Project report (Third quarter July 1, 2023, to September 30, 2023)

Project funded by North Central Soybean Research Program

Project title - Field phenotyping using machine learning tools integrated with genetic mapping to address heat and drought induced flower abortion in soybean

Participating institutions – Texas Tech University, Kansas State University, University of Missouri, and University of Tennessee

Goals & Objectives

Long-term Goal – Develop soybean cultivars with 20 to 30% lower flower abortion under favorable to challenging environmental conditions, leading to about 10-15% increase in yield potential

Objectives (Year 1)

• Explore the genetic diversity in flower abortion under different soil moisture and climatic conditions using a large diversity panel

• Develop an image-based field phenotyping system and deep-learning tool to precisely document temporal dynamics in flower abortion and pod retention in genetically diverse soybeans

• Discover environmentally stable and region-specific genomic regions controlling flower abortion in diverse soil types, moisture, and climatic conditions

Progress achieved

Objective 1 - Explore the genetic diversity in flower abortion under different soil moisture and climatic conditions using a large diversity panel.

Texas Tech University

In early July, the soybean plants of all 228 lines were at the V2 developmental stage. In preparation for imaging and flower counting, several steps were taken. Labels were added to the plots, and measures were implemented for weed control. Additionally, cameras were mounted on the tractor for testing purposes (Figure 1). As the month progressed, the plants transitioned to the reproductive stage, which was delayed and occurred around the 20th of July due to stressful conditions i.e., Lubbock had over 5 weeks of 100 plus ^{o}F with no rain. Once they started to flower, manual flower counting and imaging was started. Various aspects, including camera angles, lens types, and camera numbers, were systematically adjusted and tested on the tractor to determine the optimal position and speed for imaging. Also, some pictures to document the diversity of the genotypes were taken.

Figure 1. Tractor being tested in various positions to enhance the imaging quality.

To date, we have completed the 13th round of flower and pod counts for the 228 diverse genotypes. These counts were conducted every 4 to 5 days in conjunction with the imaging

process. Most genotypes have now completed their flowering stage and have reached the R7 stage of development. As we approach the harvesting phase, a final round of imaging will be conducted on the dried plants for developing and counting pods using machine learning models. Subsequently, each plot will be harvested manually (3 feet per genotype per rep). To ensure representative data, we will select the tagged plant used for flower counting, along with an additional four plants, for other yield related parameters. From these five plants we will gather information on plant height, the number of branches, internode size, pod count, seeds per pod, 1000-seed weight, seed size, and grain weight per plant. Data from the tagged plant will be used as ground truth data for validating machine learning models for flower and pod numbers. Plants from 3-foot row length will be used for yield determination. Lodging scores will be recorded at harvest. Figure 2 displays the current flower count progress at TTU, indicating the range in maximum flower counts in 228 lines. Some of the very low numbers could be a result of rabbit damage.

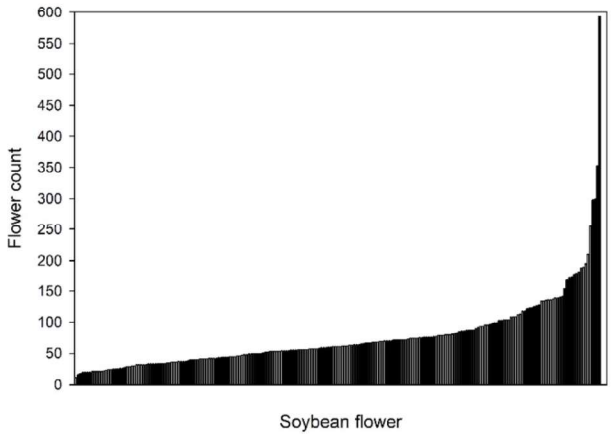

Figure 2. Maximum flower count of the 228 Soybean lines.

July 2023, Dr Jagadish aired a radio interview of the Dakota Farm Talk to highlight the project and indicated the benefits that the progress made will have on the US and global soybean industry.

October 29th, 2023, Dr. Espíndola will deliver an oral presentation titled "Advancing Phenotyping for Flower Abortion in Soybeans through Image Analysis and Machine Learning" at the 2023 annual meeting of ASA-CSSA-SSSA in St Louis, Missouri.

