October 1, 2019

SOYGEN: INCREASING THE RATE OF GENETIC GAIN FOR YIELD IN SOYBEAN BREEDING PROGRAMS

OBJECTIVE 1: Increasing selection intensity and decreasing non-genetic sources of variability through improved progeny row testing

Currently, our preliminary yield tests based on selections from 2018 are ongoing or just being harvested. This will be the second preliminary yield test based on the progeny row selection models implemented in this study.

1.e. Deliverables.

- Implementation of selection models unique to each breeders' needs.

1.f. Key Performance Indicators or performance measures (year 3).

- Additional data were collected on progeny rows in 10 programs.All programs reporting have collected additional phenotypic data for selection on progeny rows for two years.
- Selections completed before harvest for programs electing onto to use yield data.
- Preliminary yield trials organized by each breeder to test selection accuracy for the 10 breeding programs is ongoing for the 2019 season but was achieved in 2018.
- A selection accuracy assessed from 3 years of data, and manuscript and reports prepared.
    - This KPI is pending completion of 2019 season. A PhD intern will join the Rainey lab for 9 months in November and will be responsible for summarizing this experiment.

OBJECTIVE 2: Increasing selection coefficient and decreasing length of breeding cycle through genomic selection

Part of our efforts this reporting period were in transferring knowledge and data from a postdoc, Ben Campbell, who took a permanent job in industry to a new graduate student, Cleiton Wartha, who is assuming primary responsibility of this project.

We continued to contribute to the UST genomic selection training population by genotyping 342 additional public breeding lines this summer. Additionally, we extracted DNA from 299 lines with the intention of genotyping these once plates are complete. Phenotypic data from 2018 has been compiled, QC'ed and added to the training set. Training models including the 2018 data and new genotype data are being created.

We did extensive testing of multiple genotyping methods, specifically MIP protocol parameters, to try and reduce the amount of time to perform the protocol. Currently the hybridization and extension steps are performed separately and can take over 24 hours. We also tested all these conditions with both CTAB extracted DNA and the nanoparticles DNA extraction method ,as well as various oligo:DNA ratios. Overall the CTAB DNA outperformed the nanoparticle extracted DNA for all conditions tested.

We have started to run the genotyping (MIPs) protocol on a greater number of samples. A total of 770 samples provided from co-PI Lorenz was run with the 1k MIPs protocol and data was sent to Lorenz. We have also achieved multiplexing of 1000 samples in a single sequencing run, which minimizes costs for the sequencing part of the MIPs protocol.

Based on our current implementations, the cost of the MIPs reaction is down to $4.88. This cost does not include the CTAB DNA extraction. The CTAB DNA extraction estimate was increased to $1.43 per reaction due to the need to include the cost of using service equipment for the extraction. Due to the high cost of CTAB we will continue to increase testing of the nanoparticle DNA extraction to try and reach the goal of getting DNA extraction costs below $0.50.

2.e. Deliverables.

- A community resource for genomic prediction consisting of a set of soybean lines that can be used to establish genomic prediction to help expedite genetic gain for yield has been expanded as described above.
- Novel inexpensive and rapid genotyping method developed that can be used for genomic prediction and selection. From the different conditions tested we have been able to reduce the time it takes to prepare MIPs reactions for sequencing by more than half. This will significantly increase throughput of the MIPs method. We have also continued to demonstrate that the lower reaction volumes and enzyme concentrations produce good data helping to get the MIPs reaction to under $5.
- Towards the development of genotype imputation methods, we have Published cluster-flexible structural variant calling workflow using Cortex-var on Github, completed whole genomic SNP and short indel calls via Sentieon Haplotyper algorithm for 481 samples, completed whole genomic structural variant calling via Cortex-var for 481 samples, completed whole genomic structural variant calling via Sentieon DNA-scope algorithm.

2.f. Key Performance Indicators or performance measures (year 3).

- GBS method developed that can genotype 200-1000 markers with less than 10% missing data and greater than 95% accuracy has been achieved and implemented on 770 soybean lines with the MIPs 1k probe set.

OBJECTIVE 3: Increasing additive genetic variance

The group is currently growing the second year yield tests of our confirmation population from the USDA soybean germplasm collection. In addition to yield and agronomic data, multi-sensor data was collected from field sites when available. Due to the unique weather conditions of 2019, of the 16 locations grown this year, at least four were unable to be planted or were considered un-harvestable. The confirmation population was selected based on genomic estimated breeding values derived from the sampled individuals (training population) from the USDA soybean germplasm collection. We have performed genome-wide association analyses and subsequently identified QTL for yield, plant height, maturity, seed weight and lodging data using the set of 500 accessions (training population) for which we had previously collected data.

