Genome Wide Association Mapping and Genomic Selection for Agronomic and Disease Traits in Soybean

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Yong Bao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Nevin D. Young and James H. Orf

September 2014

Acknowledgements

I would first like to give sincere thanks to Drs. Nevin Young and Jim Orf, who offered me the opportunity to pursue a Ph.D. degree in Plant Breeding and Genetics at the University of Minnesota. And I would have never gone through the whole process without the unyielding support from both of them. I thank my graduate committee: Drs. Rex Bernardo, Peter Tiffin, Jim Kurle, and Senyu Chen for their valuable contribution to my thesis development, and insightful and constructive feedbacks during the course of my Ph.D. study. The appreciation also goes to Drs. Tri Vuong, Henry Nguyen, and Clinton Meinhardt for their generous help with screening soybean lines against soybean cyst nematodes for us at the University of Missouri. I feel so lucky to have the talented and great folks to work with me on my thesis projects: Roxanne Denny and Dr. Kevin Silverstein in genomics and bioinformatics lab, Gerald Decker and Darcy Weston in soybean breeding group, Grace Anderson in soybean pathology lab, and many other members from each group. Thank you all for your support, communication, and companionship. Thank faculty and staff at Department of Agronomy and Plant Genetics: Drs. Jim Anderson, Craig Sheaffer, Don Wyse, and Lynne Medgaarden.

I appreciate Minnesota Soybean Research & Promotion Council for its financial support in my thesis research and graduate assistantship. Additional financial support from Jean and Mary Lambert Agronomy and Plant Genetics Fellowship, and travel grants from The Microbial and Plant Genomics Institute, Graduate School Metric Fellowship, and Summer Institute of Statistical Genetics are greatly appreciated too.

Above all, the people I will never forget are my parents. Probably they will never read this, but I know my debts to them are beyond measure. My illiterate father save up enough money to send me to school, and the education changed my life. I also feel grateful to my wife, who’s been a close partner and steadfast supporter with every step of the way.

ABSTRACT

Genome-wide association mapping and genomic selection are two emerging genomic approaches for investigating genetic architecture and improving breeding efficiency for complex traits in crop species. The objectives of our study were to: 1) dissect the genetic basis of resistance to soybean cyst nematode (SCN) and sudden death syndrome (SDS) through association mapping (AM) and 2) evaluate genomic selection (GS) as an improved marker-based selection tool for predicting agronomic and disease traits in a public soybean breeding program. For AM, we genotyped 282 common breeding parents from the University of Minnesota soybean breeding program using a genome-wide panel of 1,536 single nucleotide polymorphism (SNP) markers and evaluated plant responses to SCN and SDS in the greenhouse. AM rediscovered reported resistance genes (*rhg1* and *FGAM1* for SCN resistance; c*qSDS001*, c*qRfs4*, and *SDS11-2* for SDS resistance) and also identified novel loci. For GS, average prediction accuracy through cross-validation studies was 0.67 for SCN resistance and 0.64 for root lesion severity associated with SDS resistance. We also empirically assessed the prediction accuracy and responses to GS for agronomic traits. Soybean lines in the AM panel were used as a training set and a validation set consisting of 273 breeding lines were selected from the ongoing breeding program. Existing historical trial data were used to train the GS model. GS was then conducted to select the top 20% individuals from the validation set based on a comprehensive consideration including genomic estimated breeding values. Our GS model predicted yield with a significant positive accuracy in only two MN x MN crosses, while the prediction accuracy was near to zero or negative for protein and oil, and for the

rest of crosses. Moreover, one generation of GS didn't significantly change the population mean of yield, seed protein and oil content. Overall, our study suggested AM holds promise to be used as an alternative approach for mapping QTL in soybean breeding germplasm, and GS deserves further investigation prior to implementation in genetic improvement in existing soybean breeding programs.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 1

Potential of Association Mapping and Genomic Selection to Explore PI88788 Derived Soybean Cyst Nematode Resistance

The potential of association mapping (AM) and genomic selection (GS) have not yet been explored for investigating resistance to soybean cyst nematode (SCN), the most destructive pest affecting soybean. We genotyped 282 representative accessions from the University of Minnesota soybean breeding program using a genome-wide panel of 1,536 single nucleotide polymorphism (SNP) markers and evaluated plant responses to SCN HG type 0. After adjusting for population structure, AM detected significant signals at two loci corresponding to *rhg1* and *FGAM1* plus a third locus located at the opposite end of chromosome 18. Our analysis also identified a discontinuous long-range haplotype of over 600 kb around rhg1 locus associated with resistance to SCN HG type 0. The same phenotypic and genotypic datasets were then used to access GS accuracy for prediction of SCN resistance in the presence of major genes through a six-fold cross-validation study. GS using the full marker set produced average prediction accuracy ranging from 0.59 to 0.67 for SCN resistance, significantly more accurate than marker-assisted selection (MAS) strategies using two *rhg1*-associated DNA makers. Reducing the number of markers to 288 SNPs in the GS training population had little effect on genomic prediction accuracy. This study demonstrates that AM can be an effective genomic tool for identifying genes of interest in diverse germplasm. The results also indicate that improved MAS and GS can enhance breeding efficiency for SCN resistance in existing soybean improvement programs.

Introduction

Soybean (*Glycine max* (L.) Merr.) is the world's foremost source of vegetable protein and oil with total crop value exceeding $43.2 billion in the U.S. in 2012 (*www.soystats.com*). However, soybean is a host to several challenging pathogens and pests. Soybean cyst nematode (SCN) (*Heterodera glycines* Ichinohe) is a highly recalcitrant endo-parasite of roots that causes the most damaging disease in soybean (Koenning and Wrather, 2010). Planting SCN-resistant soybean cultivars in combination with crop rotation is the principal way of managing SCN (Chen et al., 2001). Consequently, a more thorough understanding of the genetic basis of SCN resistance will enable soybean scientists to develop more effective resistant cultivars and improved marker-based selection strategies can accelerate the breeding process for complex quantitative traits like SCN resistance.

Previous bi-parental mapping studies have identified a total of 164 quantitative trait loci (QTL), many only weakly supported, conferring SCN resistance (*www.soybase.org*). *Rhg1* on chromosome 18 and *Rhg4* on chromosome 8 were the two genes repeatedly mapped in multiple resistance accessions in bi-parental populations (reviewed by Concibido et al., 2004). However, a major limitation with this type of genetic mapping is that it only captures a portion of soybean's overall genetic variability in SCN resistance because it is based on limited numbers of parents. With the advance of high throughput, cost-effective marker genotyping platforms, association mapping (AM) has proved to be a powerful genomic tool for whole genome analysis and genetic dissection of complex traits in crop species (Huang et al., 2010; Jia et al., 2013; Li et al., 2013; Mamidi et al.,

2011). Given high-density marker panels, AM provides an opportunity to identify QTL at a higher mapping resolution by taking advantage of historical linkage disequilibrium (LD) with diverse germplasm collections including unstructured populations from breeding programs (Asoro et al., 2013; Mamidi et al., 2011; Sukumaran et al., 2012). In the University of Minnesota soybean breeding program, *rhg1* is the only known resistance gene characterized and deployed in the breeding germplasm collection. Potentially, AM can identify novel resistance loci in addition to the known *rhg1* through more precise dissection of genetic architecture of SCN resistance not possible in the previous bi-parental populations. Compared with out-crossing species, fewer markers are needed in AM to cover the entire genome of self-crossing species such as soybean because LD extends over a longer distance (Lam et al., 2010). One of the constraints to AM is the existence of subpopulations in a mapping population, which can cause spurious associations when trait variation is correlated with subpopulation structure. Mixed linear models have been developed and applied to AM to reduce the number of the false positive associations caused by population structure and relatedness (Yu et al., 2006; Zhang et al., 2010).

To deploy the major SCN resistance genes in breeding germplasm, several molecular markers have been developed and implemented in breeding programs by means of marker-assisted selection (MAS) (Concibido et al., 1996; Cregan et al., 1999; Mudge et al., 1997). Although the use of molecular markers can potentially reduce resources and time needed for breeding disease resistance (Young, 1999), only QTL with large effects are selected and ultimately deployed in improved cultivars. Additionally, the

intensive use of large-effect resistance QTL can shift the virulence phenotypes of nematode populations to overcome resistance. Therefore, an alternative marker-based selection strategy that accesses a broader range of variation while breeding for durable SCN resistance is highly desirable. Genomic selection (GS) is a promising selection method for complex traits based on the use of all marker information to capture the genetic variance generated by numerous small-effect loci (Bernardo and Yu, 2007; Meuwissen et al., 2001). GS involves two phases: model training and prediction. In training phase, data mining algorithms are employed to train a GS model by fitting both phenotypic and genotypic data. In the prediction phase, genomic estimated breeding values (GEBVs) of experimental breeding lines can be calculated using only genotypic data. These GEBVs are then used to select the individual breeding lines for advancement in the breeding cycle. Since 2007, a rapidly increasing number of studies have evaluated the performance of GS for important traits in crop species. In maize (*Zea mays* L.), barley (*Hordeum vulgare L.*), and *Arabidopsis thaliana* datasets, the accuracy of predicting genotypic values was consistently higher with GS than with QTL-based selection (Lorenzana and Bernardo, 2009). Heffner et al. (2011) found that average prediction accuracies using GS could be 28% greater than with MAS and were 95% as accurate as phenotypic selection for single traits in wheat (*Triticum aestivum* L.). The advantage of genomic over phenotypic prediction of traits in oats (*Avena sativa* L.) was larger under lower heritability and a larger training dataset (Asoro et al., 2011). More recently, Lorenz et al. (2012) and Rutkoski et al. (2012) evaluated genomic prediction models for *Fusarium* head blight resistance in barley and wheat, respectively.

The objectives of this project were to apply two state-of-the-art genomic approaches, AM and GS, in order to: 1) explore the diversity of SCN resistance currently present in a public soybean breeding germplasm by assembling a representative collection of soybean accessions; 2) identify novel resistance loci beyond the known *rhg1* gene, which might contribute to breeders' efforts in improving SCN resistance in soybean breeding; 3) evaluate the potential of GS as an improved breeding approach for complex disease resistance in soybean.

Material and Methods

Germplasm and Genotyping

After pedigree analysis of historical crosses collected from the University of Minnesota soybean breeding program using PediTree (van Berloo and Hutten, 2005), we selected a panel of 282 representative accessions based on "footprint values" that were calculated as the sum of the weighted contribution of individual accessions. The selected accessions included ancestral lines, plant introductions, elite lines, advanced breeding lines, and released public cultivars (Table S1). We conducted both AM and GS with the same set of 282 accessions. All selected accessions were planted in the field to increase and generate pure seeds in 2012. DNA was extracted from young leaves of each accession using DNeasy 96 Plant Kit (QIAGEN, Valencia, California). We genotyped DNA samples using an Illumina GoldenGate SNP assay with the Universal Soy Linkage Panel 1.0 (Hyten et al., 2010). SNPs with greater than 5% minor allele frequency (MAF) and a missing data rate less than 50% were retained, followed by imputation of missing SNP

data based on population mean of each marker. Total 1,247 SNP markers passed the filters and were used in the subsequent analysis.

Phenotyping

We performed a greenhouse assay of plant responses to SCN HG type 0 at the National Center for Soybean Biotechnology, University of Missouri in spring 2013. The greenhouse experiment was run twice with five plants for each accession in each run. Experiments were conducted according to the Standardized Cyst Evaluation 2008 protocol (Niblack et al., 2009). SCN resistance was determined by calculating the female index (FI) using "Hutcheson" as susceptible control line (Schmitt and Shannon, 1992). To process samples efficiently, we utilized a fluorescence-based scanner and imaging software to count cysts in run 1 (Brown et al., 2010), and counted cysts in run 2 using a stereomicroscope. Previous studies had proved the method as reliable and robust evaluation of SCN resistance (Guo et al., 2005). We then fitted FI estimates into a linear model: $y_{ij} = u + g_i + r_j + g_i*r_j + \varepsilon_{ij}$, and performed the analysis of variation (ANOVA) with basic package "anova" in R (R Development Core Team, 2010), where y_{ij} was the FI of each plant, u was the intercept, g_i was the genetic value of i_{th} accession, r_j was the mean effect of j_{th} run, g_i*r_j was the interactive effect of i_{th} accession and j_{th} run, and ε_{ij} was the residual. We represented the phenotypic value of each accession as the mean of FI across ten individual plants.

Association Mapping

We first assessed population structure by a model-based approach as implemented in *STRUCTURE* (Pritchard et al., 2000) and principal component analysis (PCA) implemented in TASSEL (Bradbury et al., 2007). To avoid overestimation of subpopulation divergence due to tightly linked SNP markers (Falush et al., 2003), two subsets of 227 SNP markers with approximately 10 cM spacing were chosen for use in *STRUCTURE*. For each marker subset, k=1 to 10 subgroups with each k value were modeled 10 times with a burn-in period and number of replications equal to 10,000 using an admixture model in *STRUCTURE*. The optimal k was then determined based on the rate of $\ln Pr(X|k)$ change from k-*1* to k. Soybean accessions were assigned to subpopulations based on the highest mean of membership probability. All SNP markers were used to investigate the population structure with PCA implemented in TASSEL (Bradbury et al., 2007). In addition, we estimated the extent of LD and illustrated the pair-wise measures of LD (r^2) as a heatmap using *Haploview4.2* at the genomic region of interest including *rhg1* and *FGAM1* (Barrett et al., 2005). We performed the AM for SCN resistance with Kinship (K) and Population Structure + Kinship (Q+K) models, respectively, in the "rrBLUP" package (Endelman, 2011) in R (R Development Core Team, 2010). In Q+K model, the first three principal components from PCA were used to control the population structure. A false discovery rate (FDR) of 0.05 was used to control for false positive associations in AM. Manhattan plots were created based on the AM results with SNPEVG (Wang et al., 2012). By assuming the identified candidate genes act additively, a forward stepwise linear regression model with the FI estimates as dependent variables and the significant SNP markers as explanatory variables was

constructed in R (R Development Core Team, 2010). Adjusted R^2 values were estimated from the linear regression model representing the percentage of phenotypic variation explained by the candidate genes. These R^2 values were likely up-biased estimations because the population structure and relatedness were not adjusted and these significant SNPs were pre-selected in the simple linear regression model.

Genomic Selection Model

Besides identifying candidate genes, the same set of phenotypic and genetic data was utilized to access the genomic prediction accuracy for SCN resistance. The GS models we employed were: ridge-regression best linear unbiased prediction (RR) (Bernardo and Yu, 2007; Endelman, 2011; Meuwissen et al., 2001), Bayesian LASSO regression (BLR) (de los Campos et al., 2009; Park and Casella, 2008; Pérez et al., 2010), Bayes Cπ (BCP) (Habier et al., 2011), support vector machine (SVM) (Long et al., 2011), and random forest (RF) (González-Recio and Forni, 2011). The RR, BLR, SVM, and RF were implemented with the respective package "rrBLUP", "BLR", "Kernlab" and "randomForest" in R (R Development Core Team, 2010). The BCP was implemented with *GenSel* software (*BIGS.ansci.iastate.edu*). Specifically, a total of 10,000 burn-ins and 40,000 saved iterations of Markov-Chain Monte Carlo (MCMC) were used in both BLR and BCP; "vanilladot" kernel and "eps-svr" type were used in SVM; and 500 trees and 4 branches were used in RF. All other parameters in each model were adopted from the guidelines and examples in the corresponding references and R package description. To account for major resistance genes identified through AM analysis, we constructed a

ridge-regression best linear unbiased prediction with major genes fitted as fixed effects (RRF) model by treating the significant SNPs tagging the major genes as fixed effects while the rest of SNPs as random effects. For each fold in the subsequent cross-validation study, the significant SNP markers were first detected using the AM analysis within training set and the significant SNPs then fitted as fixed effect in the RRF model. Additionally, we extended the RRF model to FGAM, rhg, and rhg+FGAM models by fixing the significant SNP(s) tagging the corresponding gene as fixed effects.

Marker Assisted Selection Model

To set up reference methods to compare with, we constructed two MAS models: multiple linear regression model fitted with two *rhg1*-associated SNP markers as fixed effect to represent conventional MAS (cMAS) approach frequently implemented in the soybean breeding program for SCN resistance; and multiple linear regression model fitted with 35 top SNP markers selected based on their association with phenotypic variation as fixed effect to represent an improved MAS approach (iMAS). For each fold in the subsequent cross-validation study, SNP effects were estimated from the linear regression models in the training set and then used to predict phenotype of SCN resistance in the validation set. Because our AM study had identified four significant SNP markers tagging *rhg1* gene, all six possible pairs of SNPs were sampled from the four and fitted as fixed effect in the cMAS model. The average of prediction accuracy from all six pairs was estimated. For the iMAS model, the top 35 SNP markers detected using the AM analysis within training

set in each fold were fitted as fixed effect in the model. The linear regression analysis was performed with basic R package “lm” (R Development Core Team, 2010).