University of Missouri

Since it was highly challenging to obtain human help to physically count flowers on all 228 lines, all other participating locations, selected a core set of 30 lines based on genetic diversity for manual flower counting. To date, we have completed the 7th round of flower counting for this core set of 30 lines in three replications (90 plots). These counts were conducted every 3 to 4 days in conjunction with the video imaging. Flower numbers are relatively consistent across 3 replications of each genotype and significant differences in flower number were observed among different genotypes. All genotypes have finished flowering and currently reached the R5 to R7 stage. As we approach the harvesting phase, a final round of imaging will be conducted on the dried plants for pod counting together with manual pod counting. Subsequently, each plot of 228 lines (684 plots) will be harvested manually (two center-rows of 8 feet/row) to estimate seed yield. Seed harvest will start at the end of September and the harvested seeds will be used for next year planting at all the locations. There are about 10 lines that may not have enough seeds, which will be included as a part of our winter nursery for seed increase.

University of Tennessee

At the University of Tennessee soybean plants imaging system was facilitated using GoPro Hero11 cameras mounted on a Traxxas Hoss ® 4x4 VXL conveyor (Figure 3). GoPro cameras were set up based on camera parameters including FPS, image ratio, boost, high quality video mode, white balance, camera angles, lens types, and positioning (10-12 inches distance). Field phenotyping was carried out throughout the flowering period wherein flowers and pods were manually counted separately every 4 to 5 days. Labeling was done in all 690 experimental plots. Plants were identified and tagged in each plot in order to facilitate manual counting of flowers and pods and imaging which was initiated on July 27, 2023. An overhead shot was taken to show the entire field including all of the 228 genotypes using an UAS platform. Remarkable differences in foliage color were observed among the genotypes (Figure 4).

When most of the genotypes completed their flowering stage and reached the R7 developmental stage (i.e., beginning maturity), we recorded another set of imaging, this time for the soybean pods for the 30 core lines (complete defoliation in a considerable number of lines). For those plots with plants reaching R8 (full maturity), harvesting has already been initiated (Figure 5). Plants were harvested manually from the two-row plots within 10.8 ft^2 (~1 m^2) to calculate the final yield. For documenting the yield components and other morphological parameters, the tagged plant that was used for manual counting of flowers and pods during the season along with other 4 plants on the same row were sampled. The harvested soybean plants are threshed using the USDA single plant thresher then the collected seeds per plot were placed inside a labeled bag for quantifying yield. Plant growth stage per soybean line were regularly monitored to estimate the harvest time. Plant height, number of branches, number of pods, seed number per plant, 100-seed weight, total seed weight per plot (1m^2), and final yield will be determined. Lodging scores were also recorded once during late August and will be recorded per soybean line at harvest.

We were able to release a podcast on the UTIAg website (available on Spotify for Podcasters) about our soybean flowers abortion project. Find the link here: https://podcasters.spotify.com/pod/show/utiag/episodes/Culture--Agriculture-Ep--4-Research-Could-Improve-Soybean-Yield-e277htq/a-aa5c7ku

Furthermore, we worked with the UTIA communication team to create and release a video about our current research project. To see the video, click here: https://www.youtube.com/watch?v=H5CVeWbiliU

The video will be broadcasted via WBBJ and Nashville TV channels during late September 2023.

Finally, we put a research abstract together titled "Image-based field high throughput phenotyping for quantifying flower abortion in genetically diverse soybean germplasm" and submitted it to the 2023 ASA-CSSA-SSA annual meeting. It will be presented at the CSSA section during the meeting in late October/early November 2023.

Figure 3. GoPro Hero11 cameras mounted on a Traxxas Hoss ® 4x4 VXL conveyor (left) the imaging technique for soybean flowers in between rows using our conveyor (right) (at WTREC)

Figure 4. Overhead shot of 230 soybean genotypes at the University of Tennessee's West Tennessee Research and Education Center (WTREC).

Figure 5. Soybean plants at R8 (full maturity) (left). Plants were scored for lodging at harvest (right)

Objective 2 - Develop an image-based field phenotyping system and deep-learning tool to precisely document temporal dynamics in flower abortion and pod retention in genetically diverse soybeans.

Texas Tech University

In pursuit of Objective #2, Texas Tech has been working on dataset preparation for the flower detection model and implementing an algorithm for flower counting. These are essential to the foundation for the successful development of our phenotyping system.

1. Dataset Preparation for Flower Detection Model Development:

1.1 Automated Video Frame Extraction

One of our initial tasks was to develop an automated pipeline to extract unique frames from videos captured at various locations. This pipeline streamlined the data collection process and ensured a consistent dataset for analysis.