Complex and simple F2 or backcross lines derived from crosses with wild soybean (*Glycine soja*) or *G. tomentella* have yields comparable to checks over 3 years; however, the genomic sequence analysis of *G. tomentella* derived lines has failed to yield evidence of *G. tomentella* introgression.

We have examined genetic variation at the gene locus (*GmBa1*) controlling branching angle and canopy coverage in ~ 800 lines with available re-sequencing data. Our data indicate that the causal mutation occurred in the promoter region. Construct overexpression *GmBa1* has been made and are still in the process of validation of the gene’s function.

We have examined the genetic variation of the major maturity genes in the ~800 lines. The information has helped to confirm causal mutations and is useful for design of individual maturity gene-specific markers for more accurate marker-assisted selection.

We have investigated the distribution of a reciprocal chromosomal translocation in representative *G. max* and *G. soja* populations. Our data suggest that such a translocation played a major role in isolating of domesticated soybean from the wild relatives.

Through genomic analyses of convention and alternative germplasm pools at ancestral and elite stages of breeding development, we have identified and tracked across populations the genomic changes caused by selection for yield in alternative and conventional gene pool.

3.e. Deliverables.

- **Identified yield-marker genotype relationships based on association mapping results from the extensive, high-quality yield dataset.** We performed GWA analysis on the complete set of data for Yield, Plant Height, Maturity (days after planting), seed weight, and lodging. The seed composition data set is being finalized and will be analyzed shortly. In general, the entire set of 500 accessions identified QTL better than any single sampling method. There may be a unique QTL for yield that was identified in the CLU sample that did not show up in the other samples or overall.
- **Developed predictive model(s) that allow selection of superior high-yield genotypes from the USDA germplasm collection**.
  Our results show that the SSD sampling method more effectively reflects the total genetic variation in the USDA Soybean Germplasm Collection that do the Random or Cluster sampling groups. Furthermore, the variance for most traits measured is nearly double in the SSD sample vs. the other sampling groups. For genomic prediction models, cross validation results showed that the SSD sampling group performed better than RAN and CLU for predicting yield, and at least as well as when the entire set of 500 accessions was used for prediction. These results support our hypothesis and verification that the SSD sampling method using genotype information is an efficient and effective way to sample the germplasm collection. This has important implications for future use of the Soybean Germplasm Collection and other germplasm collections around the world. We will highlight this in our publications.
- **Incorporated high-throughput phenotype data, plant developmental data, and environment data in the models.**
  We collected multi-sensor phenotype information on the NCSRP germplasm sampling validation plots in NE, MN, KS, IA, MO, IN and IL at two stages of development during the 2019 season: (1) near V5 and (2) just prior to R5. Research results from our prior work and others shows that image and spectral data from these two growth stages show highest correlations with yield. We have some good video from these trips that can be used for presentation, education, and PR in the future for our SOYGEN objectives. After the 2019 data are received after harvest, we will begin work on compiling and analyzing the data from the validation study, including the high-throughput phenotype information.
- High yielding lines derived from crosses with G. *tomentella* have been identified, no introgressions from *G. tomentella* have been identified.
- Multiple lines derived from wild soybean (G. soja) F2 crosses with Williams 82 or BC1 crosses with Williams 82 yield 80-97% and 89-95% of the checks over 3 years, respectively.
- We have putatively identified causal mutations for maturity genes.
- We have putatively localized the causal mutation in the GmBa1 gene for branching angle to the promoter region.
- A scientific manuscript that reports the contribution of wild type soybeans (*G. soja*) as a source of genomic diversity in soybean elite lines is in prepartaion.

- A scientific manuscript that reports the candidate regions under selection for yield in the alternative and conventional gene pools, and the potential application of this results for increasing the yield in the elite lines is in preparation.
- Workflow scripts for studying population genomics in soybeans using SNP array data have been developed. Using this resource, the researcher/breeder will be able to select samples in populations based on the results of genetic diversity, structure, genetic association, and selection analysis.
- We have identified the direction of selection for each candidate haplotype in elite populations. This result might contribute for increase the efficiency in the selection of parental combinations and decrease the length of breeding cycle.

