North Central Soybean Research Program

Increasing the rate of genetic gain for yield in soybean breeding programs

Leah McHale, project leader (Ohio State University), Matthew Hudson, Brian Diers and Steven Clough (University of Illinois-Carbondale), Andrew Scaboo,and Pengyin Chen (University of Missouri), George Graef (University of Nebraska), Katy M Rainey (Virginia Tech), Aaron Lorenz, (University of Minnesota), William Beavis, Asheesh Singh, Silvia Cianzio and David Hyten (Iowa State University), William Schapaugh (Kansas State University), Dechun Wang (Michigan State University),Jianxin Ma (Purdue University)

Objectives

1. To Increase selection intensity and decrease non-genetic sources of variability through improved progeny row testing
2. To increase selection coefficient and decrease length of breeding cycle through genomic selection
3. To Increase additive genetic variance
4. To develop a metric to estimate genetic gains on an annual basis

Progress report for 2017

Earlier this year, the group had brain-storming session over email in which we decided that it would be wise for us to "brand" our project with a name that can envelope this group, but also serve as an umbrella that could potentially include future projects (hopefully from other funding sources, including federal grants). The name that was chosen was SOYGEN: Science-Optimized Yield Gains across ENvironments. A project website is under development.

Specific updates for each objective follow.

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

Task 1: Collection of additional data in all progeny rows.

Participating breeders are currently growing progeny rows and collecting data additional to their usual data collection on 5000 of these rows. Data collected can vary from yield, to maturity, to canopy coverage, etc. Two breeders experienced loss of some portion of the 5000 rows due to dicamba drift or poor quality seed. However, do to the high participation in this objective, loss of two sites is not expected to affect the outcome. Multiple collaborators were able to capture images of canopy coverage around canopy closure time.

Increasing selection coefficient and decreasing length of breeding cycle through genomic selection

Task 1. Complete study on genomic selection using the nested association mapping of soybean population (SoyNAM) and apply findings to ongoing genomic selection effort.

All data analysis is complete. Preparation of the manuscript is underway. We hope to

have a manuscript submitted by 2018. In summary, genomic prediction appeared to work very well within the NAM population, but a substantial amount of accuracy was lost when going from the NAM population to the new breeding populations. While this is a disappointing result, we feel it illuminates our approach to incorporating genomic selection into practical breeding programs and gives us direction on how to build the phenotype-genotype databases described below.

Task 2. Compile existing phenotypic, genotypic, pedigree, and environmental data (weather, soil) from various projects on yield and diversity conducted in the North Central Region.

Historical phenotypic data (yield, maturity, lodging, protein, oil, disease resistance, etc.) going back to 1989 has been completely cleaned and compiled and is ready for upload to a relational, publicly available database (described below). Pedigree data going back to 1989 has been cleaned and compiled and is ready for upload. We have a total of 8120 strains in this dataset and a thorough pedigree cleaning was a lengthy task. Line names and pedigrees were curated to 1) identify and eliminate typos and duplicate names created by typos; 2) generate an alias database of lines with duplicate names; 3) generate machine readable and traceable pedigree structures for complex crosses. Metadata and trait ontology has been refined and is ready for upload.

Regarding the database development, we initiated a database using the T3 platform on a local machine at the time of the last report, and have since migrated that to a University of Minnesota College of Food, Agriculture and Natural Resource Sciences server. This will serve as a long- term home for the database as well as make it publicly available. The platform construction is underway and is expected to be completed shortly. Once the platform is constructed, we will immediately upload all URT data. We have discussed this with Soybase curator David Grant, and he is enthusiastic about pointing Soybase to our new database. Finally, we have hired a part-time data curator to help with the task of cleaning and curating data.

Another activity we have been involved with is working with BensonHill Biosystems in uploading our data into their database. We are one step away from uploading this data to them. BensonHill has a unique system where they take available genotype-phenotype data and use it to help enable predictive plant breeding in the private sector. We feel this is one direct way in which the data we generate from this project can have impact on soybean improvement.

Next steps for this part of the project include: 1) Upload all available URT data to database, 2) Begin compiling and cleaning all other publicly available genotype-phenotype datasets in soybean.

