Final Report

2/22/2018

Implementing and Vetting Genomic Selection in Soybean

PIs: George Graef[1], Aaron Lorenz[2], and David Hyten[1]
[1]University of Nebraska-Lincoln and [2]University of Minnesota

Historically, the Nebraska soybean breeding program has made tremendous progress improving yield in soybean based on phenotypic evaluation and selection; that is, yield testing over multiple locations and years to identify high-yielding lines that are consistent over environments. We use pedigree information to maximize and maintain genetic diversity in the crosses and lines that are developed, but have not used DNA genotype information. Now, that DNA-level information is available at affordable cost and may enhance our ability to identify superior lines for advancement and crossing. There are different DNA genotyping platforms available, including fixed-array chips or various whole-genome sequencing methods. Based on cost per sample and the amount of information gained, we decided to use the genotyping-by-sequencing approach for our genomic selection study.

The objectives of this project were:
Put into practice genomic prediction and selection in the UNL Soybean Breeding Program. It will be thoroughly compared to phenotypic selection to assess its true potential for increasing rate of genetic gain for grain yield.

Sub-aims include
1) Develop in-house genotyping methods to reduce costs and improve turn-around time, and 2) Develop and test models for enhancing prediction accuracy through modeling the interaction between environment and cultivar.

The use of genotyping-by-sequencing (GBS) for genomic selection holds good potential for improving soybean grain yield. For this project, **we looked at three main comparisons**:

(1) **Understand the size and composition of the "training population" that is optimal for successful genomic selection in the soybean breeding program**. That is, we first need to calibrate our genotype and phenotype information. So we have a group of high-yield lines that were tested in multiple environments and years to obtain good yield, maturity, and other performance information. We then obtained the DNA genotype information on those lines and through statistical analyses obtained information on the relationship between the genotype information for each line and its average yield performance. With that information relating overall genotype information to average yield performance, we could then apply just the genotype information to subsequent lines to predict their yield performance, based just on the degree to which their genotype information relates to that of the high-yield lines in our calibration set.

(2) **Compare phenotype selections, genotype selections, and a set of randomly selected lines in two validation populations to evaluate genomic selection relative to our normal process of phenotypic selection, and compare both to a random control set**. Three hundred and one soybean experimental lines in advanced stages of the University of

Nebraska-Lincoln Soybean Breeding Program were used as a training population to predict genetic values for yield, maturity date, and plant height on two soybean test populations comprised of 373 breeding lines (UX2862) and 415 breeding lines (UX2872), respectively **(Fig. 1)**. The test populations and training population were both genotyped with GBS. Filtering and alignment of the same GBS SNPs (single-nucleotide polymorphisms) between training and test populations was done sequentially. The SNPs with 80 percent missing values (PMV) and with minor-allele frequency (MAF) > 0.05 were used, giving a total of 3,669 SNPs for UX2862 and 3,107 SNPs for UX2872. After filtering, remaining missing values were imputed using naive imputation method. This method is not expected to add information, but rather serves the purpose of ensuring unchanged allele frequencies after imputation and provides a marker matrix containing no missing data so that analytical operations can be performed. The prediction equation (RR-BLUP model) was then used to evaluate individuals in the test population, which has ***both*** **genotype** (DNA SNP) and **phenotype** (multi-location yield test) information. The best breeding lines were then selected based on their **(1) phenotype** – that is, average yield over seven environments during 2014, and **(2) genotype** – that is their marker-predicted genotypic values for yield, maturity date and plant height. The top ten predicted highest yielding breeding lines for UX2862 are shown in **Table 1**.

