November 2016 Final Report

This is the final report covering years 1 and 2 of an initial 3-year proposal, “Identifying high-yield genotypes in the USDA Germplasm Collection.” Year 3 and subsequent validation will follow as part of the new NCSRP project “Increasing the rate of genetic gain for yield in soybean breeding programs.” Progress is documented in the previously submitted semi-annual reports and technical reports. Because the most recent progress report submitted October 2016 brings the current project mostly up to date, this final report will focus on the status related to KPIs and Deliverables for the overall project. As of this date, cooperators are still collecting and processing their 2016 data, which is due to be submitted to me at Nebraska by December 8, 2016. After that date, we will work on data checking and analyses for 2016 and over years.

The overall aim of this project is to facilitate selection of superior accessions from among the more than 20,000 soybean accessions available in the USDA Soybean Germplasm Collection. With so many soybean accessions to choose from, and limited or no yield information, that becomes a sampling question. So we employed some statistical methods to sample the collection in three different ways, using the 50K SNP genotype information to assess the genotypic diversity among accessions. Two different sampling methods were aimed at maximizing the diversity among the selections in the sample. The third method was a random sample for comparison. We developed three sets of soybean accessions selected from the USDA collection based on the three selection methods and conducted extensive multi-environment yield trials on those selections during 2015 and 2016. We grew 505 total accessions of MG I to MG IV in NE, MN, IA, IL, and MO, and included a MG III and a MG IV location in Kansas during 2016.

Our 2015 multi-location yield tests provided very good data, especially considering the unadapted nature of most of the soybean lines we are testing, and the variation in environments for 2015 across the region. We just completed harvest for the 2016 season, and in general conditions were excellent. We now have high-quality phenotype data for yield and other agronomic traits from 14 environments over 2 years. We can now relate the phenotype information (yield, plant height, maturity, lodging, seed weight, and seed composition) to the genotype information in the USDA germplasm collection database.

The amount of yield and agronomic data we now have exceeds the quality of any yield plot data currently available on this number of accessions in the collection. Through this coordinated effort, we have 28 replications worth of data from 14 environments in five major soybean producing states in the north-central region. That is valuable, and very important to get good estimates of yield and other traits across environments to relate to the genotype information and develop good prediction models. The prediction models we develop based on this sample of genotype and phenotype information from 500 accessions in the collection will be used to go back to the entire collection and resample new, untested soybean genotypes in MG I, II, III, and IV, that are predicted to have superior genotypes for improving yield. The next phase of the project will evaluate those reselected lines from the collection to validate the models and provide superior soybeans for use in public and commercial 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. This offers the potential to 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 compared to those without more genetic diversity.

**Progress related to Milestones listed in the proposal**: We have met all expected milestones for this proposal.

**Progress related to Deliverables listed in the proposal**:

**Deliverables:**

1. **High-quality, multi-environment yield and other agronomic performance data for 500 accessions in the USDA Soybean Germplasm Collection**. High-yielding accessions with unique yield genes will be used in public and private breeding programs to increase yield.

*Progress: The 2016 data will be processed, analyzed, summarized and added to the 2015 data for a total of 14 environments of yield, agronomic, and seed composition data on 500 accessions.*

1. **Identify yield-marker genotype relationships based on association mapping results from the extensive, high-quality yield dataset.** Information from the 2015-2016 2-year, multi-location yield tests and the 50K SNP genotype data will be used to identify genomic regions, or haplotype blocks, that are associated with yield in soybean. This information will be used in public and private breeding programs to increase yield, rate of genetic gain, and genetic diversity of the commercial soybean germplasm pool.

*Progress: this analysis depends on the 2016 data, and will be completed during the November 2016 through March 2017 period as part of the continuation project.*

1. **Develop predictive model(s) that allow selection of superior high-yield genotypes from the USDA germplasm collection**. From each sampling group (SSD, CLU, and RAN), as well as for the group of 500 accessions overall, we will develop predictive models to allow us to go back into the germplasm collection and select untested lines based on genotype. Validation of the models with yield and other phenotype testing will be a follow-up project.

*Progress: We evaluated three different models using the 2015 data, and will redo that process for the 2016 data and overall 2-year dataset. Some results from the use of the 2015 data for yield predictions were submitted as part of the technical report for the October 2016 semi-annual report.*

1. **Public use of data, documentation of results.** Results will be published in refereed scientific journals. Data from all tests will be made available to all users through SoyBase and possibly GRIN. Details will be worked out with USDA and SoyBase administrators to facilitate availability and use.

*Progress: This deliverable is at the end of the project, so will be available after all the data analyses are completed. We aim to have the first paper submitted to a scientific journal by June 2017. I will work with USDA contacts regarding addition of the data to GRIN and/or SoyBase.*

**Progress related to KPIs listed in the proposal:**

**KPIs/Performance Metrics**

* 1. High quality yield and seed composition data on 500 soybean accessions from the USDA Soybean Germplasm Collection from 14 environments, 7 environments in each of 2 years.

*Progress: Done for year 1. Year 2 (2016) is currently being processed.*

* 1. Preliminary model to predict yield and seed composition on accessions in the USDA Soybean Germplasm Collection.

*Progress: Done for YIELD, MATURITY, PLANT HEIGHT for 2015. Working on 2016 data and predictions for other traits.*

* 1. Based on association mapping analyses, identify at least one QTL related to each of following traits: seed yield, increased protein concentration and increased oil concentration.

*Progress: Association analyses not done yet. These studies will follow processing of the 2016 data.*

* 1. Possibly identify QTL for other seed composition traits that we measure, as well as maturity, height, lodging, R1 date, and length of reproductive period.

*Progress: Not done yet. Will be completed after processing 2016 data.*