Task 3. Genotype all available soybean lines grown in the USDA Northern Uniform Tests beginning in 2004.

An additional 376 lines have been extracted and submitted for genotyping to the UMN Genomics Center, including all newly entered lines from the 2017 URTs. During this summer, we inquired with breeders about additional seed available and we received an additional ~200 lines from previous trials. We have cleaned and organized this seed and are in the process of extracting DNA.

Next steps: 1) Continue to identify additional missing lines and genotype them (for example, we are traveling to SDSU next week to obtain seed from their defunct breeding program); 2) QC and upload new genotype data in relational database and link to phenotypic data.

Task 4. Unify genotypic data collected from the multiple platforms, using a single flexible data management system, capable of adapting to any genotyping platform.

Thus far, the group has been focused on genotyping from "genotyping-by-sequencing" (GBS) methods. We compared the performance of five GBS pipelines using low-coverage Illumina sequence data from three soybean populations. To address issues identified with existing methods, we developed GB-eaSy, a GBS bioinformatics workflow that incorporates widely used genomics tools, parallelization and automation to increase the accuracy and accessibility of GBS data analysis. Compared to other GBS pipelines, GB-eaSy rapidly and accurately identified the greatest number of SNPs, with SNP calls closely concordant with whole-genome sequencing of selected lines.

We showed that GB-eaSy is approximately as good as, or better than, other leading software solutions in the accuracy, yield and missing data fraction of variant calling, as tested on low- coverage genomic data from soybean. It also performs well relative to other solutions in terms of the run time and disk space required. In addition, GB-eaSy is built from open-source modular software packages that are regularly updated and commonly used, making it straightforward to install and maintain. A manuscript describing GB-easy is under review currently.

Task 5. Development of ultra-cheap low-density marker system for genomic prediction applications.

In the development of the genotyping methods, we have screened 14 different methods of DNA isolation that could have the potential for producing DNA suitable for sequencing. From these methods we selected the most promising one that met all the criteria of being cheap, automatable and would likely produce data with the GBS protocol. We compared the new DNA extraction method to a standard high quality DNA extraction method using a 4k-plex genotyping-by- sequencing method. For the high quality DNA extraction, our on-target sequence was 60-70%. This matches the expected on-target rate for this GBS protocol. The new method produced an on- target sequence rate of 30%. While this is lower than the expected 60% this is a good indication that this new DNA extraction method has potential for GBS. We will continue to work on optimizing this method and test methods that could help purify the DNA further to increase on- target sequence without adding a significant amount of cost.

We have also ordered and begun to test a SNP probe set created from individual oligo synthesis. This is different from the 4k-plex which was created by a pool oligo synthesis method. The individual oligo synthesis will provide an ability to adjust oligo concentration based on the efficiency of each oligo within a reaction. Since each oligo is individually synthesized we will also be able to create low-density GBS marker sets specific to germplasm that would need to be genotyped. The individual oligo synthesis method has an initial large up-front cost for oligo purchase but the amount of oligo received is enough to do 1 million reactions. With the initial cost spread out over the lifetime of using the oligo in GBS reactions, the individual synthesis method will not add a significant cost per reaction for genotyping.

Increasing additive genetic variance

Task 1: Exploration of a retrospective analysis

A part time post-doc, Dr. Mao Huang, has recently been hired to determine to the feasibility of a retrospective analysis using existing public data from the URT (compiled as part of objective 2). The retrospective analysis is designed to predict the success of parental combinations and evaluate their success based on the relative performance of

their progeny (total progeny tested/total progeny advanced). We have thus far determined that the number of lines tested in the URT which were derived from parents which were tested in the URT and are part of the set of lines to be genotyped as part of objective 2 is small (~80) and insufficient for a valid statistical analysis. Thus, any retrospective analysis would require further data acquisition from breeders on

the number of progeny tested from specific cross combinations. A survey is currently being designed to determine the ability and willingness of public soybean breeders to provide this information.

Task 2: Evaluation of germplasm mined from the USDA Soybean Germplasm Collection using genomic prediction

We compared models with and without genotype data, and with and without GxE effects, to predict yield performance of soybean germplasm accessions. We performed predictions using each sampling group as a training set (SSD, CLU, and RAN) as well as using all the data for all 500 PI over all sampling groups as the training set. The SSD set performed as well as the complete set, and addition of GxE effects in the model did not improve prediction accuracy.

