North Central Soybean Research Program

Identifying high-yield genotypes in the USDA soybean germplasm collection

George Graef (Project leader) Ggraef1@unl.edu, Asheesh Singh (Iowa State University), William Schapaugh (Kansas State University), Randall Nelson (University of Illinois at Urbana-Champaign), Brian Diers (University of Illinois-Carbondale), Andrew Scaboo (University of Missouri), Aaron Lorenz and Kent M Eskridge (University of Nebraska)

The USDA Soybean Germplasm Collection contains over 21,000 accessions including wild relatives, landraces, and cultivars from around the world. The majority of unimproved accessions come from China, where soybean was domesticated, as well as Japan and Korea, other areas of ancient cultivation. Domestication resulted in a loss of genetic diversity, with landraces retaining only about 63% of the diversity found in the wild *Glycine soja*.

Furthermore, 86% of the parentage of US commercial soybean cultivars released between 1947 and 1988 are accounted for by only 17 ancestral PI accessions. Because it is limited, we need to more effectively use the available diversity in soybean. The goal of this project is to identify and use soybean germplasm with positive alleles for yield and other traits that can be bred into commercial cultivars to effectively increase productivity and expand the genetic base of US soybean varieties. A major challenge in plant breeding is how best to sample a large germplasm collection where phenotypic information for traits such as yield is absent or very limited.

Project Objectives

- Determine how we can effectively and efficiently select from among the large number of accessions in the collection to identify lines to use in a breeding program to increase yield.

- Identify specific loci and alleles (genes) related to improved productivity in both US cultivars and the untapped soybean germplasm.

Benefit to Soybean Farmers

The ultimate benefit to soybean farmers is faster development of higher yielding soybean varieties with superior agronomic and quality traits. We will be able to make more informed decisions on choice of PI accessions from the collection to use in breeding programs. The final information from this project will be publically available, so both public and industry research and development programs will be able to use it. We also could greatly expand the diversity of the commercial soybean germplasm base, which may provide more resilience in varying climate

conditions as well as allow for longer term gains in yield vs. those possible without added diversity.

Progress Report October 2016

We grew seven environments of yield tests for each of the Maturity Group sets representing 500 soybean accessions from the USDA Soybean Germplasm Collection. Each experiment was grown in two replications of an augmented block design at locations in MN, NE, KS, IA, MO, and IL. Harvest for 2016 is just beginning. Reports from cooperators indicate that plots are in good condition. If all 2016 data are of high quality, our final dataset will include 13 environments (26 reps) for the MG1, MG3, and MG4 sets, and 14 environments (28 reps) for the MG2 set of entries. This kind of extensive, high-quality yield and agronomic data does not exist for such a large collection of PIs. That is one important and major deliverable of this cooperative NCSRP project.

That extensive, high-quality yield information is needed to develop effective models that can predict yield of the 19,000+ untested accessions in the collection based on their genotype information. We need excellent phenotype data to enable better prediction of performance of the untested lines. One test of the merit of a prediction model is how well it can estimate phenotype performance of a set of untested lines from among the lines that we tested. Using different cross-validation methods that leave out a percentage of the lines and attempt to predict their performance based on the remaining data, we evaluated different models with our current set of data. So far, we have only the first year of data from 2015 harvest available for comparisons. Addition of the 2016 data will double the amount of phenotype information, which should improve prediction accuracy.

Results from comparisons of prediction models using the 2015 yield information are summarized in **Table 1** (below).

The main points from the results are:

(1) including genotype information significantly increased prediction accuracy of the model (compare results of model 1 vs. model 2),

(2) adding genotype x environment interaction information did not improve prediction accuracy (compare results of model 2 vs. model 3),

(3) having at least some phenotype information on a line greatly increased prediction accuracy of all models (compare results of CV1 vs CV2), and

(4) the Super-Saturated Design (SSD) group generally resulted in greater prediction accuracy than the Cluster or Random sets of entries.

Cross Validation=>		CV1							CV2						
Model =>		E+M:E+L			E+M:E+G			E+M:E+G+GE	E+M:E+L			E+M:E+G		E+M:E+G+GE	
		mean	sd		mean	sd		mean	sd	mean	sd		mean	sd		mean	sd
Sampling Method																	
CLUSTER		-0.244	0.126		0.511	0.039		0.516	0.044	0.667	0.030		0.691	0.680		0.689	0.029
RANDOM		-0.251	0.140		0.471	0.048		0.484	0.051	0.692	0.028		0.695	0.025		0.709	0.026
SSD		-0.245	0.154		0.552	0.040		0.547	0.043	0.718	0.027		0.727	0.024		0.724	0.027

Table 1. Summary of mean prediction accuracy (mean) and standard deviation (sd) based on soybean accessions selected by three different sampling methods using two different cross validation methods (CV1 and CV2) and three different models. The models are (1) E+M:E+L, which uses Environment (E), Maturity Group within environment (M:E), and lines (L); (2) E+M:E+G, which uses Environment, Maurity Group within environment, and Genotype (G). Genotype is the 50K SNP data; and (3) E+M:E+G+GE, which adds genotype-environment interaction information (GE) to the second model. For the two cross validation methods, CV1 is predicting the new line that has not been observed (phenotyped). The CV1 method used 5 folds, assigning 20% of the lines randomly to each fold, and repeated 50 times for each of the models. The CV2 method simulated incomplete field trials, so it is using phenotype information from lines tested in a subset of locations to predict their performance overall; again with entries assigned randomly to 5 folds, repeated 50 times.

Distribution of average yield for entries in each sampling method was pretty similar across methods, with a slightly smaller range in the RAN group vs. CLU and SSD **(Figure 1).** Prediction accuracy estimates for maturity, height, and lodging were good, but slightly lower than for yield. Prediction accuracies for seed weight were high, >0.95 for cross validation method 2 for all three models.