Cross Validation

We conducted a 6-fold cross-validation study to avoid inflated estimation of the prediction accuracy of GS and MAS for SCN resistance. All soybean accessions first were randomly divided into six subsets. In each fold, five subsets of lines were used as training sets and the remaining subset was a validation set. Marker effects were estimated from eight different models by fitting the genotypic and phenotypic data in the training set. Marker-based prediction of line performance in the validation set was calculated by summing all marker effects of that line using only genotypic data. Prediction accuracy was calculated as the correlation between marker-based prediction and phenotypic values. The mean, standard deviation, and confidence interval of the six folds were calculated for each model. One-tail paired t-Test was used to compare the prediction performance of all assayed models to reference models cMAS and iMAS, respectively.

Marker Number

We also determined the effect of marker numbers on GS accuracy through 6-fold cross-validation by including random samples of 96, 192, 288, 384, 768, and 1152 SNPs from the full marker set in the respective RR and RRF model. Within each fold, this was repeated 100 times to avoid sampling bias for markers and the average of 100 replications was used to represent the prediction accuracy of each fold. The mean of the six folds was

calculated. All prediction accuracies were estimated with R package "rrBLUP" (R Development Core Team, 2010).

Results

Phenotypic Analysis

The assayed soybean accessions had a mean FI=84 with a range from 2 to 154 (Fig. 1). ANOVA for the FI indicated that the effect of accession, run, and accession by run had significant effects (Table 1). The vast majority of accessions were moderately to completely susceptible to SCN HG type 0, with only 11 accessions being highly resistant (FI < 10) and 8 accessions being moderately resistant (10 < FI < 30) (Fig. 1). All 11 resistant accessions share a single resistance source of PI 88788 with no other resistance sources uncovered through pedigree searching.

Marker Profile, Population Structure

Distances between adjacent markers ranged from 0 to 29.4 cM with a mean of 1.4 cM (Table S2). Approximately 80% of adjacent markers were within 2 cM of one another and only 6% were > 5 cM apart (Table S2). Among a total of 1,521 polymorphic markers, 69 markers had MAF < 0.01, and 232 markers had MAF < 0.05 (Fig. S1).

Each subset used in *STRUCTURE* included a random set of 227 SNP markers spanning all 20 linkage groups with an average gap size of 10 cM. Plots of natural logarithm probability difference k and k-1 (Δk) against k were similar in the pattern for two subsets (Fig. S2). We observed the first rapid drop in posterior probability from k=3

to *k*=4 in both subsets suggesting the presence of three theoretical subpopulations in our germplasm pool (Fig. S2). The pair-wise PCA also suggested a pattern of three clusters in the germplasm pool (Fig. 2). Based on the incomplete pedigree information and breeder's knowledge, we identified three distinct genetic backgrounds corresponding to the three subpopulations, namely, "High Protein", "High Yield", and "Small Seeds" (Fig. 2). Specifically, subpopulation "High Protein" contains mostly University-released high-protein specialty cultivars and high protein ancestors such as Kasota, Kato, Proto, and Toyopro. Subpopulation "High Yield" contains mostly University-released commodity cultivars and North American elites such as Agassiz, Alpha, Amsoy, Archer, Bell, Capital, Chico, Clay, Evans, Freeborn, McCall, and Traill. Subpopulation "Small Seeds" contains largely University-released food-type cultivars such as Minnatto and numerous Plant Introduction (PI) lines. These three groups were approximately coincident with the three major categories of breeding materials the University of Minnesota soybean breeding program has created in the history. Soybean accessions with moderate to high resistance to SCN were all included in subpopulation "High Yield" (Fig. 2), indicating the breeder's historical efforts of stacking SCN resistance with high yield commodity cultivars.

Association Mapping

Since the results from both *STRUCTURE* and PCA indicated the presence of three theoretical subpopulations in our germplasm pool, we fixed the first three PCs in Q+K model. The quantile-quantile plots showed the Q+K model performed slightly better than

the K model for controlling type I error caused by population stratification (Fig. S3). Thus, the Q+K model was implemented in AM for SCN resistance in the subsequent analysis. We detected a total of six significant SNPs, all on chromosome 18, in AM with FDR of 0.05, but observed no additional significant SNPs on any other chromosome. The linear regression using the six significant SNPs as the explanatory variables collectively explained 49% of phenotypic variation, which was likely biased upward by using the pre-selected SNPs based on their significant association with phenotypes in AM (Stanton-Geddes et al., 2013).

Four of the six highly significant SNPs were located 3 kb to 258 kb away from the center of *rhg1* gene that has been reported to confer the resistance to SCN previously in numerous studies (Table 2; Fig. 3; reviewed by Concibido et al., 2004). The four significant SNPs were in a cluster of high LD (Fig. 4). By comparing the haplotypes among resistant and susceptible lines, we identified a > 600 kb haplotype block of SNP markers including the four significant ones that appear consistent with intensive selection for SCN resistance in modern breeding efforts (Table 3).

SNP (BARC-047665-10370) with the $-log_{10}(P)$ value of 11.92 (Table 2) was located in the coding region of *FGAM1* gene, which has previously been shown to exhibit differential gene expression in response to SCN feeding (Vaghchhipawala et al., 2004). This SNP was not in the same haplotype block with *rgh1* (Table 3, haplotype blocks defined by complete LD between markers), but it was in the strong LD ($r^2 = \sim 0.8$) with all significant SNP markers at *rhg1* (Fig. 4).

Additionally, we identified a significant SNP (BARC-019001-03050) at *Glyma18g46201* on the opposite end of chromosome 18 that is predicted to encode a ring finger protein (Table 2; Fig. 3). BARC-019001-03050 was in the genomic regions significantly associated with PI567516c (Vuong et al., 2010) and PI 209332 resistance (Concibido et al., 1996) in previous studies.

With less stringent cutting-off –logp value of 2, we identified an additional signal that includes two nearly significant SNPs on chromosome 7: BARC-04209-07684 and BARC-028385-05858 (Fig. 3). No apparent nucleotide-binding site-leucine-rich repeat (NBS-LRR) or other potential resistance genes are known to be located nearby (Schmutz et al., 2010).

Prediction Accuracy

The same set of phenotypic and genotypic data used in the AM analyses was used to assess the genomic prediction accuracy for SCN resistance through a 6-fold cross-validation. The prediction accuracy for SCN resistance with the RR model using the full marker set ranged from 0.44 to 0.80 with a mean of 0.66 (Table 4). When major genes were fitted as fixed effects, the RRF model resulted in a mean of prediction accuracy of 0.61, not significantly different from any other GS models (Table 4). Three to eight SNP markers were detected as significant through AM analysis in the training set in each fold, and subsequently fitted as fixed effects in the RRF model. The signals at *rhg1* and *FGAM1* were always significant across 6 folds. Both BLR and BCP models assume that only a small portion of loci are causal loci with large effects and non-causal loci have

infinitesimal to no effect. The prediction accuracy ranged from 0.59 to 0.75 with a mean of 0.67 in the BLR model and ranged from 0.48 to 0.71 with a mean of 0.62 in the BCP model (Table 4). The SVM and RF models with capacity to capture non-additive sources of genetic variability including dominance and epistasis did not outperform the additive linear regression models neither in the case of SCN resistance (Table 4).

We also compared the six GS models with two reference MAS models: conventional MAS (cMAS) model represented by multiple linear regression fitted with two *rhg1*-associated SNP markers, and improved MAS (iMAS) model represented by multiple linear regression fitted with 35 top SNP markers detected through AM within training set. The mean prediction accuracy of cMAS model was only 0.49, significantly lower than that of all GS models and iMAS model (Table 4). All GS models and iMAS model performed equivalently (Table 4).

To further determine the sufficient number of markers for sustaining high accuracy and meanwhile reducing genotyping cost, we compared the prediction accuracy estimated from RR, rhg, FGAM, and rgh+FGAM models with different sizes of marker set. The prediction accuracy generally increased as the number of SNP markers increased, and the gain in accuracy became minimal when more than 288 SNPs were used (Fig. 5). The rhg+FGAM model using as few as 96 genome-wide SNPs as random effects produced the prediction accuracy of 0.60, while the RR model was only 0.50 (Fig. 5). Using more markers substantially increased the prediction accuracy in the RR model, but had little effect in the rhg+FGAM model (Fig. 5).

Discussion

Limitations of the SNP Array

The relatively low SNP density on our genotyping panel might have limited the power of AM to identify causal variants or even generated biased results, so our study should be interpreted with caution. With only 1,247 polymorphic SNPs spread across the approximately 1,100 Mb soybean genome, markers generally would not be in complete LD with all the causal genes controlling variation in SCN resistance. Stanton-Geddes et al. (2013) empirically compared the AM gene candidates identified with resequencing data in *Medicago truncatula* versus reduced-representation SNP arrays and showed that SNP arrays could bias AM results. With more recently developed high-density SNP chip (Song et al., 2013) and sequencing-based genotyping including genotyping-by-sequencing (Elshire et al., 2011; Xu et al., 2013), AM should enable improved dissection of genetic variation and pinpoint causal variants more accurately in future investigations.

The average significant LD extent was 29.3 cM based on the marker panel used in this population (Data not shown). LD was extensive with long range LD (>20 cM) observed in over half of all significant pair-wise intra-chromosomal LD (Table S3). This agrees with the LD pattern identified in soybean with resequencing data by Lam et al. (2010). Extremely high LD is one of the distinctive characteristics of soybean genome compared with most other crop species. In a high LD species, less number of markers is needed for mapping and MAS while more linkage drag and low mapping resolution are expected.

SCN Resistance Genes

The robustness of AM still enabled us to rediscover the SCN resistance gene: *rhg1* using moderate-density SNP markers in a diverse set of breeding germplasm. In previous linkage mapping studies, *rhg1* on chromosome 18 was repeatedly mapped in resistant accessions such as Peking, PI 88788, PI 209332, PI 437654, etc., while second gene, *Rhg4* on chromosome 8, was identified only in Peking, PI 209332 and PI 437654 (reviewed by Concibido et al., 2004). Kim et al. (2010) pinpointed the *rhg1* locus in a 67 kb genomic region on chromosome 18. More recently, Cook et al. (2012) performed fine mapping at *rhg1* locus and found that copy number variation (CNV) of a genomic segment spanning three genes was required to confer resistance.

By contrast, we failed to observe *Rhg4* in our AM, presumably due to the lack of resistant accessions with Peking or other sources distinct from PI 88788 in the germplasm pool. We performed a pedigree search for all soybean lines with FI < 10 in our germplasm pool and only identified PI 88788 in their ancestry, but no other sources of resistance. To reveal the genetic basis of SCN resistance more thoroughly in further AM studies, larger number of soybean accessions including PI lines with resistance sources distinct from PI 88788 is essential. Another possibility could be the confounding effects of two effective *rhg* genes, because we only evaluated plant response to SCN HG type 0 that is avirulent to both PI 88788 and Peking resistance sources. We might expect different responses if plants were inoculated with SCN HG type 2, which is virulent to PI 88788, but avirulent to Peking.

In addition to the strong signal at *rhg1* locus, we identified a significant SNP with MAF of 8% within the coding region of *FGAM1* gene. FGAM synthase expression has been shown to respond to nematode feeding in the *Arabidopsis thaliana – Heterodera schachtii* system (Vaghchhipawala et al., 2004). Interestingly, *rhg1* was reported to disrupt the formation and/or maintenance of nematode feeding sites (Niblack et al., 2006), while the *FGAM* promoter redirected the gene expression within feeding sites to benefit the nematodes (Vaghchhipawala et al., 2004). Although the *FGAM1* gene resides about 1.1 Mb away from *rhg1* gene, its co-localized SNP exhibited extensive LD (r^2 = ~0.8) with almost all *rhg1*-associated SNPs in the population (Fig. 4). To determine whether this significant SNP detected at *FGAM1* was actually tagging the *rhg1* locus, we tested a mixed model fitting *rhg1* locus as a fixed effect in the AM analyses, and found the SNP at *FGAM1* locus was still significantly associated with SCN resistance (Figure S4). This suggests that the *FGAM1* locus independently accounts for a considerable amount of SCN resistance variation in our germplasm panel.

The significant SNP identified at the opposite end of chromosome 18 was in the non-coding region of gene *Glyma18g46201* annotated as ring finger protein (Schmutz et al., 2010). Although no apparent nucleotide-binding site-leucine-rich repeat (NBS-LRR) or other potential resistance genes are known to be located nearby (Schmutz et al., 2010), BARC-019001-03050 was in the genomic regions which had been reported significantly associated with PI567516c (Vuong et al., 2010) and PI 209332 resistance (Concibido et al., 1996) in previous studies. It is still possible that this SNP may tag a distant causal gene within the nearby resistance QTL interval or other uncharacterized genomic regions

associated with SCN resistance. The candidate gene *Glyma18g46201* deserves further confirmation and molecular characterization.

Genomic Selection Models

More sophisticated but computation-intensive Bayesian and machine learning models did not outperform additive linear models such as RR and RRF in prediction accuracy (Table 4). GS studies conducted in maize, wheat, oat, and barley for both agronomic and disease traits also suggested slight differences among various genomic prediction algorithms (Asoro et al., 2011; Lorenzana and Bernardo, 2009; Lorenz et al., 2012; Rutkoski et al., 2012). Since all soybean lines included in our germplasm pool are homogeneous or near-homogeneous inbred lines, the advantage of SVM and RF models with capacity to capture non-additive sources of genetic variability including dominance and epistasis was not observed in our study (Table 4).

By taking into account of the presence of major genes, the RRF model didn't increase prediction accuracy compared with the RR model (Table 4), consistent with the findings in a simulation study in maize (Bernardo, 2013). In that study, when R^2 was 50%, heritability of trait was 0.5 and population size was 250, similar to the case of SCN resistance we studied here, selection response from GS with major gene(s) fixed was only 6% greater than GS treating major gene(s) as random effects (Bernardo, 2013). Additionally, we compared the RRF models with different major gene(s) fixed and found the higher prediction accuracy with more genes fixed (Fig. 5). The difference of

prediction accuracy among different RRF models was more distinct when few random markers were used for GS modeling (Fig. 5).

Genomic Selection versus Marker Assisted Selection

In soybean breeding populations targeting SCN resistance, MAS is frequently used to evaluate the SCN resistance of breeding candidates by detecting the flanking markers of major resistance genes such as *rhg1* and *Rhg4*. However, breeders might still miss a large proportion of total genetic variation in MAS caused by numerous loci of small-moderate effects. Potentially, this additional variation in SCN response can be captured through the use of genome-wide SNPs implemented through GS. As indicated in Table 4, breeders may be able to improve average prediction accuracy up to 0.67 through the use of all marker information, which is significantly more accurate than the conventional MAS approach with accuracy of only 0.49.

Interestingly, the iMAS model using top 35 SNPs performed equivalently to the RR model using full marker set in our study (Table 4). This is in contrast to the results from empirical GS recurrent selection for yield and stover index in maize and cross-validation study for agronomic traits in wheat (Heffner et al., 2011; Massman et al., 2013). A likely reason for the high predictive accuracy of iMAS in our study could be that two genes of fairly large effect control considerable variation in SCN resistance in our germplasm pool. Among the top 35 SNPs used in iMAS, we observed the 6 significant SNPs detected in the full panel across all the 6 folds. This is consistent with the narrow genetic base of SCN resistance in U.S. breeding programs, with more than

90% of SCN resistant cultivars grown in the Midwest coming from just a single source, PI 88788 (Concibido et al., 2004). Continuously selecting for the same major resistant genes and growing the same resistant cultivars is likely to lead to nematodes that overcome resistance (Mitchum et al., 2007). Thus, introduction of exotic soybean germplasm to improve SCN resistance diversity should remain an important goal in current SCN breeding programs.

In the presence of major genes in a breeding population, the preferred marker-based selection strategy, either MAS or GS, depends on the breeding goal and resource allocation. For example, when soybean breeders would like to select candidate lines with moderate resistance or horizontal resistance besides the *rhg1*-conferring resistance, GS tends to be more accurate than MAS in predicting the quantitative difference in plant responses to SCN. In contrast, the haplotype information we identified (Table 3) appeared to be sufficient for breeders to select for the highly resistant candidates in the breeding population. Since the iMAS model using 35 top SNPs had a mean of prediction accuracy of 0.63, equivalent to the GS model using full marker set (Table 4), the iMAS obviously is more cost-effective and likely implemented in the current breeding program for SCN resistance. However, the iMAS wouldn't be sustainably effective as GS in a high throughput breeding program because many of 35 SNP markers tend to be fixed in breeding germplasm after few cycles of selection. With continuously declining genotyping costs, the iMAS might be replaced by GS considering increased genetic gain per unit cost.

Both iMAS and RRF models are probably germplasm-specific because they only capture the resistance loci currently present in our germplasm pool. In other words, the predictive ability of our model might be limited if applied to a germplasm pool where SCN-relevant loci are distinct from the ones in ours. The model should therefore be updated to integrate the markers associated with the novel resistance gene(s) that the breeders desire to introgress into the germplasm pool. To maintain high prediction accuracy, model updating also applies to GS to ensure that the target resistance genes are represented in GS models.