From the set of ~9,400 new accessions whose yield was predicted based on our 2015-2016 2-year yield evaluations in 14 environments and 28 reps, we selected 250 accessions to use in the validation set. The distribution of yield predictions was divided into quintiles, and we sampled 50 lines from each of the five quintiles of the distribution so we would end up with a sample of lines spanning the entire range of yield predictions. To select the 50 lines within each quintile, we used the supersaturated design analysis to identify 50 lines with maximum diversity within each group. Those 250 lines were then obtained from the collection and are currently growing in Nebraska for seed increase. We collected descriptive data on all lines, as well as maturity date and lodging and shattering scores. We just completed harvest of the MG1 accessions, and will complete harvest of all accessions by November 1. This is the seed source that will be prepared and distributed to cooperators for the 2018 and 2019 multi-location yield evaluations for these 250 PIs to obtain actual yield and agronomic data to validate the predictions.

Task 4: Identify signatures of selection in G. max derived lines selected for high yield

We have assessed population structure and signatures of selection in founder lines and high yielding elites of both conventional US cornbelt varieties and the alternative gene pool ancestors and elite lines developed in Randy Nelson's breeding program. When assessed by selection pedigree, the lines generally cluster with prominent overlap between ancestor lines suggesting that those genetic groupings are not completely genetically distinct, but share a pool of most common alleles.

Analysis of Fst agrees with the cluster analysis where ancestor and elite lines from each selection effort differ more from each other than ancestor lines differ between each other. A novel finding from the Fst analysis was that elite lines also do not seem to differ from each other across the genome except at specific markers (e.g., no selective peaks, but individual SNPs). Linkage-based assessments (Rsb) mirror Fst findings, and further suggest that elite lines from both selection efforts differ from respective ancestors at similar genetic regions. Another linkage-based analysis (H1) showed few regions under selection shared between elite lines, more regions unique to conventional elites while a few regions unique to alternative elites. We are now grouping the genome into haplotype blocks to identify specific haplotypes under selection for elite traits.

Task 5: Identify introgressed regions from wild soybean in regions of domestication genes and regions associated with high yield

We completed the identification of QTL controlling key domestication-related traits (DRT) using 151 RILs from Williams 82 x PI 468916 and 510 RILs from Williams 82 x PI 479752. We measured 11 phenotypic traits: flowering date, maturity date, main stem length, lodging, stem

diameter, growth habit, leaflet length, leaflet shape, shattering, pubescence form and 100-seed weight. A total of 97 QTL where identified with 2 to 11 QTL per trait. The majority of the domestication related traits examined in this study were controlled by minor QTL, with QTL explaining over 50% of the variation only detected for flowering date, maturity date, shattering, and pubescence type. The 97 QTL were distributed across all 20 chromosomes within 36 genomic regions. These findings identified additional QTL not detected in previous studies using smaller populations while also confirming the quantitative nature for several of the important domestication related traits in soybeans.

Additionally, we have isolated a gene controlling soybean seed coat bloom and seed oil content. We have also fine-mapped a gene associated with leaf shape, which appears to be an important yield component.

The domestication QTL regions have also been compared with the previously identified over 200 large selective sweeps (i.e., genomic regions showing dramatic reduction in genetic diversity in
G. max compared with overall genetic diversity along individual chromosomes). Some of these selective sweeps overlaps with domestication QTLs and some of the selective sweeps do not. Next, we will coordinate with the UIUC team to integrate these comparative genomics data with the yield data from the 416 lines to identify genomic regions positively or negatively associated with yields, and to understand how the domestication-related QTLs may affect yields, and which of the selective sweep regions may be related to yields. In an attempt to identify/validate genetic variation responsible high yields, the genomic regions associated yields as defined by analysis of the 416 G. soja derived lines will be examined in the 500 highly diverse PI accessions described in Objective 3 – Task 2.

We found three QTL that were related to both lodging and stem diameter, which would be expected; but there were three lodging QTL that were not related to either main stem length or stem diameter. These may be related to specific characteristics that affect stem strength.