1.2 Dataset Compilation and Annotation

We compiled a new dataset consisting of ***1314 images from four diverse locations***, namely Missouri, Tennessee, Texas, and Kansas. Collaborating with annotation teams from Texas Tech and Kansas State University, these images were annotated for flower detection. Before these annotated images could be used for model development, they need to be validated by a domain expert which involves confirming, removing, or adding annotations, enhancing the dataset's quality and consistency.

To date, 1341 images (1037 from the previous dataset and 277 from the new dataset) have been validated by Dr. Espíndola, our domain expert and the post-doctoral fellow on the project, encompassing 9367 confirmed flower annotations.

Furthermore, an additional ***1037 annotated images are currently undergoing validation***. This expansion aims to increase dataset diversity and enable the development of better-generalized models.

2. Dataset Preparation for Flower Detection Model Development:

Accurate flower counting in captured videos is essential for Objective #2. To achieve this, we focused on tracking detected flowers across frames to prevent overcounting.

2.1 Implementation and Modification of Tracking Algorithms

We implemented and modified three state-of-the-art multi-object tracking algorithms: SORT, OC-SORT, and OC-SORT with Byte. These algorithms were selected for evaluation in the context of flower counting.

2.2 Annotation of Tracking Data

Evaluating these algorithms required annotating a series of consecutive frames in a video. This process was challenging and time-consuming, as it necessitated identifying and tracking flowers through frames, even in the presence of occlusions. We annotated 211 consecutive

frames from a Kansas field video, encompassing a total of 35 flowers. This segment was chosen for its complexity, involving both long-term and short-term occlusions.

2.3 Algorithm Evaluation

We evaluated the three tracking algorithms with various parameter combinations. Surprisingly, our findings indicate that the choice of algorithm is not the critical factor for accurate tracking and counting of flowers. All three algorithms yielded accurate results when specific parameter settings were used.

To validate and consolidate our conclusions, further videos need to be annotated for tracking and subsequently used for evaluation.

Kansas State University

In this phase of the project, we made a significant shift in our approach to flower detection. Specifically, we switched from models for node detection followed by flower/pod detection to a single model that performs flower detection directly on full images or frames extracted from videos taken in the greenhouse. This shift was motivated by a preliminary exploration of the flower model on full greenhouse images, which showed that the model was capable of detecting flowers directly in those images. Furthermore, by training a flower detection model without the need for node detection, we aimed to streamline our workflow and improve efficiency.

To train an accurate model on full images, we labeled flowers in a set of 1200 images/frames based on guidelines from the TTU domain experts. The images that we labeled were taken by the K-State team in the greenhouse in the beginning of the flowering season, and exhibited many buds and small flowers. We used 800 labeled images for training a new model, 300 images for development and 100 images for testing. Some examples of images predicted by the model are shown below. As can be seen, the model can accurately detect flowers directly in the greenhouse images, without the need to detect and extract the nodes in the first place.

Figure 6. Images from the field used for model detection.

While the model worked well on images and frames from videos taken in the greenhouse early in the flowering season, we encountered challenges when attempting to apply the model to field images (Figure 6). The model's performance suffered because the flowers presented significant

differences in their characteristics (including color, shape and texture), as compared with the flowers in the greenhouse images. To account for such differences, we needed to enhance the labeled dataset by incorporating additional frames from videos taken at various flowering stages, which captured a large variety of flowers. A total of 1200 new frames sampled from videos from all four institutions were annotated and added to the original dataset. The original model was fine-tuned with the additional images and showed good performance overall in our testing as can be seen below. However, the performance on blurry frames and frames from videos taken at higher speed can still be improved.

Annotated Predicted

Annotated Predicted

As we are moving towards annotating pods in the next quarter of the project, and we may also need to annotate more flowers, we have also started to explore the use of large pre-trained foundation models, such as the recent Segment Anything model, to annotate images in a zero-shot setting with human-in-the-loop to improve its annotations. We have also worked on a script to identify differences between ground truth bounding boxes and predicted bounding boxes, with the goal of identifying mistakes in the human annotations as well as identifying challenging images that can help improve the robustness of the model.