Implementing MAS/GS in SCN Resistance Breeding

More accurate marker-based prediction, either iMAS or GS, can effectively accelerate SCN resistance breeding by increasing genetic gain per unit time and reducing the need for extensive phenotyping in soybean breeding program. Since the 1950s soybean breeders have consistently introduced production lines with a limited set of exotic sources of SCN resistance to introgress with high-yielding elite lines in soybean breeding programs, including the program at the University of Minnesota. The breeding program has successfully developed and released numerous SCN-resistant soybean cultivars adapted to various soybean growing regions across short maturity groups through conventional MAS followed by extensive field trials. Despite the success in SCN resistance breeding, breeders agree that a cost-effective solution for sustainable resistance is required to meet the challenge of emerging SCN races of more virulence. To achieve the benefits of marker-based selection strategies over phenotypic selection, recurrent

selection for multiple cycles is an effective method to increase genetic gain per unit time for complex traits. Given the high estimate of prediction accuracy, MAS with 35 most important SNPs or GS with hundreds of genome-scale SNPs can be developed as a molecular tool to assist breeders in selecting SCN resistant lines in soybean breeding program. The more accurate prediction of phenotypes allows breeders to increase the selection intensity leading to fewer candidate lines to be tested in the following expensive yield trials.

In an actual breeding program, breeders usually select promising lines based on the evaluations of multiple traits. A multi-trait GS model can be developed by simultaneously fitting phenotypic data from the evaluations of yield, SCN resistance, and other traits of interest as dependent variables in the model. Without the need for including additional markers specifically for SCN resistance, the multi-trait GS model using the existing marker panel leads to simultaneous prediction of genetic values for several traits including SCN resistance. Moreover, the prediction accuracy for low-heritability traits, such as yield, could be improved by taking the advantage of their genetic relationship with high-heritability traits in the multi-trait GS (Jia and Jannink, 2012).

Empirical recurrent selection with GS has been conducted to improve agronomic traits and introgress desired exotic traits to elite lines, exhibiting superiority to phenotypic selection and MAS in cross-pollinated crop species, such as maize (Combs and Bernardo, 2013; Massman et al., 2013). Recurrent selection in self-crossing species like soybean requires laborious pollination to obtain sufficient F_1 seeds in each recombination. Bernardo (2010) proposed a “select-combine-self” scheme for self-crossing crops and

indicated that the selection response from GS with minimal crossing was comparable to maize by intensive use of a year-round nursery. Future studies to empirically determine the genetic gain per unit cost and time in GS or improved MAS compared with conventional MAS in an SCN breeding population will facilitate the adoption of improved marker-based selection strategy in the existing breeding program.

Conclusion

We present the first study using AM to dissect the genetic architecture of SCN resistance and evaluate the ability of GS in predicting SCN resistance. Significant signals were detected at two SCN resistance genes: *rhg1* and *FGAM1*, plus the third locus located at the opposite end of chromosome 18. Estimates of high prediction accuracy suggest that improved MAS and GS has the potential for accurate prediction in SCN resistance breeding. Overall, our results indicate that advanced genomic tools provide insights into the genetic basis of complex disease traits as well as offer the possibility for greater genetic improvement in developing improved cultivars.

Table 1. Analysis of variance (ANOVA) for female index (FI %) on soybean accessions in greenhouse assay.

Source of variation	Df†	MS‡	F value	Pr(>F)
RUN	1	371171.43	119.63	< 0.001
ACCESSION	281	21716.07	7.00	< 0.001
RUN*ACCESSION	280	4971.154	1.60	< 0.001

†Df, degree of freedom.

‡MS, mean of square.

Table 2. The significant SNPs detected from association mapping (AM) for SCN resistance.

Marker	MAF†	Chromosome	Position (bp)	Nearby Gene	-log$_{10}$(P)	R^2
BARC-048271-10520	0.47	18	1705138	*rhg1*	5.04	0.19
BARC-G01477-00243	0.34	18	1710320	*rhg1*	4.34	0.17
BARC-012259-01773	0.16	18	1776719	*rhg1*	4.92	0.18
BARC-012295-01800	0.14	18	1970944	*rhg1*	3.84	0.06
BARC-047665-10370	0.08	18	2833147	*FGAM1*	11.96	0.32
BARC-019001-03050	0.28	18	55961797	*Glyma18g46201*	4.72	0.22

†MAF, minor allele frequency

Table 3. Haplotype analysis of *rhg1* locus conferring resistance to SCN HG type 0.

		BARC-054083-12329	BARC-040479-07752	BARC-012237-01756	BARC-048277-10538	BARC-048275-10534	BARC-G01477-00243	BARC-048271-10520	BARC-048801-10723	BARC-012259-01773	BARC-012289-01799	BARC-012295-01800	BARC-015067-02556	BARC-025777-05064	BARC-047665-10370	BARC-014395-01348
Line†	FI‡	-660	-484	-28	-17	-8	-8	-3§	6	64	245	258	442	584	1120	1735
R	**2**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	GG	AA
R	**2**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
R	**4**	CC	CC	TT	AA	AA	AA	GG	GG	AA	--	CC	GG	GG	AA	GG
R	**5**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
R	**5**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
R	**5**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
R	**6**	CC	CC	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
R	**6**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	AA
R	**7**	GG	GG	TT	AA	AA	AA	GG	AG	AA	AA	CC	GG	GG	GG	--
R	**7**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	GG	GG
R	**7**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	GG	AA
MR	**12**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	AG
MR	**14**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	GG	AA
MR	**14**	GG	--	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
MR	**14**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
MR	**22**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	GG	GG
MR	**23**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	AA
MR	**24**	GG	GG	TT	AA	AA	AA	--	--	AA	AA	CC	GG	GG	GG	AA
MR	**29**	CC	CC	AA	GG	GG	AA	AA	AA	TT	GG	CC	CC	GG	AG	--
MS	**30**	CG	CG	--	--	--	AA	AG	--	AT	AA	CC	CC	GG	GG	GG
MS	**39**	CC	CC	--	--	--	--	--	--	--	AA	--	--	--	--	AA
MS	**42**	GG	GG	TT	GG	GG	GG	AA	GG	TT	GG	CC	CC	GG	AA	GG
MS	**44**	GG	GG	TT	GG	GG	GG	--	AA	--	AA	AA	CC	GG	GG	--
MS	**44**	CC	CC	TT	AA	AA	AA	GG	GG	AA	AA	CC	GG	GG	AA	GG
MS	**44**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	--	CC	GG	GG	GG
MS	**53**	GG	GG	TT	AA	AA	AA	GG	GG	AA	AA	AA	CC	GG	GG	GG
MS	**54**	GG	CC	TT	GG	GG	GG	AA	GG	TT	GG	CC	GG	GG	GG	GG
MS	**57**	CG	CG	TT	GG	GG	GG	AA	AA	TT	AA	AA	CC	GG	GG	AA
MS	**58**	CC	CC	TT	GG	GG	GG	AA	AA	TT	AA	AA	CC	GG	GG	AA
S	**63**	CC	CC	TT	AA	--	AA	GG	AG	--	AA	AA	GG	--	GG	AA
S	**63**	CC	CC	TT	GG	GG	GG	AA	AA	TT	AA	AA	CC	AG	GG	GG
S	**66**	GG	CC	TT	GG	GG	GG	AA	AA	TT	AA	AA	CC	GG	GG	AA
S	**67**	CC	CC	TT	AA	GG	AA	AA	AA	TT	AA	AA	CC	GG	AA	GG
S	**67**	CC	CC	TT	--	--	--	AG	GG	AT	--	--	GG	GG	GG	GG
S	**68**	CC	CC	AA	AA	AA	AA	AA	--	TT	AA	CC	CC	AA	--	AA
S	**68**	CC	CC	TT	AA	AA	AA	GG	GG	AA	AA	AA	CC	GG	AA	GG
S	**68**	CC	CC	TT	GG	GG	GG	AA	GG	TT	GG	CC	GG	GG	GG	AA
S	**68**	GG	GG	TT	GG	GG	GG	AA	AA	TT	AA	AA	CC	GG	GG	AA
S	**69**	CG	CG	TT	GG	GG	GG	AA	GG	TT	GG	CC	CC	GG	AA	GG
S	**69**	GG	GG	TT	GG	GG	GG	AA	AA	TT	AA	--	CC	GG	GG	GG
S	**70**	CC	CC	TT	GG	GG	--	AA	AA	TT	--	AA	CC	AG	GG	AA
S	**70**	CC	CC	TT	GG	GG	GG	AA	GG	TT	GG	CC	GG	GG	GG	AA
S	**70**	GG	GG	TT	GG	GG	GG	AA	AA	TT	--	AA	CC	GG	GG	--
S	**70**	CC	CC	TT	GG	GG	GG	AA	AG	TT	GG	AA	CC	GG	GG	AA

†R, resistant (FI < 10); MR, moderately resistant (10 < FI < 30); MS, moderately susceptible (30 < FI < 60); S, susceptible (FI > 60). None of the accessions with FI > 70 have resistance haplotype, and are not included in this table.

‡FI, female index of SCN HG type 0.

§The values in kb in first row represent the SNPs with varying distances from *rhg1*.

Table 4. The mean, standard deviation and confidence interval of prediction accuracy with various models estimated from 6-fold cross-validation.

Model†	Fold						Mean	SD‡	CI§	p-value¶	
	1	2	3	4	5	6				cMAS	iMAS
cMAS	0.55	0.46	0.56	0.42	0.43	0.51	0.49	0.06	0.05		0.03
iMAS	0.81	0.68	0.54	0.72	0.5	0.51	0.63	0.13	0.10	0.03	
RR	0.77	0.8	0.64	0.64	0.44	0.67	0.66	0.13	0.10	0.01	0.25
RRF	0.76	0.75	0.62	0.59	0.32	0.64	0.61	0.16	0.13	0.04	0.41
BLR	0.64	0.66	0.75	0.65	0.59	0.74	0.67	0.06	0.05	0	0.26
BCP	0.68	0.71	0.68	0.58	0.59	0.48	0.62	0.09	0.07	0.01	0.44
RF	0.74	0.82	0.68	0.63	0.46	0.71	0.67	0.12	0.10	0	0.21
SVM	0.72	0.67	0.58	0.59	0.35	0.62	0.59	0.13	0.10	0.04	0.19

†cMAS, conventional marker assisted selection; iMAS, improved marker assisted selection; RR, ridge-regression best linear unbiased prediction; RRF, ridge-regression best linear unbiased prediction with major genes fixed; BLR, Bayesian linear regression; BCP, Bayesian Cπ; RF, random forest; SVM, support vector machine.

‡SD, standard deviation.

§CI, confidence interval with $\alpha = 0.05$.

¶p-value obtained from one-tail paired t-Test each of assayed models compared to cMAS and iMAS model, respectively.

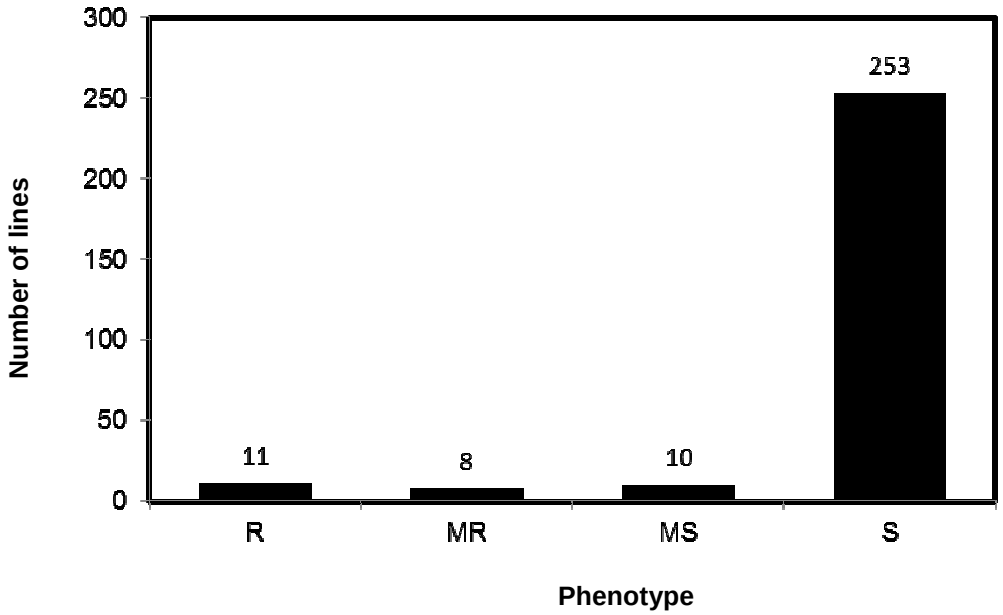

Figure 1. Soybean cyst nematode (SCN) female index (FI %) for 282 soybean accessions
R, resistant (FI <10); MR, moderately resistant (FI < 30); MS, moderately susceptible (30 < FI
60); S, susceptible (FI > 60).

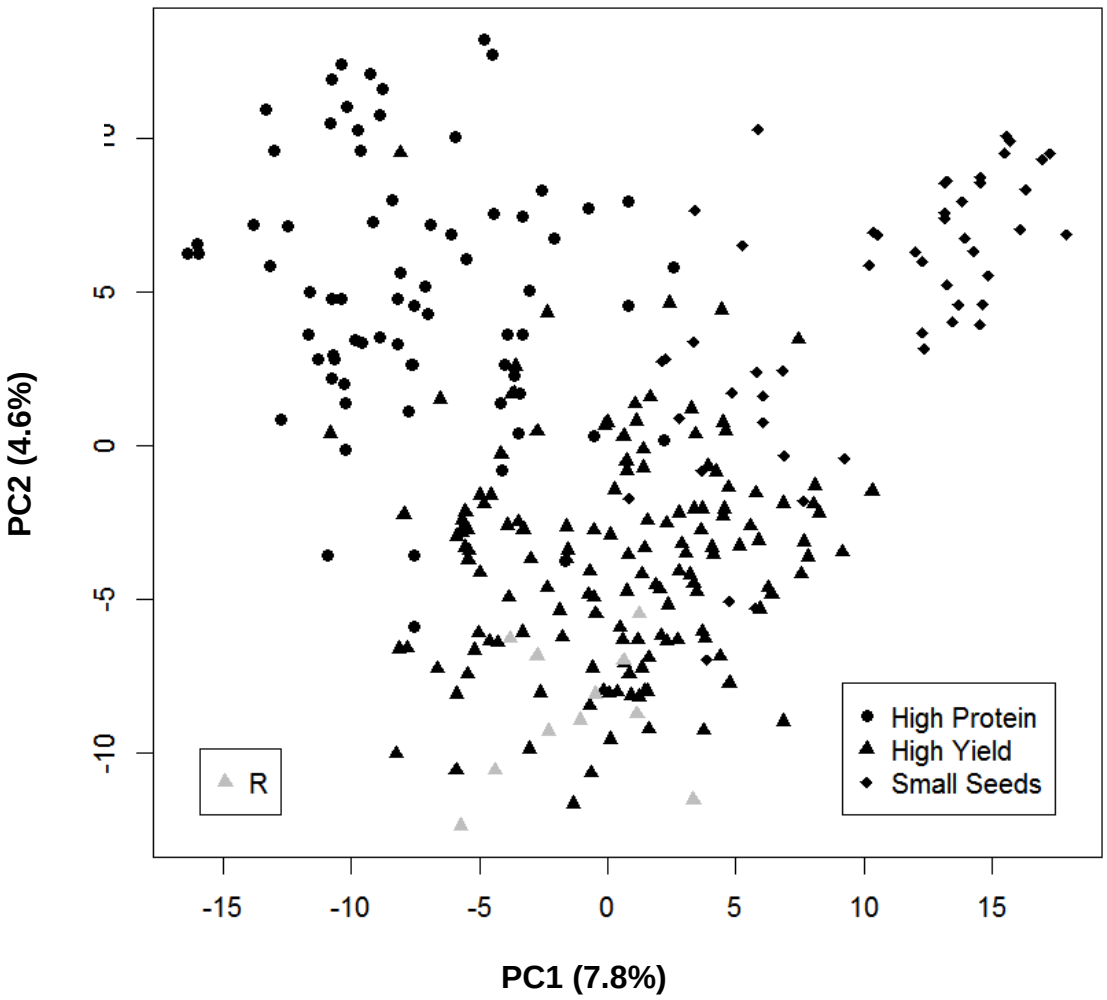

Figure 2. Principal component analysis (PCA) of the germplasm pool. Soybean accessions were assigned to 3 subpopulations based on their highest membership probability estimated in *STRUCTURE*, namely, High Protein, High Yield, and Small Seeds, represented by circles, triangles, and diamonds, respectively. The SCN resistance phenotype: resistant (R) was represented by the color of grey.