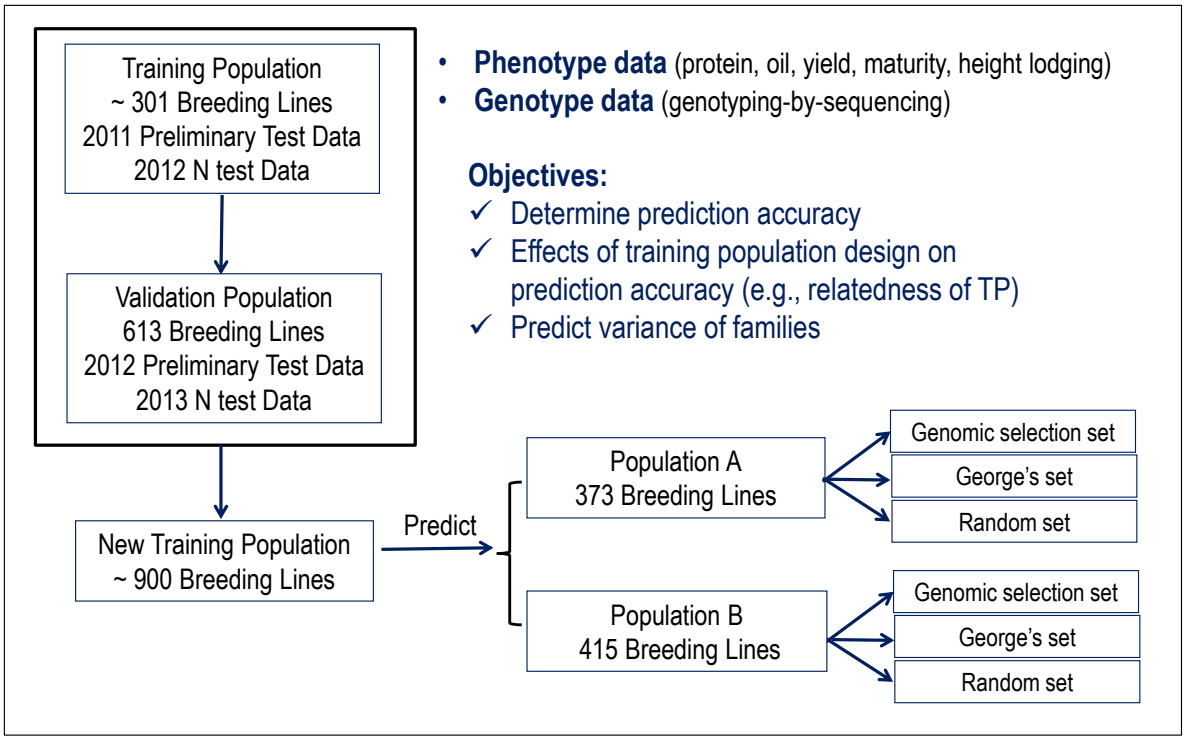

Fig.1 A general scheme on evaluating the potential of genomic selection to enhance the speed and accuracy of soybean breeding programs at the University of Nebraska-Lincoln. Training, validation, and test populations for genomic selction are presented.

Table 1. The top ten highest yielding breeding lines of UX2862 test population based on genomic-estimated breeding values. This population was genotyped by GBS.

Genotype ID	Yield	Height	Maturity
UX2862-229	4596.18	106.11	32.02
UX2862-363	4590.54	102.24	27.01
UX2862-096	4566.37	108.19	32.67
UX2862-409	4534.33	96.66	25.14
UX2862-115	4532.71	109.25	32.31
UX2862-348	4523.03	99.42	29.40
UX2862-014	4518.79	106.29	30.53
UX2862-435	4511.68	95.92	26.25
UX2862-401	4506.98	94.62	24.51

A **validation set** comprised of 613 breeding lines (i.e., an independent set of entries used to test the accuracy of predictions) was also constructed using progenies derived from crosses between selected lines included in the training population. The validation set was genotyped with the GBS and phenotyped in at least eight environments (four locations, two years) with two replications per environment. This dataset forms the best available validation set in the genomic selection community for giving more power to make inferences. With the complete GBS data, the genetic value of lines contained in the validation set will be predicted using standard genomic selection models. The accuracy of the predictions will be evaluated by correlating them with the observed phenotypes. Due to some delays in final genotype data, this analysis is currently underway.

Finally, we will combine all of the genotpe information obtained to date in the original training set plus the validation set to make a new training set with genotype and phenotype information on nearly 1,000 soybean lines tested over environments and years from multiple different crosses. That new training set with recalibrated genotype to phenotype information, will be used to predict yields of our current newly developed breeding lines based on genotype.