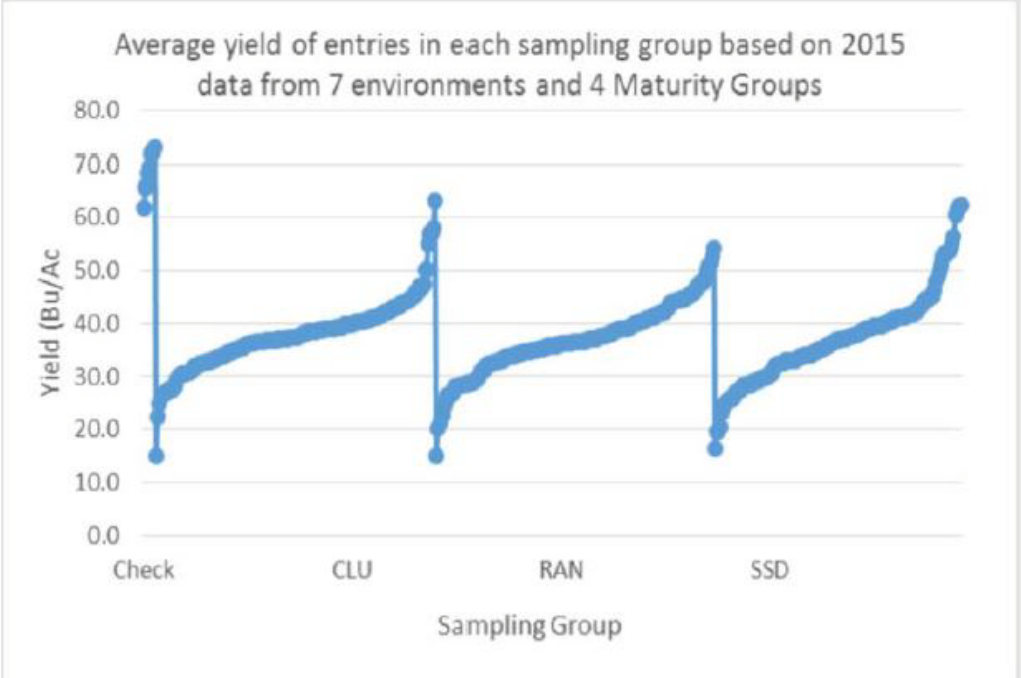

Figure 1. Distribution of average yield for entries in each sampling group, based on 2015 data from 7 environments for each of 4 Maturity Groups I to IV. Check=standard yield check entries in each block, CLU=Cluster, RAN=Random, SSD=Supersaturated Design.

After 2016 harvest, we will complete the 2016 and 2-year combined analyses for all the data. We also will complete the genome-wide association (GWA) analysis for each year and combined data, and for each sampling method and over all sampling methods. The GWA mapping of traits will provide some indication if there are particular loci with significant effects on yield or some of the other traits measured. Through that analysis, we can identify particular alleles, or haplotype blocks (small regions in the genome), that are correlated with the phenotype of interest (yield). Another analysis is based on the overall genotype of the individual, or its genomic estimated breeding value (GEBV), determined by using both the 14 environments of yield and other phenotype data along with the 50K SNP genotype information.

The breeding value is kind of an indication of how good that line will be as a parent to produce offspring with superior phenotypes – higher yielding lines. Together with the specific allele information that we find from association mapping, the GEBV will be useful in identifying lines that are more likely to be good parents for enhancing yield in our soybean breeding programs. It is this information that we will use to go back to the remaining untested accessions in the USDA Soybean Germplasm Collection to select 200 or more lines (based on their 50K SNP genotype information) for our validation study.

The fact that the SSD sampling method resulted in better prediction accuracy could be important to improve our efficiency and effectiveness of sampling this and other large germplasm collections to identify superior genotypes. If we can make the same predictions from good phenotypic evaluation of 160 individuals vs. 500 or more, we can be more efficient with limited resources and obtain improved phenotype information on a smaller set of more diverse genotypes.

Additional summary data from 2015 tests are shown below. Lines were pretty equally split among sampling methods, from 155 lines in the SSD group to 169 lines in the RAN group **(Table 2).** Distribution among maturity groups was not quite as even, but the percentage representation of sampling methods within maturity group was pretty consistent **(Figure 2).**

MG	CLU	RAN	SSD	TOTAL
I	34	28	31	**93**
II	41	56	31	**128**
III	32	30	26	**88**
IV	61	55	67	**183**
Total	**168**	**169**	**155**	492

Table 2. Distribution of soybean accessions by maturity group and sampling method. CLU=Cluster, RAN=Random, SSD=Supersaturated Design.

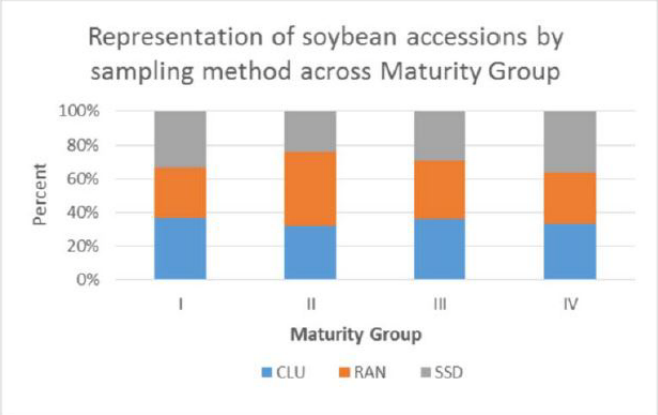

Figure 2. Distribution of soybean accessions by sampling method across maturity group. CLU=Cluster, RAN=Random, SSD=supersaturated design.