University of Missouri

We used GoPro cameras to take 3 rounds of videos of the selected core set of 30 lines. There are some challenges, including steady walking speed of the camera, shade of soybean branches and leaves, and plant lodging issues. Group is designing a uniform imaging platform based on the experience of this year field studies, aiming to unify walking speed and avoid shade from leaves. Meanwhile, we are taking notes on lodging score and maturity date for the whole set of 228 lines, which will be used to select upright genotypes (low lodging) with similar maturity dates for our next year field studies. Initial observation indicated that maturity group (MG) III lines showed significant lodging issues compared to the MG IV lines. We will discuss with the group to focus on the set of lines that had minimal lodging across locations. Thus, we can solve the other 2 major issues caused by plant lodging and different flowering peak time.

Objective 3 - Discover environmentally stable and region-specific genomic regions controlling flower abortion in diverse soil types, moisture, and climatic conditions.

In our prior analysis, we meticulously selected six pivotal genes that are well-documented in their roles pertaining to the initiation of the abscission zone (AZ), the facilitation of AZ development through ethylene signaling, the activation of tissue separation mechanisms, and the subsequent deposition of protective layers following organ detachment from the plant. Furthermore, we incorporated genes known to be involved in soybean maturity and flowering processes. To perform a robust gene-based clustering analysis and to identify alleles with substantial effects, we leveraged a state-of-the-art gene-haplotype analysis framework, which was executed on the high-performance computing servers at TTU. In a preliminary study, we executed the gene-based haplotype analysis on a cohort comprising 481 lines as a means of testing our analytical pipeline using major flowering genes. During this analysis, we successfully pinpointed four significant haploblocks, with particular emphasis on haploblocks H1 and H4 (as illustrated in Figure 7), which exhibited pronounced allelic variations possessing substantial effects on the observed traits. While the haplotype analysis of additional genes remains an ongoing endeavor, our ultimate objective is to compare the lines that overlap with these haploblocks to field data especially flower number and aborted flowers. Following the data collection from all locations will overlay the phenotypic data with haplotype analysis to identify most diverse accession for further analysis.

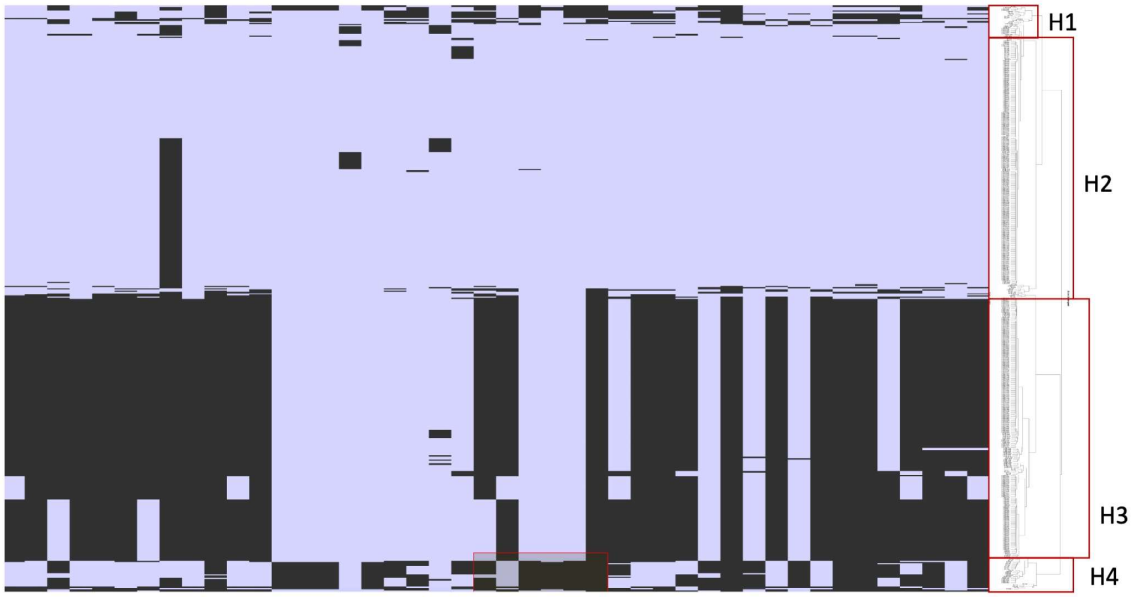

Figure 7. Identification of major haplotypes associated with flowering related traits using whole genome resequencing data. SNPs were positioned relative to the genomic position in the genome version W82.a2. The SNPs in black background are different to the reference genome ('Williams 82').