021–2023), which utilized public soybean breeding germplasm from the north-central U.S. region.

Biomass prediction was approached using two modeling strategies:

1. A simplified model using two robust drone-derived predictors — canopy coverage and canopy height — both known to exhibit strong correlations with biomass.

2. A comprehensive model leveraging a broader suite of structural and textural features, capturing more detailed aspects of plant architecture and canopy structure.

Machine learning models developed with these approaches achieved high performance, with R^2 values reaching 0.89 and RMSE as low as 68.70, demonstrating strong predictive accuracy. Our current focus is on refining these models by minimizing time-dependent biases, enabling reliable biomass predictions on any given day regardless of growth stage or sampling date.

Structural features derived from 3D models, like canopy volume, and the top and side geometry of the canopy, are also good indicators for biomass. We anticipate that the marker data will facilitate the development of a biomass prediction model. We expanded the ground truth biomass dataset in 2024, and plan to do so again in 2025.

Preliminary analysis has been conducted on yield estimation using image-derived features. The results show promising potential for predicting yield, with an R^2 value around 0.45. To improve prediction accuracy, efforts will focus on expanding the dataset and incorporating additional data such as weather variables, soil characteristics, and genetic marker information. The integration of expanded data sources and advanced modeling techniques is anticipated to significantly strengthen the robustness and predictive power of the models.

Status of 2024 season UAS data collection and processing

S. N.	Collaborator	SOYGEN Site	Planting date	Number of UAS flights	GSD cm/pixel	Metrics	Ground Truth data
1	Aaron Lorenz	StPaul	05/13/2024	6	0.94 – 1	CC, GLI, H, Texture, Structure	Yield, R1, R8, Lodging
2	Asheesh Singh	JuliousCAD	05/19/2024	4	0.32		
3	Eliana Monteverde	Burwash	05/23/2024	4	1.44		Yield, R1, R8, Lodging
4	Carrie Miranda	Casselton	06/08/2024	7	0.84		Yield, R1, R8, Lodging
		Colfax	06/06/2024	6	0.85		
5	Katy Martin Rainey	ACRE	05/12/2024	11	0.39 – 1.84		Yield, R1, R8, Lodging, Height, Biomass
		Remington	05/18/2024	5			
		Validation		10			
		Calibration B		12			
6	William Schapaugh	Ashland Bottoms	06/10/2024	6	0.50		Yield, R1, R8, Lodging
7	Dechun Wang	MSU Agronomy Farm	05/30/2024	1	0.74		
		Ingham County	05/21/2024	1	0.72		

Objective 3. Discover structural variants (SVs) and test whether modelling structural variants improves genomic predictions for yield and seed composition.

Recent improvements in our SV calling pipeline have significantly enhanced the accuracy and resolution of SV detection. By integrating five distinct SV callers, we are now able to identify a greater number and diversity of structural variants with higher confidence. These improvements were implemented using the *Williams 82 version 6* (a6) reference genome, a more complete and accurate assembly with no scaffolds compared to its predecessors. GWAS was conducted using both the SVs and SNPs identified from the a6 reference genome, yielding highly significant associations between structural variants and key agronomic traits. These traits include, but are not limited to, yield, stress resistance, and plant architecture. The use of a refined SV dataset has reduced background noise and increased the power to detect true associations. Gene Ontology (GO) analysis of significant GWAS markers located within gene models revealed that many of the associated SVs and SNPs influence genes known to control trait development and physiological processes. These functional annotations provide valuable biological insights into how structural variation contributes to phenotypic diversity in soybeans.

Comparative analyses demonstrate that the use of the Williams 82 a6 reference genome produces more meaningful GWAS results than the earlier version 4 (a4). GWAS performed with a4, which contains more unresolved scaffolds, tends to yield less consistent and potentially spurious associations. This underscores the importance of reference genome quality in structural variation-based GWAS.

Our findings highlight the critical role of high-quality reference genomes and comprehensive SV calling pipelines in conducting accurate and biologically relevant GWAS. The integration of multiple SV callers, coupled with the Williams 82 a6 reference, has led to improved detection of trait-associated structural variants. These results demonstrate that unresolved genomic regions can compromise GWAS outcomes, potentially leading to random or misleading associations.

Figure 5. Comparison of marker-trait associations between different versions of different genomes. The a4 genome is an older genome, while the a6 genome is a newer genome version. It can be seen that the marker-trait associations (dots above horizontal green line) are stronger using a6 (lower righthand corner). They are particularly strong for structural variants. This potentially means these structural variants are important for controlling yield, and finding them and understanding could benefit yield improvement in the future. Moreover, genomic prediction models incorporating structural variants may be more predictive than those models only using SNP information.