Introgressions from PI 479752 were not randomly distributed across the genome, but were detected in regions of low and high frequency. Only a few regions contained *G. soja* alleles at frequencies near or above 50%, predominantly on chr 3, 4, 7, 8, 15, and 18. Areas of higher *G. soja* frequency most often occurred in regions without domestication-related QTL. This would indicate that these regions contain neutral or potentially positive alleles. With a few exceptions, all 32 regions containing DRT QTL were located at or adjacent to regions of low *G. soja* allele frequency. Regions with DRT QTL with relatively high frequency of *G. soja* alleles were QTL that had small effects.

We have begun the process of identifying experimental lines with *G. soja* introgressions close to the regions of DRT QTL in lines that do not have the *G. soja* phenotype for the DRT. These lines are likely to continue introgressions from *G. soja* that were lost during domestication. For example, *qSH-16* is a major QTL that explains 53% variance in shattering. Only two lines contained *G. soja* alleles at this locus, and they were also the most shattering susceptible. A total of 20 lines contained *G. soja* introgressions within 500 kb of the *qSH-16* locus, and all displayed low levels of shattering. Two lines with introgressions adjacent to *qSH-16* yielded the same as Williams 82. These lines could be useful for increasing diversity within the selective sweep at *qSH-16* without introducing undesirable traits.

Selection for 100-seed weight occurred only indirectly as agronomically good

phenotypes would generally have larger seeds so recovering the 100-seed weight of the soybean parent was more difficult than reducing shattering. Lines from the Williams 82 x PI 479752 pedigree averaged 9

g/100 seeds, considerably less than the 16.1g 100-seed weight of Williams 82. Unlike shattering, 100-seed weight is controlled by many QTL with small to moderate effects. This caused weaker selection against negative *G. soja* alleles at the seed weight QTL. While *qSW-17-1* accounted for the largest phenotypic variance in the Williams 82 x PI 479752 mapping population, the *G. soja* allele at that locus still persisted in about 25% of the lines. From Williams 82 x PI 479752, 31 lines had 100-seed weights of 10 g or more. Of these, eight contained *G. soja* introgressions within one or more of the major seed weight QTL, *qSW-12, qSW-17-1,* or *qSW-19*. Two lines carried *G. soja* alleles at two of the three seed weight QTL while still maintaining 100-seed weights of 10 g.

Task 6: Map yield QTL in G. tomentella*-derived lines*

This summer we are growing two tests. A maturity group II test has 225 entries derived from Dwight x LG11-12313. LG11-12313 is from Dwight (5) x PI 441001 and in tests at 12 locations yielded 5.9 bu/a more than Dwight and was only one day later in maturity. A maturity group III test has 225 entries derived from Dwight x LG11-3187. LG11-3187 is from Dwight (4) x PI 441001 and in tests at 15 locations yielded 7.6 bu/a more than Dwight and was six day later in maturity. Both tests are being at grown at 9 locations. These tests are designed to identify the specific introgressions from *G. tomentella* that are associated with these increasing in yield.

Task 7: Development of breeding lines from perennial Glycine

This summer we are yield testing 186 new lines derived from crossing with *G. tomentella* PI 441001. Approximately 1/3 of the lines have *G. tomentella* as the female parent and thus *G. tomentella* cytoplasm.

Because of a hiring freeze within ARS, we were unable to make a new hire to continue the research on making successful crosses between soybean and other perennial *Glycine* species. Some cultures are still being maintained but most of the research is on hold. The new person has been selected and will be brought into the project as soon as the hiring freeze is lifted.

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

Task 2: Communication and outreach.

A group of students was engaged to address the question of genetic gains and what it means to farmers. This question was answered by carrying out interviews with farmers. The students are building a video utilizing these videos to explain genetic gain. This video will be hosted on the NCSRP website and cross-hosted on the SOYGEN website. The take-home message from the interviews was that farmers believe that breeding is primarily maintenance breeding and have not resulted in yield gains *per se*.

Task 3: Engage a commercial plant breeding organization to participate in this project.

Unfortunately, a collaborator at Syngenta has left. Following the merger of Syngenta with sim- china, data management is not high priority. However, it is still

expected that Syngenta will engage with this project.