(3) **The longer-term comparison and evaluation of the value of implementing genomic selection in the breeding program is ongoing. It involves comparison of progenies from crosses made among parent lines that were selected based on either *genotype* or *phenotype*.** We made selections based on seven environments of evaluation in Nebraska, Iowa, and Illinois during 2014 to identify the 20 highest-yielding lines that were used for crossing in our Puerto Rico nursery during the 2014-2015 winter season. These crosses comprise the "Genomic Selection Phenotype Crosses," and progeny lines from the 75 populations that were developed have been advanced through our breeding program thorugh progeny rows and multi-location preliminary yield tests in Nebraska during 2016 and 2017. The comparison set of parental selections based on genotype information was delayed due to delays in receiving the genotype informaiton from the external lab that was being used. Consequently, the "Genomic Selection Genotype Crosses" were made during the 2015-2016 winter seasson in Puerto Rico, one year behind the phenotype crosses. Progeny lines from the 75 populations that were developed from the genotype-selected parents have been advanced through populations and progeny rows in our breeding program during 2016

and 2017. They will be in their first multi-location yield tests in Nebraska during the 2018 season. The outome of lines from the genomic selection crosses vs. the phenotype selection crosses will be followed through the program during the next 3 years to document performance of genomic selection vs. phenotypic selection in the soybean breeding program.

Summary findings, coments, and plans going forward

Some key findings from our work indicate that we have an efficient breeding program that provides high-quality phenotype data from our high-yield Nebraska locations, and that inclusion of genotype information may help improve selection of superior lines for crossing and production. It is important to have high-quality phenotype information to calibrate the genotype-phenotype relationship in the training population and be able to make more accurate predictions. Genotype representation in the training set and target populations also is important, so there is a question on the size of the training set that is needed to adequately represent the genotypes that will be encountered in the target populations. A brief summary of main findings is listed here:

- We don't need more than ~150 lines in the training set to effectively predict performance in the target population.
- More closely related training and target populations result in improved prediction accuracy.
- Cost of genotyping per sample still is above where we can realistically process 10,000 samples in our program, but we are making progress on decreasing costs per sample and improving throughput and quality of genotype information with development of the in-house genotyping methods.
- Prediction accuracy is relatively high (~0.60), indicating that we could potentially improve overall efficiency by adding genotype information to the phenotype information from our line evaluations.
- Genotype information may improve our breeding progress by allowing testing of more lines in preliminary tests through sparse testing and prediction of missing lines. I will outline this more in a later update on the breeding program as we go forward. But briefly, say we normally advance 10,000 plants from F4 populations to progeny rows. Then I select (visually) about 1,500 or at most 2,000 of those progeny rows to advance to our multi-location yield tests in Nebraska. All lines are tested at four locations, for a total of 8,000 yield plots at that stage. Compare that to selection of lines with genotype information. Suppose we can genotype all 10,000 plants before sending them to progeny rows. So, with 60% prediction accuracy, say we retain 40% of the plants and send the top-ranked 40% (based on their genomic predictions for yield) to progeny rows in Chile. That is 4,000 rows instead of 10,000. Then I harvest all 4,000 and test them in multi-location tests in Nebraska. Except that now we have genotype information on all those individuals. So for the yield testing, we could potentially test only a portion of the 4,000 lines, with an experimental design that tests all of the lines at one or more locations. With the genotype information, we can then *predict* the performance of the lines that are missing at individual locations, using their performance at the other locations *plus* the genotype information from all the other lines. So, assuming that we could have up to 50% missing lines, then for the same 2,000 yield test plots at a location, we could effectively evaluate 4,000 lines

instead of 2,000. This is one part of the overall implementation of genomic selection in the breeding program that we will evaluate going forward.

- We will continue with the comparison of lines derived from crosses based on genotype selection vs. crosses based on phenotype selection, and make appropriate modifications to the overall breeding program based on the findings of that comparison.
- By the end of 2018, the new marker system should be more cost-effective and throughput more efficient, so we will be able to accommodate the numbers needed for closer to full implementation in the breeding program going forward.
- We are working on finalizing the manuscript for the studies outlined in Figure 1, and the paper should be submitted for journal review before the end of April 2018.

Thank you for your support of this project. We will be able to provide valuable information on implementation of genomic selection in soybean breeding to the research community, and to move ahead with implementation of key findings in our own breeding program to enhance our breeding progress for yield and quality. Ultimately, this will enhance progress and provide soybeans with superior yield and quality to soybean producers.