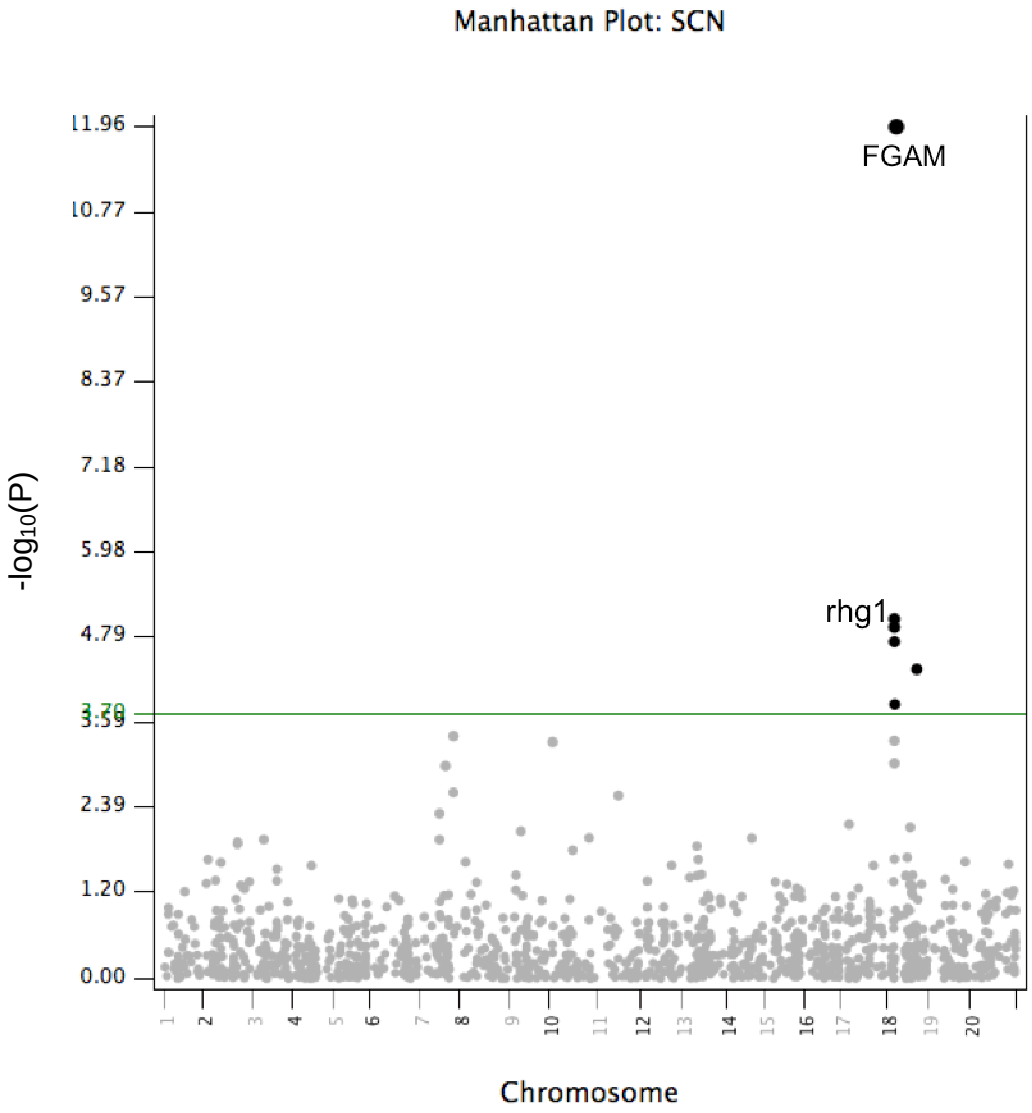

Figure 3. Manhattan plot of association mapping with Q+K model for SCN resistance. The green line represents the false discovery rate (FDR) of 5%.

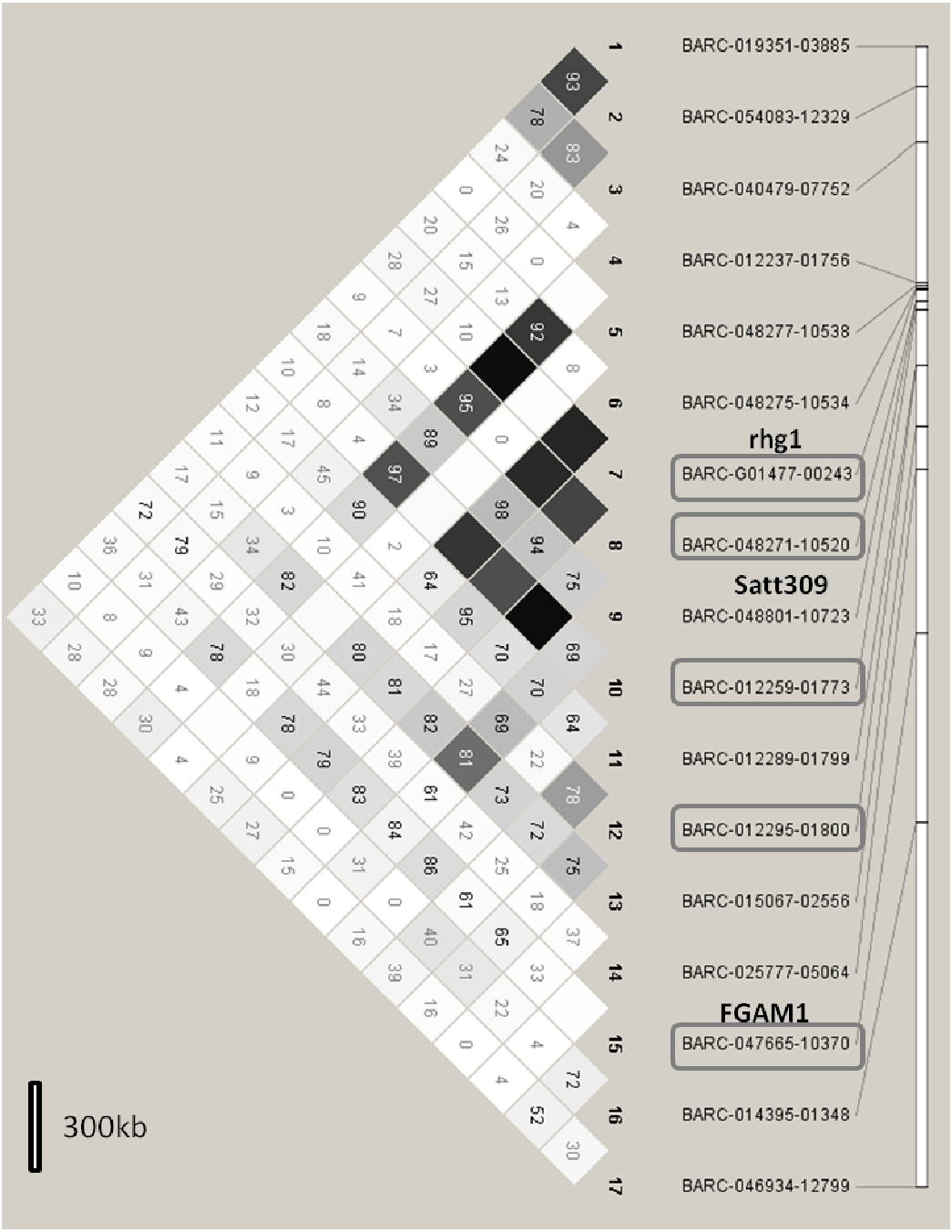

Figure 4. Linkage disequilibrium (LD) heatmap of 17 SNPs near *rhg1* and *FGAM1* on chromosome 18. Number in each cell is the pair-wise LD calculated as r^2 * 100. Color scheme: r^2 = 0, white; 0 < r^2 < 1 shades of grey; r^2 = 1, black. SNPs within box are significantly associated with resistance to SCN HG type 0.

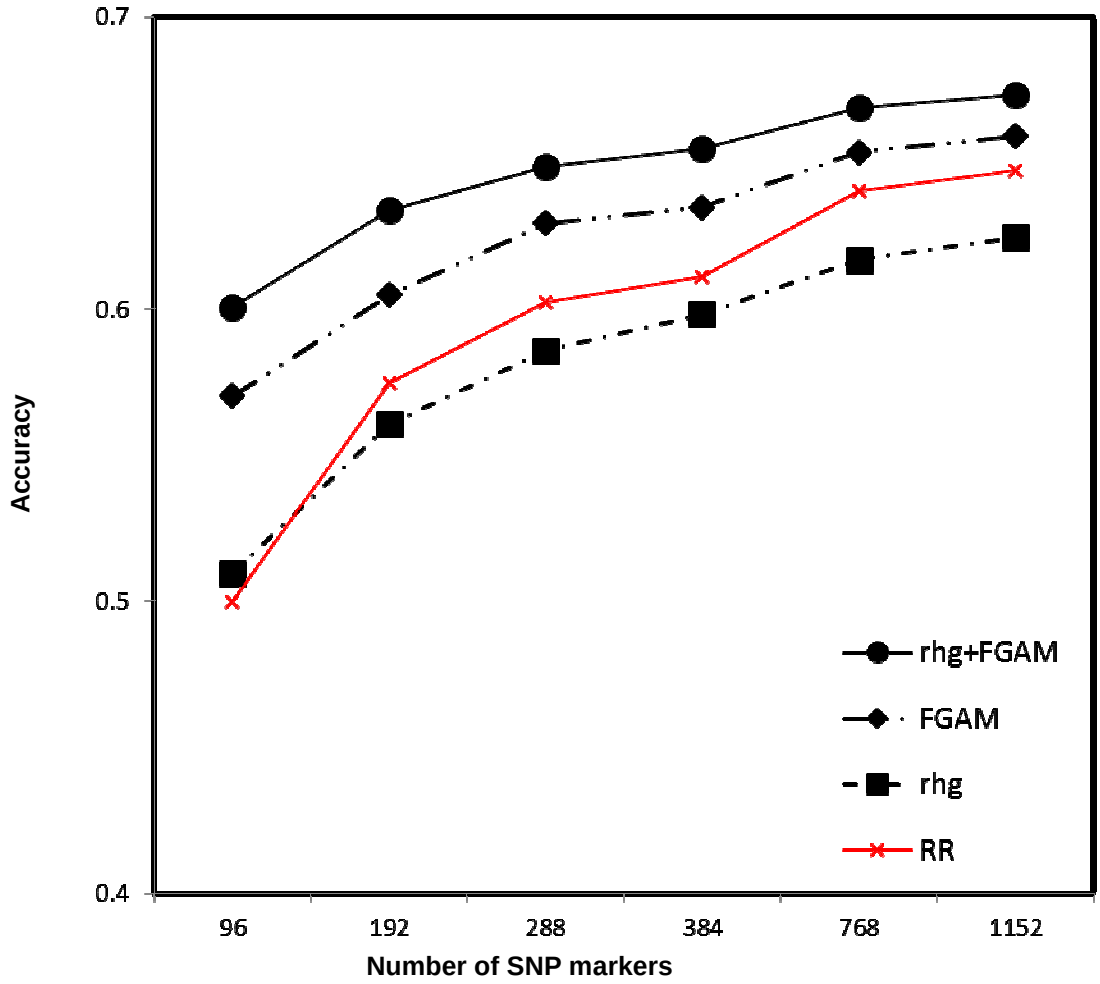

Figure 5. The mean of prediction accuracy with different major gene(s) fixed and different number of SNP markers in genomic selection for SCN resistance. The prediction accuracy was the mean of six folds estimated from 6-fold cross-validation with 100 replications within each fold. RR, ridge-regression best linear unbiased prediction; rhg+FGAM, RR with both *rhg1* and *FGAM1* fixed by treating 4 significant SNPs at *rhg1* and 1 significant SNP at *FGAM1* as fixed effects; FGAM, RR with *FGAM1* fixed by treating 1 significant SNP at FGAM1 as fixed effects; rhg, RR with *rhg1* fixed by treating 4 significant SNPs at *rhg1* as fixed effects.

Chapter 2

Assessing Potential of Association Mapping and Genomic Prediction for Resistance to Sudden Death Syndrome in Early Maturing Soybean Germplasm

Sudden death syndrome (SDS), caused by *Fusarium virguliforme*, has spread to northern soybean growing regions in the U.S. causing significant yield losses. The objectives of this study were to identify loci underlying variation in plant responses to SDS through association mapping (AM) and to assess prediction accuracy of genomic selection (GS) in a panel of early maturing soybean germplasm. A set of 282 soybean breeding lines was selected from the University of Minnesota soybean breeding program, and then genotyped using a genome-wide panel of 1536 single nucleotide polymorphism (SNP) markers. Four resistance traits, root lesion severity (RLS), foliar symptom severity (FSS), root retention (RR), and dry matter reduction (DMR), were evaluated using soil inoculation in the greenhouse. Association mapping identified significant peaks in genomic regions of known SDS resistance quantitative trait loci (QTL) c*qSDS001*, c*qRfs4*, and SDS11-2. Additionally, two novel loci, one on chromosome 3 and another on chromosome 18, were tentatively identified. A nine-fold cross-validation scheme was used to assess the prediction accuracy of GS for SDS resistance. The prediction accuracy of single-trait GS (ST-GS) was 0.64 for RLS but less than 0.30 for RR, DMR, and FSS. Compared to ST-GS, none of multi-trait GS (MT-GS) models significantly improved the prediction accuracy due to weak correlations among the four traits. This study suggests both AM and GS hold promise for implementation in genetic improvement of SDS resistance in existing soybean breeding programs.

Introduction

Sudden death syndrome (SDS), caused by *Fusarium virguliforme* (Aoki et al. 2003), is an important disease that continues to spread across northern soybean (*Glycine max* (L.) Merr.) growing regions in the United States (Bernstein et al. 2007; Chilvers et al. 2010; Malvick and Bussey 2008; Navi et al. 2008; Kurle et al. 2003), causing significant yield losses in SDS-infected fields (Wrather and Koenning 2009). Hyphae penetrate soybean roots and eventually colonize the vascular tissue of the plant causing the development of root rot (Jin et al. 1996). Subsequently, phytotoxin FvTox1 is produced by *F. virguliforme* and translocated to plant leaves during reproductive stages, causing diagnostic foliar sympotoms such as leaf scorch (Brar et al. 2011; Jin et al. 1996). Both the root rot and leaf scorch lead to yield losses varying from 5% to 80% in individual soybean fields greatly affected by environmental conditions (Roy et al. 1997).

Crop rotation is generally ineffective in reducing the occurrence and severity of SDS in fields because *F. virguliforme* in the form of chlamydospore or macrocondia can persist in crop residue and soil for many years (Roy et al. 1997). Although seed treatment fungicides were recommended for managing early season root rots associated with *Fusarium* spp., seed treatments are not considered to be effective against *F. virguliforme*. Therefore, SDS management relies heavily on planting resistant or tolerant cultivars complemented by optimal cultural practices. To date, soybean cultivars with partial resistance to SDS have been identified and developed (Hartman et al. 1997; Mueller et al. 2002; Mueller et al. 2003; Njiti et al. 2002; Schmidt et al. 1999). However, no highly-resistant soybean cultivars adapted to northern growing regions are yet available for

soybean growers to use. Consequently, there is an urgent need to develop early maturing soybean cultivars with effective and durable resistance to SDS.

Developing SDS resistant soybean cultivars has proven difficult mainly due to the complex genetic basis of SDS resistance, the interaction of pathogen and plant with the environment, and/or imperfect screening methods. Both the pathogen and disease are greatly influenced by environmental factors such as temperature, soil fertility, soil texture, rainfall, and planting date, which makes characterization and evaluation of cultivar performance extremely challenging (de Farias Neto et al. 2006; Gongora-Canul and Leandro 2011a, b; Jin et al. 1996; Sanogo and Yang 2001; Vick et al. 2003). For example, rainfall and temperature in the early season can lead to great variability in evaluation of SDS resistance in soybean genotypes because cool and wet conditions favor the initial infection of soybean roots by *F. virguliforme* while weather during reproductive stages influences foliar symptom expression. In order to accurately screen for resistance to SDS, extensive field trials of soybean genotypes across multiple locations and years are necessary.

The genetic architecture of (partial) resistance to SDS is complex. A total of 58 QTL have been reported as providing resistance to SDS in bi-parental mapping populations (www.soybase.org, verified 11 June 2014) and only a few of them have been consistent across mapping populations from different genetic backgrounds (Kazi et al. 2008). However, the resistance loci c*qRfs4* on linkage group C2 (chromosome 6), *cqSDS001* on linkage group D2 (chromosome 17), *cqRfs1*, *cqRfs2*, *cqRfs3* on linkage group G (chromosome 18), and *cqRfs6* on linkage group N (chromosome 3) were repeatedly mapped in multiple populations (de Farias Neto et al. 2007; Hnetkovsky et al.

1996; Iqbal et al. 2001; Kassem et al. 2006; Kazi et al. 2008; Lightfoot et al. 2001; Njiti et al. 1998; 2002; Prabhu et al. 1999). Potentially, the genetic markers identified from previous QTL mapping studies can assist in the selection of SDS-resistant cultivars in a timely and resource-efficient manner (Prabhu et al. 1999). Luckew et al. (2013) recently evaluated ten confirmed SDS QTL in F_2-derived lines from six populations and suggested the possibility of stacking QTL to achieve durable SDS resistance.