3.f. Key Performance Indicators or performance measures (year 3).

- A detailed list of 18 candidate haplotypes/genomic region underlying yield-related traits has been identified that are under selection in the elite soybean lines from conventional and alternative pools.
- A list of Alternative Elite Lines belonging to the MG III and MG IV has been identified containing the combination of most favorable haplotypes that could be used for increasing the performance in the elaboration of parental combinations in soybean breeding crosses

OBJECTIVE 4: Development of a metric to estimate genetic gains on an annual basis

In order to estimate the realized genetic gain (RGG) on an annual basis, a simulation software based on the package AlphaSimR was implemented in the R environment. The simulation pipeline consists of a simulation of 14 years of multi-environment trials (MET) of a soybean breeding program (Figure 1). The parameters of the simulation, such as heritability, number of genotypes and locations, selection intensity, etc., will be based on real public breeding programs. These parameters were obtained in interviews with Brian Diers, Aaron Lorenz, George Graef, and Leah McHale. The main goal of the simulation is to have the true genetic values of genotypes across generations and years, and therefore to estimate the RGG. The following six models were also implemented in the R environment to estimate the RGG: (*i*) meta-analysis for yield correction due to a reference year; (*ii*) linear regression with unadjusted averages of genotypes; (*iii*) linear mixed model with simple effects; (*iv*) linear mixed model with a contrast of checks against regular genotypes; (*v*) three-way linear mixed model with subsequent simple linear regression of adjusted means of genotypes and years; and (*vi*) GBLUP model. We are working on evaluation criteria for evaluating the models.

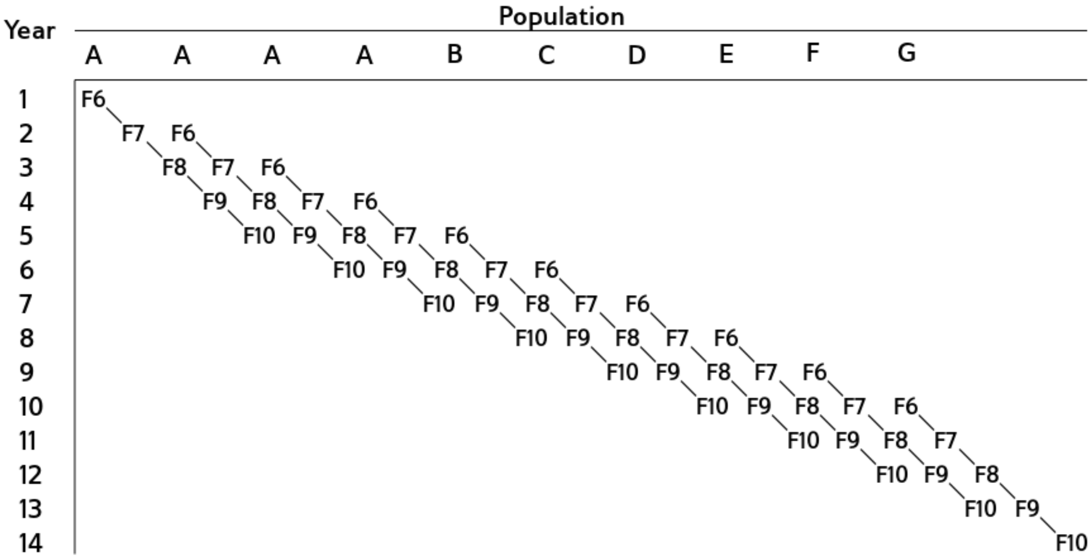

Figure 1 – General scheme of the simulation pipeline.

4.e. Deliverables.

- For the production of short videos describing history and future developments of genetic gain to non-experts, interview questions for farmers in January/February 2020 have been developed.
- Simulation of Genetic gain using genomic and phenotypic selection in Soybean breeding population structures were documented with R-markdown and are being evaluated before depositing in a public repository.
- Simulation software identified (AlphaSimR) for generating yields of potential varieties in various stages of field trials.

4.f. Key Performance Indicators or performance measures (year 3).

- Three manuscripts are in preparation to document the establishment of four objective criteria for evaluating methods that estimate realized genetic gains.
- Publically available simulation software was identified that is sufficiently flexible for a skilled graduate student to generate yields of potential varieties from multiple (>1) families grown in multiple (>1) environments in at least one stage of field trials.