QTL mapping in bi-parental populations has been limited by the specific genetic backgrounds of the population under study, which reduces the ability to detect resistance genes. By contrast, association mapping (AM) (Rafalski 2002) provides an opportunity to identify QTL at a higher resolution by taking advantage of historical linkage disequilibrium (LD) in diverse populations. With increasing numbers of single nucleotide polymorphisms (SNPs) combined with declining costs in genotyping, AM has become an attractive approach for revealing the genetic basis of target traits in crop species (Huang et al. 2012; Jia et al. 2013). Recent AM studies have proven successful in identifying QTL for quantitative traits in populations composed of advanced breeding lines and landraces from crop breeding programs (Asoro et al. 2013; Bao et al. 2014; Mamidi et al. 2012; Sukumaran et al. 2012, Zhou and Steffenson 2013; Zhou et al. 2014). To our knowledge, however, AM using genome-scale SNP analysis has not been employed in dissecting the genetic basis of SDS resistance.

Rather than utilizing only molecular markers in significant association with targeted QTLs, a new marker-based approach known as genomic selection (GS) has been developed with the aim of directly predicting genetic value for quantitative traits by taking advantage of all available genome-wide marker information (Bernardo and Yu,

2007; Meuwissen et al., 2001). In the GS scheme, QTL mapping is replaced by genomic prediction model training which involves fitting both phenotypic and genotypic data from a training population in either linear or nonlinear models. Marker effects estimated from the models are subsequently summed up to estimate genomic breeding values of individuals in a validation or breeding population with only genotypic data. Previous results in crop species, including soybean, have indicated that GS holds the potential to improve disease resistance with complex genetic architecture in breeding programs (Bao et al. 2014; Lorenz et al. 2012; Rutkoski et al. 2012, 2014). Here, we seek to investigate the potential use of GS to select SDS resistance in a typical public soybean breeding program.

Materials and Methods

Population, Genotyping, Population Structure, and Linkage Disequilibrium

Details about the population and genotyping strategy were described previously, as were characterization of and the population structure and linkage disequilibrium (LD) (Bao et al. 2014). Briefly, we selected 282 soybean lines including ancestral lines, advanced breeding lines, released public cultivars, and landraces from University of Minnesota Soybean Breeding Program (Bao et al. 2014). An Illumina GoldenGate assay with 1536 SNP markers was used to genotype the selected soybean lines (Hyten et al. 2010). A total of 1247 SNP markers with greater than 5% minor allele frequency (MAF) and missing data rate less than 50% were used in subsequent analyses (Bao et al. 2014). Both *STRUCTURE* (Pritchard et al. 2000) and principal component analysis (PCA) identified a pattern of three clusters in the population approximately corresponding to three distinct

genetic groups (Bao et al. 2014). LD was characterized and illustrated using *Haploview4.2* (Barrett et al. 2005).

Phenotyping and Data Analysis

In spring 2013, a total of 279 soybean lines (seeds of three lines were unavailable) were evaluated for SDS resistance in the greenhouse using the inoculation procedure of Luckew et al. (2012). An isolate of *F. virguliforme*, Somerset #1A, originating in Minnesota had been maintained on PDA until it was used to inoculate autoclaved sorghum for use in these screening experiments. The sorghum was prepared for inoculation by soaking 1.5 liter quantities overnight in sterilizable spawn bags (Fungi Perfecti LLC, Olympia, WA) followed by autoclaving and cooling. The cooled sorghum was then inoculated with 15 x 5 mm blocks of PDA infested with two-week-old cultures of the Somerset #1A isolate. Bags were incubated at room temperature with normal fluorescent room lighting for 30 days. The contents of each bag were mixed daily to ensure uniform infestation of the sorghum throughout the bag. At the time of soybean planting, the growth media was inoculated with a 1:20 (volume:volume) ratio of infested sorghum inoculum to media. The uninoculated control treatment contained only growth media. Each entry was planted in a Jumbo Junior (Belden Plastics Co., St. Paul, MN) square pot containing 800 ml of soil. After planting, the pots were placed in the greenhouse, watered to field capacity daily, and maintained at 22°C with 14 hours daylight.

The greenhouse experiment was conducted as six separate plantings because of space limitations. The six plantings were conducted consecutively under the same

greenhouse conditions. Each planting consisted of 34-55 soybean lines with five inoculated replications plus one uninoculated replication for each line. Two check cultivars: 'McCall' (susceptible) and 'MN0302' (resistant) were included in each planting. For each planting, each plant was evaluated for four symptoms or responses associated with SDS by the same experienced evaluator four weeks after planting. These observations included: root lesion severity (RLS), foliar symptom severity (FSS), root retention (RR), and dry matter reduction (DMR).

RLS is a measure of the severity of root lesion development caused by *F. virguliforme* infection ranging from 1 (no lesion) to 10 (most severe lesion development): 1 = No lesions visible on taproot, 2 = Lesions on 10% of the taproot, 3 = Lesions on 20% of the taproot, 4 = Lesions on 30% of the taproot, 5 = Lesions on 40% of the taproot, 6 = Lesions on 50% of the taproot, 7 = Lesions on 60% of the taproot, 8 = Lesions on 70% of the taproot, 9 = Lesions on 90% to 100% of the taproot, 10 = Lesions on > 90% of the taproot of the taproot or the taproot is completely missing.

FSS is a rating of the severity of leaf scorch caused by *F. virguliforme* (Bowen, C.R. and Slaminko, T.L. 2008, personal communication; Chawla et al. 2013): 1 = no scorch, 2 = slight symptom development, with mottling on leaves, 3 = moderate symptom development with interveinal chlorosis and necrosis, 4 = intermediate symptom development with interveinal chlorosis and necrosis, 5 = severe interveinal chlorosis and necrosis accompanied by cupping, 6 = interveinal chlorosis and necrosis accompanied by cupping with some defioliation, 7 = most leaves displaying necrosis, and 8 = dead plants.

Percentage of root or shoot dry weight change caused by *F. virguliforme* infection was calculated as RR = (root dry weight of inoculated plant) / (root dry weight of uninoculated plant) x 100. DMR = 100 - (shoot dry weight of inoculated plant) / (shoot dry weight of uninoculated plant) x 100.

We then fitted ratings of each trait into a linear regression model: $y = u + L + \varepsilon$ within each planting, and performed the analysis of variation (ANOVA) with the PROC ANOVA in Statistical Analysis System (SAS) Version 9.4 (Cary, NC), where y was one of the four trait ratings of each plant, u was the intercept, L was the effect of soybean line, and ε was the residual. The effect of line x replication was used as the error term to test significance of the effect of line. We represented the phenotypic value of each soybean line as the mean of trait ratings across five replications for each trait, and used the phenotypic values for subsequent association mapping and GS modeling. Scatter plots were made based on the pair-wise correlation between the phenotypic values of each pair of traits.

Association Mapping

We performed association mapping (AM) for RLS, FSS, RR, and DMR, respectively, with mixed linear model in the "rrBLUP" package (Endelman, 2011) in R (R Development Core Team, 2010). The mixed linear model: $y = X\alpha + P\beta + K\gamma + \varepsilon$ was used, where y is the vector of phenotypic values, X is the vector of SNP marker genotypes, α is the coefficient of marker effect being estimated, P is the matrix of first three principal components from PCA accounting for the population structure plus the covariate vector of experimental plantings, β is the coefficient of principal components

and experimental plantings, *K* is the additive relationship matrix estimated based on SNP genotypes accounting for genetic kinship among the individuals, γ is the vector of random effects corresponding to genetic kinship, and ε is the vector of random effects corresponding to residuals. The variances of γ and ε are $Var(\gamma) = 2KV_g$ and $Var(\varepsilon) = V_R$, respectively, where *K* is the genetic kinship, V_g is the genetic variance, and V_R is the residual variance. False discovery rate (FDR) of 0.05 was used to correct for multiple comparisons in AM using package "QVALUE" in R (R Development Core Team, 2010). SNP markers with FDR q-value < 0.05 were defined as significant SNPs associated with SDS resistance. Given the low SNP density on our genotyping panel, significant SNP markers are not expected to be exact locations of causal genes controlling variation of plant response to SDS. In the vicinity of the significant SNPs, we scanned previously-described SDS resistance QTL in soybean genome (www.soybase.org). Manhattan plots were created based on the AM results with SNPEVG (Wang et al. 2012).

Prediction Accuracy of Genomic Selection

To assess prediction accuracy of genomic selection (GS) for SDS resistance, the same set of phenotypic and genotypic data was used in a nine-fold cross-validation study. Specifically, 279 soybean lines first were randomly divided into nine subsets. In each fold, eight subsets of lines (248 lines) were used as training sets and the remaining subset (31 lines) was a validation set. In the training set, the marker effects were simultaneously estimated by fitting a statistical model to both phenotypic and genotypic data. The marker effects were then used to predict the genetic values of individuals in the validation set. Prediction accuracy was calculated as the correlation between marker-based prediction

and phenotypic values. The cross-validation process was repeated nine times (nine folds), with every subset of soybean lines used exactly once as the validation set.

Genomic Selection Model

Since there are four phenotypic traits associated with SDS resistance in our data set, we evaluated both multi-trait genomic selection (MT-GS) and single-trait genomic selection (ST-GS) model, and compared their accuracies for predicting SDS resistance. For ST-GS, a mixed linear model was constructed to estimate marker effects of phenotypic traits: $y = Xb + Z\alpha + e$, where *y* is the vector (n x 1) of phenotypic observations of n individuals, *X* is the design matrix (n x r) for fixed planting effects, *b* is the vector (r x 1) of planting effects, *Z* is the design matrix (n x m) for additive effects of SNP markers, α is the vector (m x 1) of additive effects of SNP markers, and *e* is the vector (n x 1) of residuals. The variances of α and *e* are $Var(\alpha) = I_m\sigma_\alpha^2$ and $Var(e) = I_n\sigma_e^2$, respectively, where I_m is the m x m identity matrix, σ_α^2 is the additive genetic variance for each maker, σ_e^2 is the residual variance, I_n is the n x n identity matrix. We employed a computationally efficient method, ridge-regression best linear unbiased prediction (RR-BLUP) to solve the mixed model. Previous GS studies suggested slight differences among various genomic prediction algorithms including G-BLUP (which is equivalent to RR-BLUP), Bayesian approaches, and machine learning algorithms (Asoro et al. 2011; Bao et al. 2014; Lorenzana and Bernardo, 2009; Lorenz et al. 2012; Rutkoski et al. 2012). The marker effects were simultaneously estimated by solving the mixed model through the restricted maximum likelihood (REML) method implemented in R package

"rrBLUP" (R Development Core Team, 2010). Variance of additive effects and variance of residual effects were estimated.

MT-GS models were developed by fitting the phenotypic observations of multiple traits (t) simultaneously in a multivariate mixed linear model:

$y = (I_t \otimes X)b + (I_t \otimes Z)\alpha + e$, where *y* is the matrix (n x t) of phenotypic observations for t traits of n individuals, I_t is the identity matrix (t x t), *X* is the design matrix (n x r) for fixed planting effects for each trait, *b* is the matrix (r x t) of planting effects for t trait, *Z* is the design matrix (n x m) for additive effects of SNP markers for each trait, α is the matrix (m x t) of additive effects of SNP markers for t trait, *e* is the matrix (n x t) of residuals, and $\otimes$ denotes the Kronecker product. The variances of α and *e* are $Var(\alpha) = G_0 \otimes A$ and $Var(\mathrm{e}) = \mathrm{R}_0 \otimes \mathrm{I}_\mathrm{n}$, respectively, where G_0 is the covariance matrix (t x t) of additive effects, A is the additive genetic relationship matrix (n x n), R_0 is the covariance matrix (t x t) of residuals, and I_n is the identity matrix (n x n). The marker effects of each trait were simultaneously estimated by solving the mixed model through REML method implemented in R package "rrBLUP" (R Development Core Team, 2010). The pair-wise genetic correlation was estimated as $\sigma_{g12}\sqrt{\sigma_{g11}\sigma_{g22}}$, where σ_g is the genetic variance-covariance matrix for multiple traits. σ_g was calculated as $\sum_{i=1}^{m_i} Var(SNP_i)\alpha_i\alpha_i^T$ (Jia and Jannink, 2012). The additive genetic variance and the residual variance were estimated. Ten types of MT-GS models were developed: RLS_FSS model for RLS and FSS; RLS_RR model for RLS and RR; RLS_DMR model for RLS and DMR; FSS_RR model for FSS and RR; FSS_DMR model for FSS and DMR; RR_DMR model for RR and DMR; RLS_FSS_DMR model for RLS, FSS, and

DMR; RLS_FSS_RR model for RLS, FSS, and RR; RR_FSS_DMR model for RR, FSS, and DMR; and FT model for all four traits. A notched boxplot was made to compare the prediction performance of MT-GS models to ST-GS models for each trait. The notch marks the 95% confidence interval for the medians. In the notched boxplot, the medians significantly differ if two boxes' notches do not overlap.

Marker Number

We also determined the effect of marker numbers on GS accuracy through nine-fold cross-validation by including random samples of 96, 192, 384, and 768 SNPs from the full marker set. Within each fold, this was repeated 100 times to avoid sampling bias for markers. All prediction accuracies were estimated with R package "rrBLUP" (Endelman, 2011). A notched boxplot was made to compare the prediction performance of GS models with different subsets of markers for each trait. The notch marks the 95% confidence interval for the medians. In the notched boxplot, the medians significantly differ if two boxes' notches do not overlap.

Results

Phenotypic Analysis

Analysis of variance (ANOVA) for each of four SDS resistance traits was conducted within each planting. ANOVA showed the effect of soybean lines was significant ($p < 0.05$) in all plantings, except for FSS in planting 4 and 6, and RR and DMR in planting 5 (Table 5). The lack of significance of line effect in ANOVA indicated that the effect of replication x line contributed a large portion of the trait variation within the planting.

Susceptible and resistant check cultivars were set up to provide the means of comparing phenotyping performance in the six plantings. As expected, the susceptible check ‘McCall’ exhibited high RLS scores ranging from 5.5 to 8.8 within plantings with an exception of 2.4 in Planting 3; the resistant check ‘MN0302’ exhibited low RLS scores ranging from 2.2 to 4.6 within plantings with an exception of 6.4 in Planting 1 (Data not shown).

In general, soybean lines showed a wider range of responses to SDS for both RR and DMR than RLS and FSS scores (Fig. 6). The phenotypic data density of RLS was more evenly distributed than that of the other three traits (Fig. 6). RLS scores ranged from 2.4 to 10 with a total of 49 lines exhibiting scores less severe than the resistant check ‘MN0302’ (Fig. 6). FSS scores ranged from 1 to 8 with a total of 81 lines that did not develop any foliar symptoms plus another 43 less severe than ‘MN0302’ (Fig. 6). The range observed in RR was 0 to 1141% with a total of 69 lines more resistant than ‘MN0302’ (Fig. 6). A total of 29 lines did not show any dry matter reduction plus another 64 lines with DMR less severe than ‘MN0302’ (Fig. 6). Based on all four traits associated with SDS resistance, 11 soybean lines consistently exhibited symptoms less severe than that of the resistant check ‘MN0302’, and have potential to be used as breeding parents in the SDS resistance improvement program (Table S4).

Pair-wise Correlations of Traits

The pair-wise correlations between the phenotypic values of each pair of traits were shown in scatter plots (Fig. 7). As expected, a strong negative correlation was observed between RR and DMR, while RLS and FSS were positively correlated with r = 0.47 (Fig.

7). However, the correlations between RR and RLS, DMR and FSS, RLS and FSS, and RR and FLS were poor (Fig. 7). We observed similar pair-wise phenotypic correlations within each of six plantings (Data not shown). The pair-wise genetic correlation of traits was consistent with the observation in phenotypic correlation (Table 7).

Association Mapping

Association mapping (AM) was performed for RLS, FSS, RR, and DMR. We identified two and eight significant (qFDR < 0.05) SNP markers for DMR and RR, respectively, but none for the other two traits (Table 6; Fig. 8). Among the eight distinct significant markers, three were in the same genomic interval as the known SDS resistance quantitative trait loci (QTL) *cqSDS001* on linkage group D2 (chromosome 17) (Table 6; Fig. 8). Another marker at position 80.28 cM on linkage group C2 (chromosome 6) was in the genomic region of *cqRfs4* (Table 6; Fig. 8). Both *cqSDS001* and *cqRfs4* have been previously identified and confirmed in multiple bi-parental populations (de Farias Neto et al. 2007; Hnetkovsky et al. 1996; Iqbal et al. 2001; Kassem et al. 2012; Kazi et al. 2008; Njiti et al. 2002). Additionally, two significant SNP markers in our study confirmed a previously identified QTL, *SDS11-2*, on linkage group D2 (chromosome 17) (Kazi et al. 2008). The rediscovery of the previously identified QTL strengthened the confidence of overall quality of AM analysis. Moreover, one SNP marker near the telomere on chromosome 3 (linkage group N) was tentatively identified as associated with RR variation, and another SNP marker near the telomere on chromosome 18 (linkage group G) was associated with the variation in both RR and DMR (Table 6; Fig. 8; Fig. 9). Since no QTL had been discovered near these two significant SNP markers, two newly

identified loci named as *SDS14-1* (on chromosome 3) and *SDS14-2* (on chromosome 18) were added to the list of QTL underlying resistance to SDS (Fig. 8; Fig. 9). For each of five loci (namely, *cqSDS001*, *cqRfs4*, *SDS11-2*, *SDS14-1*, and *SDS14-2*), RR and DMR peaks were coincident with each other (Fig. 9). This indicated that markers associated with these five loci would be potentially useful for selecting for resistance to both root and shoot reduction caused by SDS.

Single-trait versus Multi-trait Genomic Selection

Besides identifying causal loci associated with SDS resistance through AM, the phenotypic and genotypic data sets were used to evaluate the utility of GS in predicting SDS resistance phenotypes. Single-trait (ST) and ten types of multi-trait (MT) GS models were developed for SDS resistance. The prediction accuracy of ST model was 0.64, 0.20, 0.18, and 0.16 for RLS, FSS, RR, and DMR, respectively (Fig. 10). Compared to ST models, none of MT-GS models significantly improved the prediction accuracy for any trait (Fig. 10). The RLS_FSS_DMR model increased the prediction accuracy for DMR from 0.16 to 0.25 meanwhile maintaining a similar accuracy for FSS, but reduced the accuracy for RLS to 0.26 (Fig. 10). The FT model performed equivalently to ST models for all four traits (Fig. 10).

Marker Number

To determine the effect of number of markers on prediction accuracy, we compared the prediction accuracy with different sizes of marker set used in ST-GS models. The prediction accuracy generally increased as the number of SNP markers increased for

RLS, RR, and DMR, but not for FSS (Fig. 11). The rate of gain in accuracy was significant when the marker set increased from 96 to 192 for RLS; and 384 to 768 for RLS, RR and DMR (Fig. 11). With 96 random genome-wide SNPs, the prediction accuracy of ST model was only 0.25, 0.02, 0.14, and 0.04 for RLS, FSS, RR, and DMR, respectively (Fig. 11).

Discussion

Accurate assessment of phenotypic variation is essential for understanding disease biology, effective resistance breeding, and dissection of genetic architecture. The heritability of greenhouse evaluation of SDS resistance ranged from 33% to 66% in previous studies (Njiti et al. 2001). In our greenhouse experiment, the effect of soybean genotypes was significant ($p < 0.05$) in most plantings indicating an overall reliability of phenotypic data (Table 5). We, however, still observed substantial replication x genotype variation in four trait x planting experiments (Table 5). The high level of phenotypic variation among replications has also been observed in previous studies (Kazi et al. 2008; Luckew et al. 2013), and could be attributed to the complex genetic basis of SDS resistance, interactive effects of genotype with environment, and/or imperfect screening methods. Another limitation in the current study was the low throughput capacity of the phenotyping system; the evaluation of soybean lines conducted as six plantings, which might have reduced our ability to detect all causative QTL or led to biased estimation. In other words, the genetic effects might be confounded with the effect of consecutive experimental plantings conducted over time, limiting our ability to induce SDS symptoms, and as a result, reducing the explanatory power of AM. This was the result of

changing light intensity and ambient temperature variation associated with seasonal changes in sun angle and ambient temperature. To minimize the influences of these sources of variance of among plantings, we conducted the greenhouse experiments with supplemental lighting and air conditioning, and accounted for the effect of plantings as a fixed effect in the AM model.

We identified eight and two SNP markers in significant association with RR and DMR, respectively, which indicated a total of five loci underlying SDS resistance. Among the five loci identified in this study, *cqSDS001* and *cqRfs4* have been previously identified and confirmed in more than one population, which strengthens the confidence of the overall analysis. The *cqSDS001* locus was first discovered at positions 78 cM and 85 cM on linkage group D2 from the resistant sources PI567374 and Ripley, respectively (de Farias Neto et al. 2007), and was later confirmed in another population derived from Hartwig (Kazi et al., 2008). A second SDS resistance locus, *cqRfs4*, was reported to be associated with foliar resistance (Kazi et al. 2008; Luckew et al. 2013; Triwitayakorn et al. 2005), however, we identified a significant SNP marker, BARC-028177-05786, underlying variation of RR in this QTL interval. Given increasing numbers of SNPs in newly developed genotyping assay for soybean (Song et al. 2013), higher resolution of genetic mapping might pinpoint the potential candidate genes in the genomic regions underlying SDS resistance. Additionally, two SNP markers on linkage group D2 were detected as being significantly associated with RR in our study, which adds support to the *SDS11-2* locus identified previously in Kazi et al. (2008).

A cluster of SDS resistance genes: *cqRfs1*, *cqRfs2*, *cqRfs3* have been repeatedly mapped on linkage group G (chromosome 18) in earlier studies but were not detected in

our collection of soybean accessions (Chang et al. 1996; Iqbal et al. 2001; Meksem et al. 1999; Njiti et al. 1998; 2002; Prabhu et al. 1999; Kazi et al. 2008). A soybean cyst nematode (SCN) resistance gene, *rhg1*, is in the vicinity of this cluster of SDS resistance genes near the telomeric region on chromosome 18. Furthermore, the *Rfs2/rhg1* locus has been reported as being associated with pleiotropic resistance to both SCN and SDS in roots (Afzal et al. 2012; Gelin et al. 2006; Iqbal et al. 2005; Srour et al. 2012; Triwitayakorn et al. 2005). A significant signal was detected at the *rhg1* locus in an earlier AM for SCN resistance using the same markers and population (Bao et al. 2014), however, we could not confirm the presence of *Rfs2* in the current study. A possible explanation might be the differing sources of *rhg1* allele in two studies. *Rhg1* in Srour et al. (2012) was derived from SCN resistant cultivar "Peking", while PI88788 was the only resistant source in our earlier study where an association with SCN resistance *rhg1* was identified (Bao et al. 2014).

Instead of *Rfs2*, we identified a significant SNP marker BARC-024251-04812 on the opposite end of chromosome 18, which accounted for the variation in both RR and DMR. This SNP was about 1.7 Mb away from a previously-described resistance QTL *SDS4-2* (Njiti et al. 1998). Another novel locus was tagged by a SNP marker BARC-044643-08744 located near the telomeric region of chromosome 3. These two novel loci could be validated in future investigation of either a bi-parental mapping population or another AM population with a higher density of SNP markers.

To pyramid these resistance QTL into commercial soybean cultivars, the significant SNP markers identified in present study can be developed as a breeder-friendly SNP array for conducting MAS in SDS resistance breeding programs. However,

stacking multiple QTL and introgressing them to an adapted elite parent requires considerable resources and time. As an alternative to stacking major SDS resistance genes, GS may provide breeders an opportunity to integrate a broader set of causative loci underlying SDS resistance with the goal of more durable resistant soybean cultivars. Despite successful rediscovery of known QTL for RR and DMR, we failed to identify any significant signals (qFDR < 0.05) for RLS and FSS with AM. This might indicate that the genetic variation of RLS and FSS captured in the population is associated with numerous causative genes each with a small effect. In this case, genome-wide selection as implemented through GS will be more effective than MAS because GS would enable breeders to select candidate lines with higher levels of cumulative resistance to SDS conferred by numerous small effect loci. The prediction accuracy for RLS was as high as 0.64 (Fig. 10), which is comparable to that for SCN resistance in soybean (Bao et al. 2014), Fusarium head blight resistance in barley and wheat (Lorenz et al. 2012; Rutkoski et al. 2012), and northern leaf blight in corn (Technow et al. 2013). Given the high prediction accuracy, GS holds great potential for implementation in genetic evaluation of breeding candidates in an actual soybean improvement program targeting at SDS resistance.

SDS resistance breeding is further complicated by the existence of two apparently distinct resistance mechanisms involved in expression of root versus foliar responses to SDS (Kazi et al. 2008; Triwitayakorn et al. 2005). Some known QTL confer specific resistance to root rot or foliar scorch, while others confer resistance to both (de Farias Neto et al. 2007; Hnetkovsky et al. 1996; Iqbal et al. 2001; Kassem et al. 2006; Kazi et al. 2008; Njiti et al. 1998; 2002). To develop soybean cultivars with both root and foliar

resistance to SDS, multi-trait GS (MT-GS) has the potential to be an effective selection strategy for implementing an SDS resistance improvement program. An MT-GS model is developed by simultaneously fitting phenotypic data from the evaluations of root and foliar symptoms as dependent variables in the model. Subsequently, the MT-GS model using one marker panel leads to simultaneous prediction of both root and foliar symptoms.

Our results suggested that the prediction accuracy of GS model based on single traits (ST-GS) for FSS, RR, and DMR was comparatively low (< 0.3) (Fig. 10). In a simulation study, Jia and Jannink (2012) indicated that the prediction accuracy for low-heritability traits could be improved by GS models based on multiple related traits (MT-GS) models. The underlying mechanism of improved accuracy for low-heritability traits in MT-GS is presumably genetic relationship among the highly related traits (Jia and Jannink, 2012). In the case of SDS resistance, MT-GS might be capable of taking advantage of the genetic relationship between low-heritability traits: FSS, RR and DMR, and high-heritability trait: RLS. However, the FT model based on all four SDS resistance traits performed equivalently to the ST models in our study (Fig. 10). None of the MT-GS models significantly improved the prediction accuracy. We observed an increase in the prediction accuracy for DMR with the RLS_FSS_DMR model, while the RLS_FSS_DMR model failed to maintain similar prediction accuracy for RLS and FSS as that in ST-GS models (Fig. 10).

A simulation study indicated that MT-GS greatly increased the prediction accuracy only when the genetic correlation between two related traits was higher than 0.7 (Jia and Jannink, 2012). The MT-GS models performed equivalent to the ST-GS models;

this indicates that the genetic basis of FSS, RR, and DMR might not be highly correlated with that of RLS. It was consistent with weak pair-wise correlation of FSS x RLS, RR x RLS and DMR x RLS as shown in Fig. 7 and Table 7. Mueller et al. (2002) also suggested that the correlation between root rot and foliar severity was not significant. Considering that root rot is caused by direct infection of *F. virguliforme* (Jin et al. 1996), while foliar scorch is caused by phytotoxin FvTox1 produced by *F. virguliforme* (Brar et al. 2011; Jin et al. 1996), different genetic mechanisms appear to be involved in root versus foliar resistances.

Conclusion

The present study suggests AM could be used as an alternative method for mapping QTL underlying SDS resistance, and GS holds potential for implementation in genetic evaluation of root lesion severity associated with SDS. We conclude that SDS resistance is a complex of disease traits, leading to numerous challenges in evaluating and breeding for SDS resistant soybean cultivars. Firstly, improving phenotypic screening methods to ensure high quality and high throughput evaluation of SDS resistance should remain as an important component of the current SDS breeding program. Secondly, high-density genome-wide markers or sequence-based genotyping methods could be employed to dissect the genetic architecture of SDS resistance more precisely. Lastly, the realized response and cost-effectiveness of GS deserves further investigation in both greenhouse and field prior to implementing GS for developing durable SDS resistance in soybeans.

Table 5. Analysis of Variation (ANOVA) for four SDS resistance traits within each planting.

Planting	Source	RLS[a]		FSS[b]		RR[c]		DMR[d]	
		Df[e]	MS[f]	Df	MS	Df	MS	Df	MS
1	Line	53	6.82**	53	11.02**	52	0.65**	52	0.32**
	Error	210	3.73	210	6.01	206	0.35	206	0.20
2	Line	54	7.04***	54	5.68**	54	0.26**	54	0.18***
	Error	196	3.50	196	3.51	196	0.15	196	0.07
3	Line	49	12.19***	49	6.78***	49	0.61**	49	2.24***
	Error	200	4.46	200	3.33	200	0.35	200	0.15
4	Line	53	10.09*	53	6.34	53	1.07**	53	2.81***
	Error	203	7.53	203	5.21	203	0.62	203	0.97
5	Line	52	10.14***	52	10.47**	52	0.14	52	0.09
	Error	188	4.77	188	5.86	187	0.13	188	0.09
6	Line	34	13.51**	34	3.99	34	17.45***	34	14.01***
	Error	118	7.17	118	3.76	118	1.64	118	1.21

* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.

a RLS, root lesion severity.

b FSS, foliar symptom severity.

c RR, root retention.

d DMR, dry matter reduction.

e Df, degree of freedom.

f MS, mean of square.

Table 6. The significant SNPs (false discovery rate < 0.05) detected from association mapping (AM) for SDS resistance.

Trait[a]	Marker	LG[b]	Chromosome	Position (cM)	Position (bp)	P	qFDR[c]
RR	BARC-044643-08744	N	3	4.71	460387	0.00017	0.03
	BARC-028177-05786	C2	6	80.28	13550856	0.000095	0.02
	BARC-051665-11191	D2	17	72.14	14849926	0.00000019	0.0002
	BARC-023721-03465	D2	17	75.11	20352435	0.00020	0.03
	BARC-064101-18557	D2	17	75.44	25852278	0.000023	0.008
	BARC-059487-15840	D2	17	76.12	35057016	0.000012	0.006
	BARC-061049-17016	D2	17	77.39	36090548	0.0000071	0.005
	BARC-024251-04812	G	18	94.30	59472567	0.000059	0.002
DMR	BARC-051665-11191	D2	17	72.14	14849926	0.000020	0.01
	BARC-024251-04812	G	18	94.30	59472567	0.0000058	0.008

a RR, root retention; DMR, dry matter retention.

b LG, linkage group.

c qFDR, q-value of false discovery rate (FDR) estimated with R package "QVALUE". SNP markers with FDR q-value < 0.05 were defined as significant SNPs associated with SDS resistance.

Table 7. Pair-wise genetic correlation of traits associated with SCN resistance.

	FSS[b]	RR[c]	DMR[d]
RLS[a]	0.51	-0.19	0.16
FSS		-0.17	0.17
RR			-0.87

a RLS, root lesion severity.

b FSS, foliar symptom severity.

c RR, root retention.

d DMR, dry matter reduction.

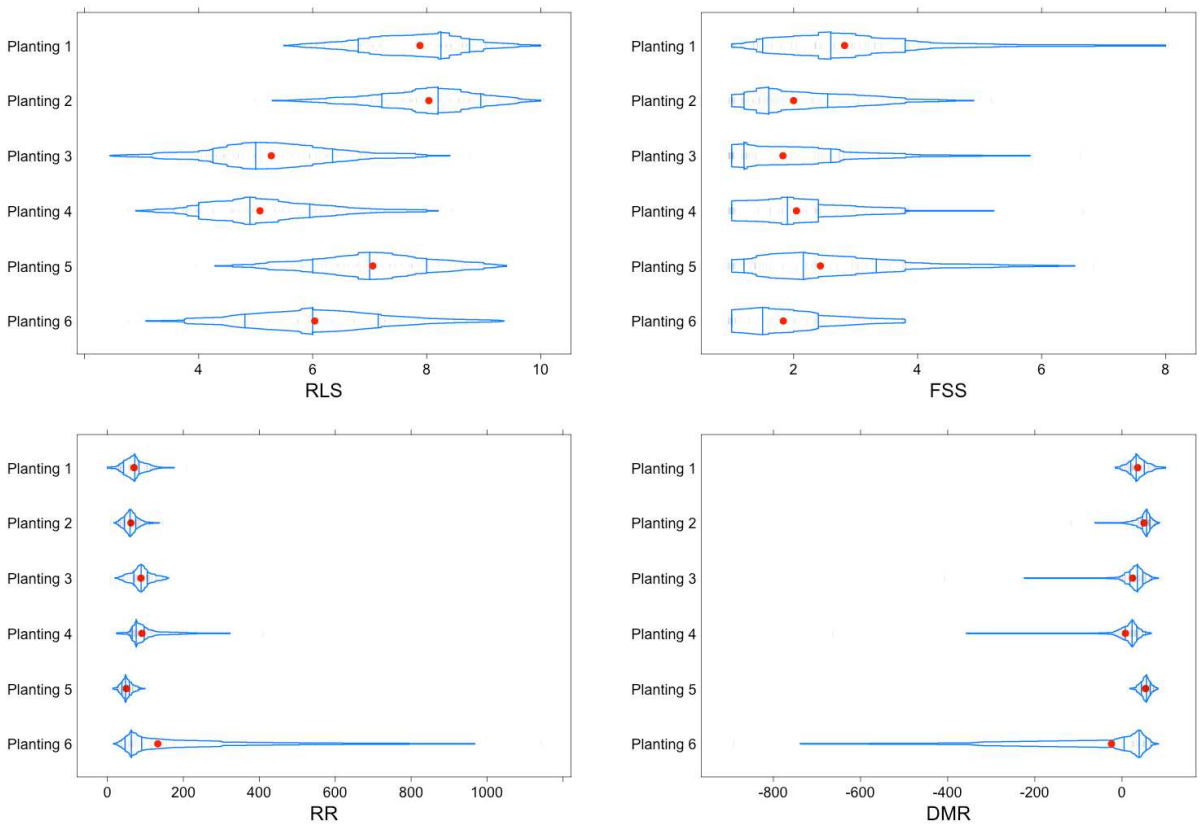

Fig. 6. Box-percentile plots with data density of four SDS resistance traits. RLS, root lesion severity; FSS, foliar symptom severity; RR, root retention (%); DMR, dry matter reduction (%).

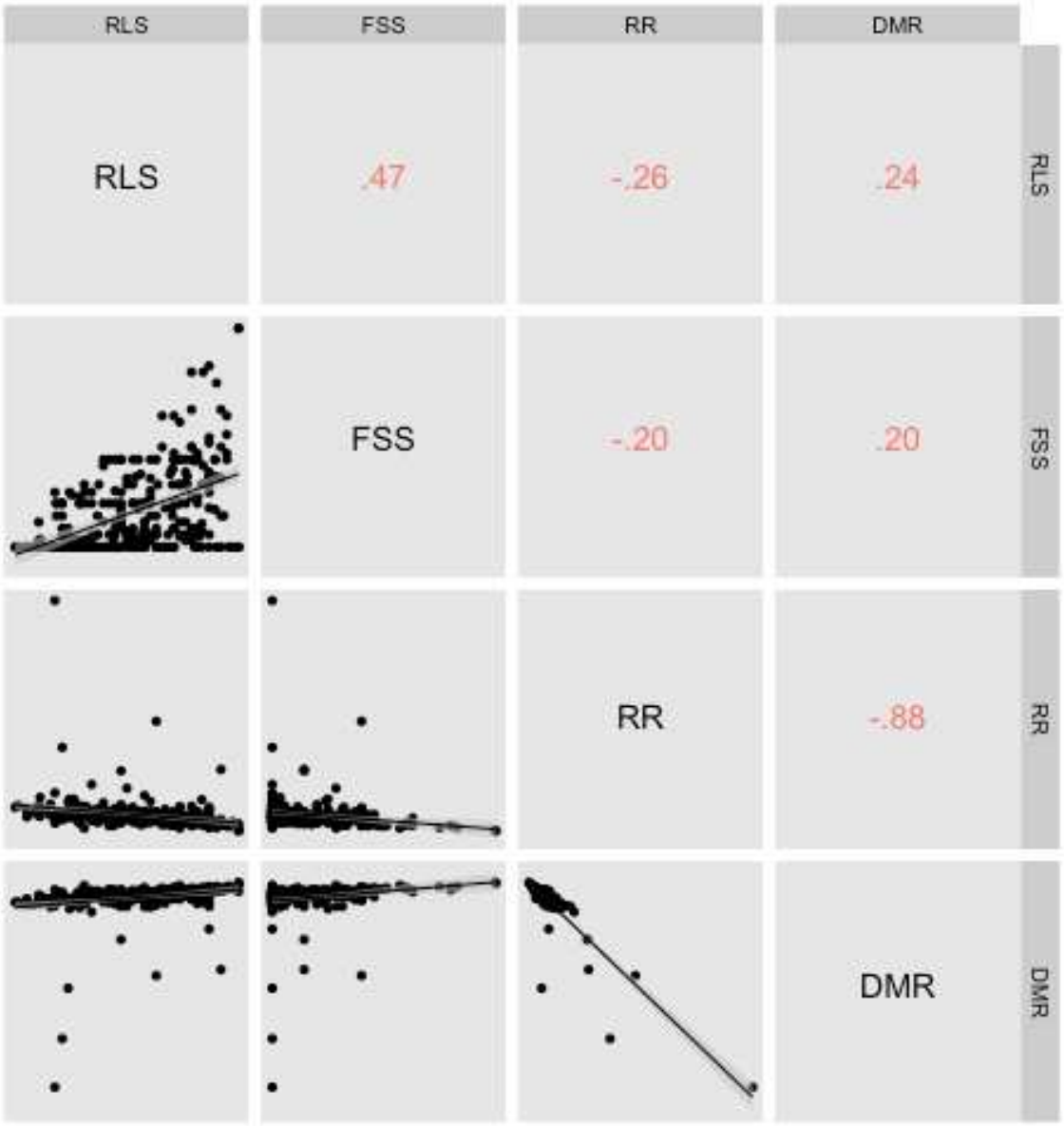

Fig. 7. Scatter plots of pair-wise correlation of traits associated with SDS resistance. RLS, root lesion severity; FSS, foliar symptom severity; RR, root retention (%); DMR, dry matter reduction (%). The values in the scatter plot matrix represent the r-values of pair-wise correlation of traits.

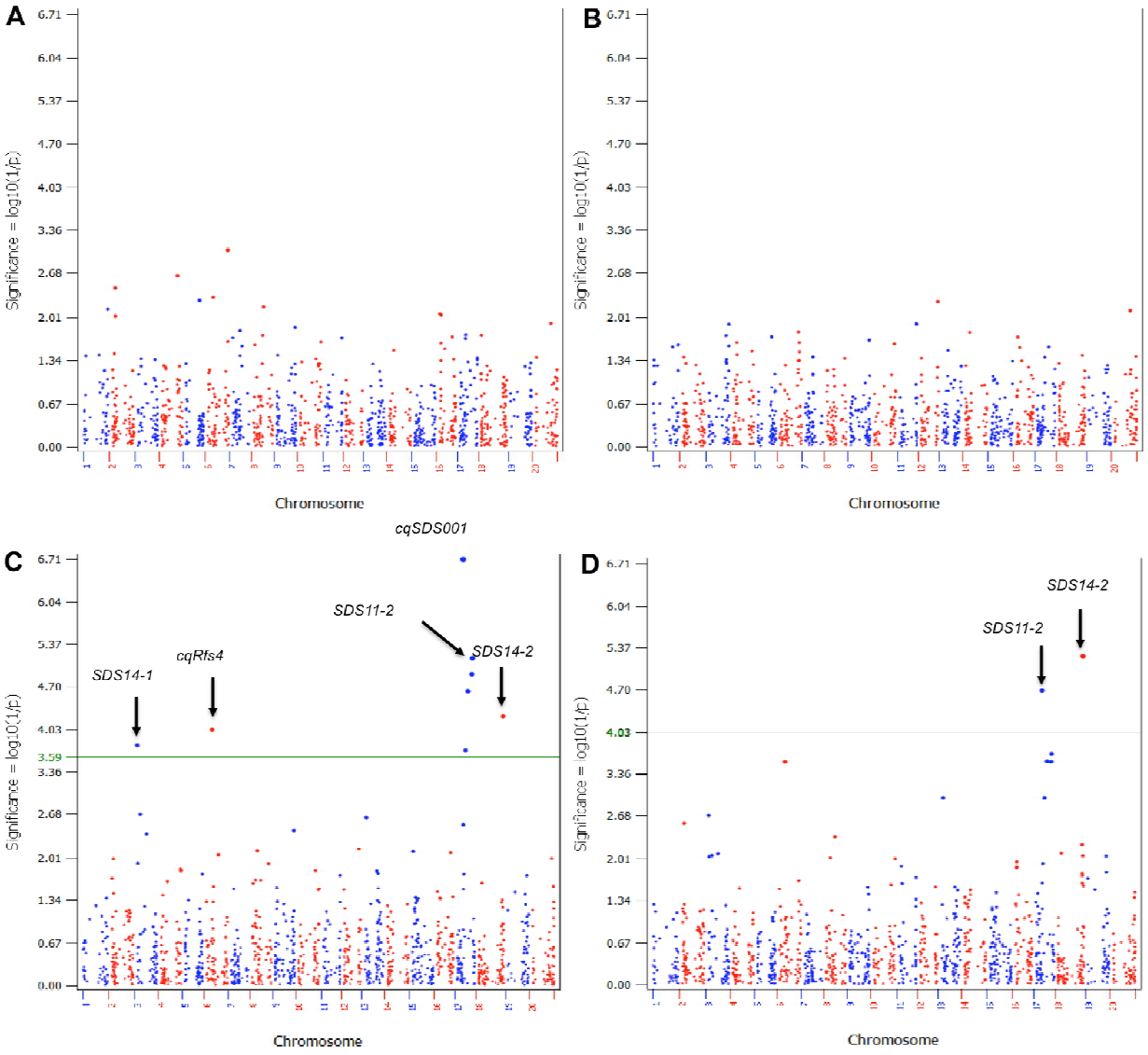

Figure 8. Manhattan plots of association mapping for four SDS resistance traits. A, root lesion severity (RLS); B, foliar symptom severity (FSS); C, root retention (RR); D, dry matter reduction (DMR). The green horizontal line represents the false discovery rate (FDR) of 5%.

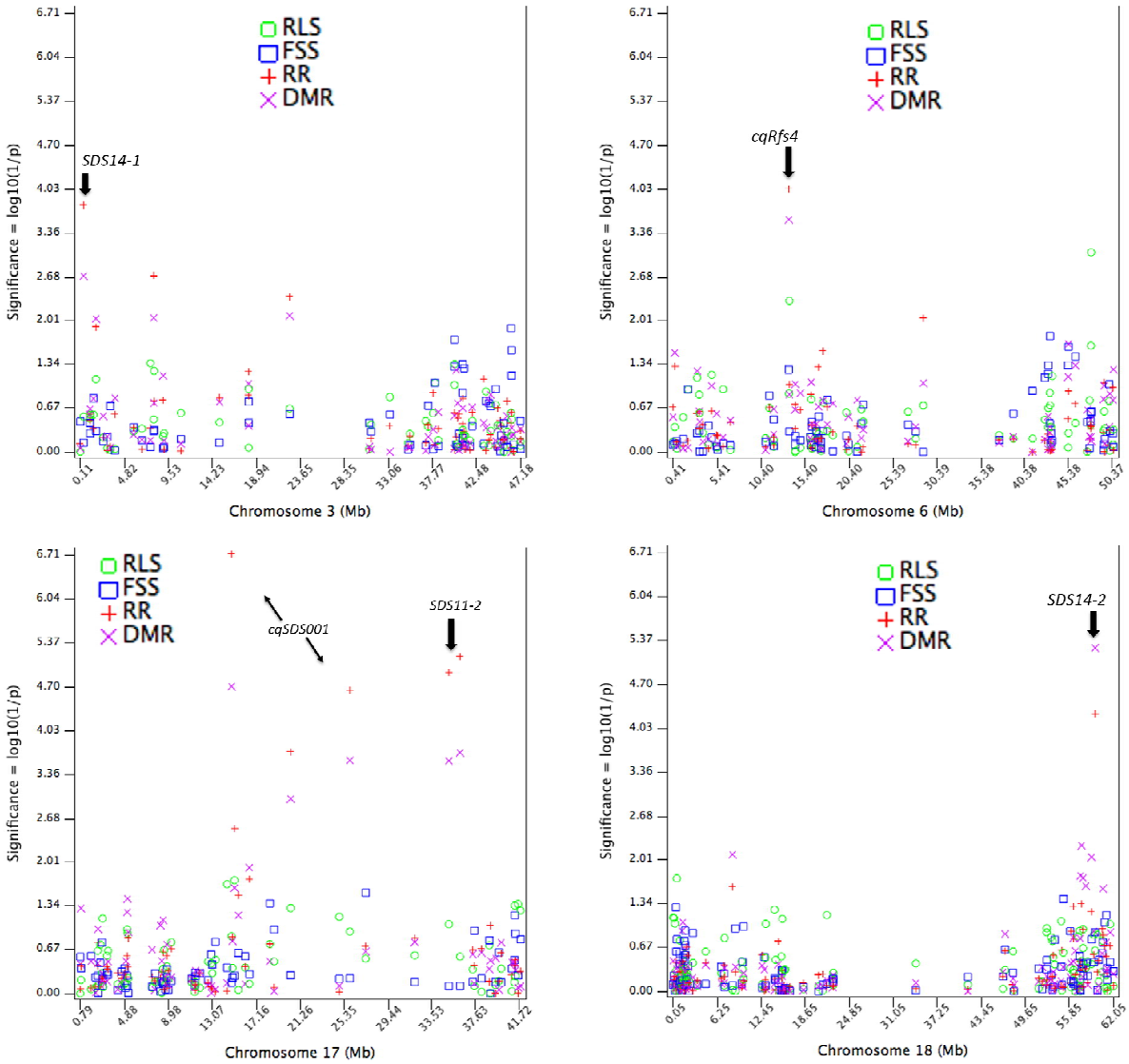

Figure 9. Manhattan plots of association mapping for SDS resistance on chromosome 3, 6, 17, and 18. RLS, root lesion severity; FSS, foliar symptom severity; RR, root retention (%); DMR, dry matter reduction (%).

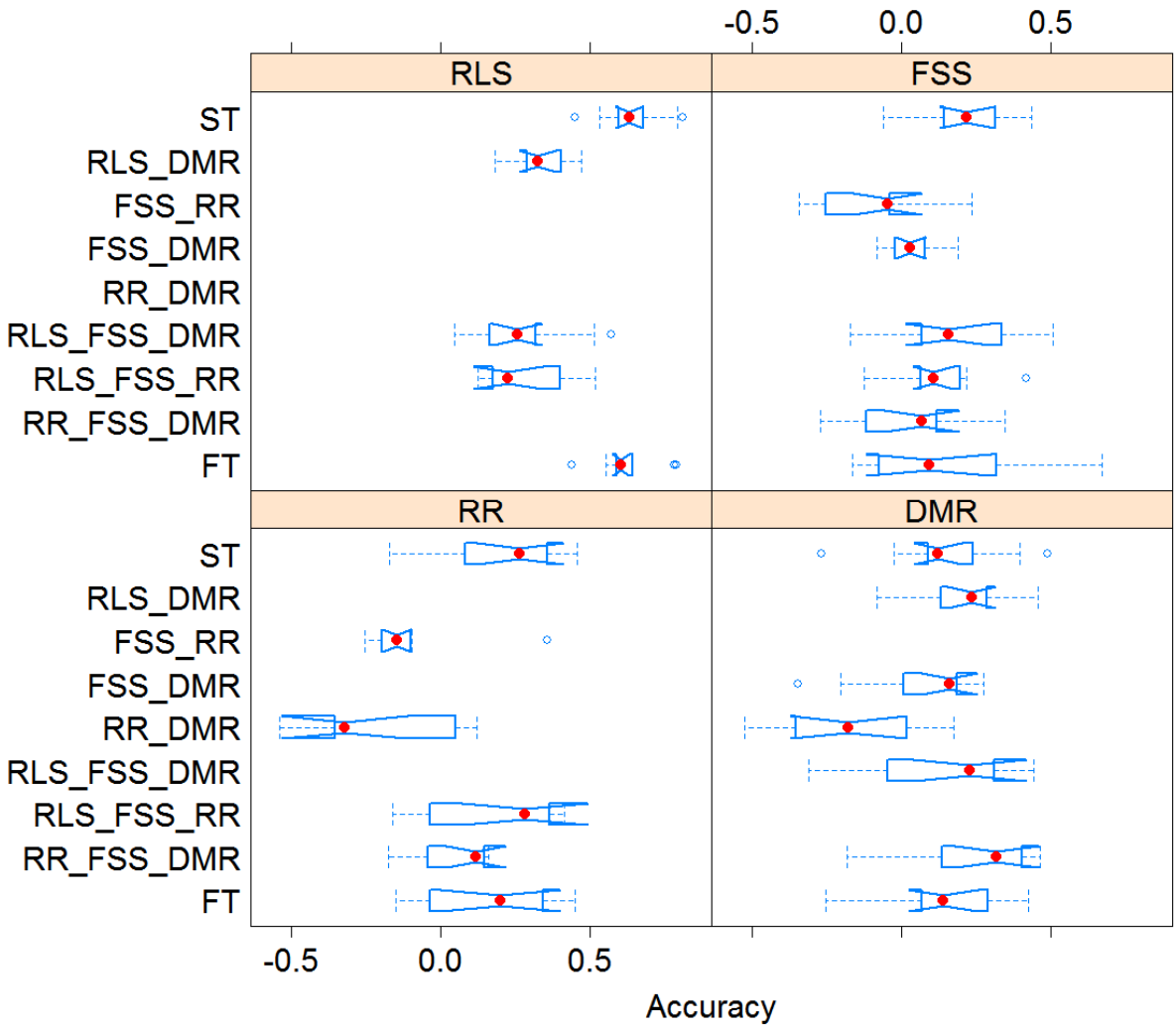

Fig. 10. Prediction accuracy with multi-trait genomic selection (GS) models compared with single-trait GS models for four SDS resistance traits. RLS, root lesion severity; RR, root retention; FSS, foliar symptom severity; DMR, dry matter reduction. ST, single-trait model; RLS_FSS model for RLS and FSS; RLS_RR model for RLS and RR; RLS_DMR model for RLS and DMR; FSS_RR model for FSS and RR; FSS_DMR model for FSS and DMR; RR_DMR model for RR and DMR; RLS_FSS_DMR model for RLS, FSS, and DMR; RLS_FSS_RR model for RLS, FSS, and RR; RR_FSS_DMR model for RR, FSS, and DMR; and FT model for all four traits. Red dot represents median of accuracies for each model. Notch marks the 95% confidence interval for the medians.

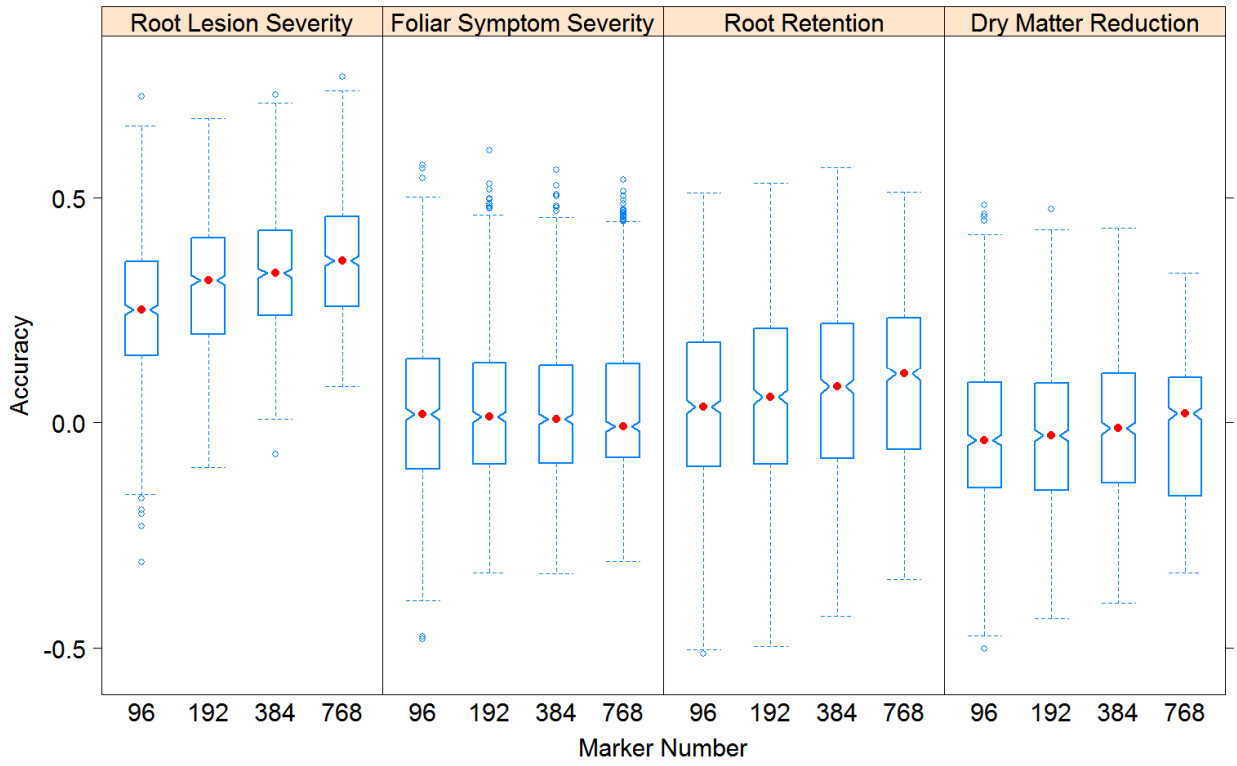

Fig. 11. Prediction accuracy with different numbers of markers for four SDS resistance traits. RLS, root lesion severity; RR, root retention; FSS, foliar symptom severity; DMR, dry matter reduction. Red dot represents median of accuracies for each subset of markers. Notch marks the 95% confidence interval for the medians.

Chapter 3

Prediction Accuracy and Response to Genomic Selection for Agronomic Traits in Soybean Breeding Populations

Recent studies suggest that genomic selection (GS) holds the potential to increase genetic gain for quantitative trait breeding in crop species. The objective of our study was to empirically assess the prediction accuracy and response to GS in a typical public soybean breeding program. A training set consisting of 282 common breeding parents and a validation set consisting of 273 current breeding lines were selected from the University of Minnesota soybean breeding program. We genotyped both training and validation sets using a genome-wide panel of 1536 single nucleotide polymorphism (SNP) markers. Existing historical trial data were then used to train the GS model. Validation set had been evaluated for yield, seed protein and oil in two-location trials in 2012. GS was then conducted to select top 20% individuals from the validation set. The selected lines were further evaluated in three-location trials in 2013. In 2012, our GS model predicted yield with significant positive accuracy in only two MN x MN crosses, while the prediction accuracy was near to zero or negative for protein and oil. In 2013, the prediction accuracy of our GS model was 0.88 for yield, 0.52 for protein, and 0.77 for oil. Moreover, one generation of GS didn't significantly change the population mean of yield, seed protein and oil content. This study suggests that the program-specific GS has erratic usefulness for predicting agronomic traits in the soybean breeding program.

Introduction

Important traits in soybean, like yield and seed composition, are genetically complex and highly polygenic. Complex traits show significant interaction with environment, so phenotypic selection (PS) for complex traits relies on extensive evaluation of replications of breeding candidates across years and locations to obtain accurate genetic potential. In addition, the reliable evaluation of complex traits typically cannot be assessed until after four to five generations of selfing in a conventional soybean breeding program. Marker-assisted selection (MAS) with flanking markers at an early breeding stage in soybean breeding has proven a useful strategy for plant breeders selecting for large-effect genes to enhance simply inherited traits (Young, 1999). However, MAS is largely ineffective in improving complex traits like yield because their genetic variance is controlled by a large number of genes with small effects. Given the difficulties and costs in complex trait breeding, genomic selection (GS) (Meuwissen et al. 2001) was developed to leverage abundant genome-wide markers to predict genetic potential of selection candidates without the need for phenotyping. Instead of detecting and utilizing molecular markers in significant association with targeted traits, GS uses the entire set of available genome-wide markers to capture the cumulative effects of all causative loci. Compared to PS, GS enables the evaluation of genetic potential of breeding candidates based on genome-wide markers in a timely and resource efficient manner, and thus is expected to increase genetic gain for the complex agronomic traits.

The utility of GS had been evaluated in actual breeding programs for a range of crop species including maize (*Zea mays* L.) (Albrecht et al., 2011; Combs and Bernardo, 2013; Crossa et al., 2013; Jacobson et al., 2014; Massman et al., 2013a, b; Riedelsheimer et al., 2013), soybean (*Glycine max* L.) (Bao et al., 2014), wheat (*Triticum aestivum* L.)

(Heffner et al., 2011a, b), barley (*Hordeum vulgare* L.) (Lorenz et al., 2012), cassava (*Manihot esculenta* L.) (Ly et al., 2013), oat (*Avena sativa* L.) (Asoro et al., 2013), rapeseed (*Brassica napus* L.) (Würschum et al., 2014), and sugarbeet (*Beta vulgaris* L.) (Würschum et al., 2013). Heffner et al. (2011a) found that average prediction accuracy using GS was 95% as accurate as PS for agronomic traits in wheat. Massman et al. (2013a) compared GS using genome-wide SNP markers with the MAS using a number of significant markers in a multi-cycle recurrent selection scheme, and suggested the GS was advantageous over MAS for selecting complex traits in maize. The moderate to high prediction accuracies as well as the competitive selection responses of GS from earlier studies have therefore indicated that GS holds great potential in genetic evaluation of breeding candidates and acceleration of breeding processes. In this part of my thesis, we seek to investigate the use of GS to select important agronomic traits as part of an ongoing public soybean breeding program.

Accurate genomic prediction relies on genetic relationships between a training set and validation set captured by genome-wide markers (Habier et al., 2007). Given the close genetic relationship within a biparental cross, high prediction accuracy and steady genetic gain can be expected in GS implementation (Combs and Bernardo, 2013; Massman et al., 2013a). However, biparental GS requires phenotypic data from a specific breeding population for model training which takes at least one additional year in commercial maize breeding program and a number of additional years for crop species without double haploid (DH) technology available, such as soybean. Moreover, the predictive ability of a biparental GS model is intrinsically restricted to its own population making this marker-based selection method less cost-effective and unfavorable to plant

breeders. There is a growing interest in multi-family (multi-cross) GS because it eliminates the need for phenotyping the specific populations and can be applied in related crosses. Multi-family GS model can be built with a training set consisting of diverse lines from several related crosses or an elite breeding program. Another advantage over biparental GS is that a multi-family GS can take advantage of existing breeding lines along with their genotypic and phenotypic data from a breeding program. Therefore, the multi-family GS model has the capacity to predict complex traits prior to phenotyping in related populations. Würschum et al. (2013, 2014) demonstrated that a training set consisting of diverse lines from a breeding program can generate comparable prediction accuracies to biparental families in rapeseed and sugarbeet. The selection response of agronomic traits to a general combining ability model constructed with breeding lines sharing one common parent were 68-76% of that with PS in maize (Jacobson et al., 2014). In the case that selection is conducted not only within a specific biparental population but also among the breeding lines across different crosses in an actual breeding program, multi-family GS is expected to be more effective than any population-specific GS models. Asoro et al. (2013) conducted two-cycle recurrent GS, MAS and PS in an oat breeding population initially consisting of 446 diverse lines and found a polygenic trait like beta-glucan content can be improved more effectively with the multi-family GS. However, pooling crosses that are unrelated or remotely related in the GS training set reduced prediction accuracies (Albrecht et al., 2011; 2014; Ly et al., 2013; Riedelsheimer et al., 2013; Würschum et al., 2013) and response to selection (Jacobson et al., 2014).

Previous studies have shown that existing phenotypic data in combination with genomic data can be exploited to understand the underlying genetics of agronomic traits (Wang et al., 2012) and improve the selection precision to increase genetic gain (Ly et al., 2013). A considerable amount of breeding data have already been obtained at multi-location yield trials in the University of Minnesota soybean breeding program over the past decades. The dataset is potentially useful for constructing a multi-family GS model since it includes thousands of genotypes tested for yield, lodging, maturity, protein, oil, seed weight, seed quality as well as disease scores in numerous locations across Minnesota. One constraint to using breeding program-derived phenotypic data is the datasets tend to be highly unbalanced. Breeding programs usually evaluate different sets of breeding lines in different experiments because they differ in stage of testing or breeding objectives. The effects of markers estimated from GS model may vary substantially among different environments (Burgueño et al., 2012; Crossa et al., 2010), which could undermine the predictive ability of GS. To accommodate historical trial data, a two-stage computational strategy was pursued in previous GS studies (Asoro et al., 2011; Heffner et al., 2011a; Rutkoski et al., 2012). In the first step, linear mixed model analysis has been commonly used to obtain best linear unbiased predictions (BLUP) for the lines tested in an unbalanced design (Smith et al., 2005). Subsequently, the BLUP value can be fitted as the response variable in GS models.

The main goal of the current study was to evaluate the utility of GS for important agronomic traits using historical trial data in a public soybean breeding program. The specific objectives of this study were to: 1) assess the prediction accuracy of a program-based multi-family GS within and across biparental populations; 2) evaluate the

responses to one cycle of GS; 3) determine the effect of training population design on genomic prediction accuracy for yield, protein, and oil.

Material and Methods

Population and Genotyping

A set of 282 common breeding parents were selected from the University of Minnesota soybean breeding germplasm collection (Bao et al., 2014) and used as a GS training set. The common breeding parents included ancestral lines, elite lines, advanced breeding lines, released public cultivars, and a number of plant introductions. The results from both *STRUCTURE* and principle component analysis (PCA) suggested three subpopulations existed in the training set corresponding to three distinct groups of germplasm generated in the University soybean breeding program: high yield, high protein, and small seeds (Bao et al. 2014; Fig. S5). The SNP-based kinship tree was coincident with actual pedigree knowledge (Fig. S6). Another set of 273 F_6-stage breeding lines from 13 recent crosses were then selected from the ongoing University breeding program as a validation set. The number of breeding lines in each population ranged from 13 to 25 with a mean of 20 (Table 8). Among the 13 crosses, seven crosses were made between MN varieties (MN x MN), four crosses were made between a MN variety and a variety from another state (MN x Other), one cross was made between a MN variety and Plant Introduction line (MN x PI), and one cross made between a MN variety and an ancestral cultivar (MN x Ancestor). All the lines were then genotyped using the Illumina GoldenGate 1536 SNP assay (Hyten et al., 2010) as described in Bao et al. (2014). The same SNP quality filter and imputation method were applied to the

training set and validation set, respectively: SNPs with greater than 5% minor allele frequency (MAF) and a missing data rate less than 50% were retained, followed by imputation of missing SNP data based on the population mean of each marker. In the training set, a genotype matrix of 1331 (SNP markers) x 271 (soybean lines) for yield, 1308 (SNP markers) x 196 (soybean lines) for protein, 1308 (SNP markers) x 194 (soybean lines) for oil passed the filters and were used in the subsequent analysis. In the validation set, only the SNP markers that passed the filters and were segregating in both training set and validation population (Table 8) were used in the subsequent analysis.

Historical Trial Data Analysis for Training Set

A two-stage computational strategy was pursued in our study: estimating the best linear unbiased prediction (BLUP) of individual lines followed by fitting the BLUPs as response variables in GS model. For the training set, archived breeding data including yield, seed protein and oil were extracted from the University soybean breeding database and Germplasm Resources Information Network (*www.ars-grin.org*). Crookston, Moorhead, Shelly, Morris, Rosemount, Lamberton, and Waseca were the seven test locations in Minnesota for the breeding lines in the validation set. Therefore, we only used the historical trial data from these seven locations for GS model training in this study. In brief, there were a total 2228 yield records for 280 lines across 30 years, and 2003 records for 243 lines across 28 years for protein and oil, respectively (Table S5). The variations of yield, protein, and oil across test locations and years were shown in boxplots (Fig. S8; Fig. S9). The full dataset for each trait is highly unbalanced because none of lines were tested in all locations and years. A linear mixed model, $y = u + Y + L$

+ G + e (Equation 1), was used to estimate the BLUP of individual lines to account for the environmental effects on the phenotypic values, in which y is a response variable, u is intercept, Y is fixed year effect, L is fixed location effect, G is BLUP of individual line, and e is a per-observation error term. The BLUP of each individual line was then used in the subsequent genomic modeling.

Phenotyping and Data Analysis for Validation Set

All breeding lines in the validation set were evaluated in preliminary yield trials (PYT) as part of Minnesota's ongoing breeding program in 2012. Because the breeding lines have different maturity, they were assigned to specific experiments in different test locations across MN based on their maturity evaluations. These test locations included Crookston, Moorhead, and Shelly for Maturity Group (MG)00; Morris and Rosemount for MG0; Lamberton and Waseca for MGI. Each line was evaluated for yield at two locations with two replications in each location. In the case of breeding populations targeting fatty acid, high protein and food-type cultivars, was each line evaluated for protein and oil at two locations with two replications. For remaining populations, seed protein and oil were only evaluated at one location with two replications for each line. The full dataset for each trait is therefore unbalanced across experiments as not all of lines were evaluated in the same locations. The mean of two replications for each line was used in the subsequent analyses. At least four check cultivars were set up in each experiment in each test location.

In each field experiment, the plot was mechanically seeded in two rows, spaced 76 cm apart. Planted plot dimensions were 3.1 m wide by 4.6 m long. Fertility and pest

management were performed according to standard management recommendations. Each plot was mechanically harvested for collecting grain weight and moisture data. Seed yield was adjusted with 13% seed moisture. Grain subsamples were collected from each plot for seed protein and oil concentration analysis using a Perten DA 7200 Feed Analyzer (Perten Instruments, Stockholm, Sweden).

A two-stage computation strategy was also pursued in analyzing the phenotypic data of validation set. To assess the prediction accuracy within each population, a linear mixed model, $y = u + L + G + e$ (Equation 2), was used to estimate the BLUPs along with the additive variance (V_a) and residual variance (V_e) within each population. In this analysis, y is a response vector, u is intercept, L is fixed location effect, G is BLUP of individual lines, and e is a per-observation error term. The BLUP of each line was used in the subsequent analysis. An entry-mean based heritability (h^2) was estimated as $V_a / (V_a + V_e)$ within each population. h^2 was tested against zero through t-Test: $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is h^2, n is the number of lines in each population.

We also assessed the prediction accuracy across populations within each maturity group. Because there were three major test regions (North, Central, and South) for testing soybean cultivars corresponding to three maturity groups (MG00, MG0, and MGI), the breeding lines evaluated in each test regions were pooled and the BLUPs of individuals were estimated using Equation 2 in each maturity group. An entry-mean based heritability (h^2) was estimated as $V_a / (V_a + V_e)$ within each maturity group. All linear mixed model analysis was conducted in R 3.0.2 (R Development Core Team, 2010). h^2

was tested against zero through t-Test: $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is h^2, n is the number of lines in each population.

Genomic Prediction Accuracy

In the training phase, a single-trait genomic prediction model for each trait was developed by fitting the phenotypic BLUPs estimated from Equation 1 as response variables and all genome-wide SNP markers as random variables in a ridge-regression best linear unbiased prediction (RRBLUP) model (Bernardo and Yu, 2007; Endelman, 2011; Meuwissen et al., 2001). Marker effect of each SNP was estimated from the model. In the prediction phase, the genomic estimated breeding value (GEBV) based on genotypic data was used to predict yield, protein, and oil for breeding lines in the validation set. Accuracy was calculated as correlation between GEBVs and phenotypic BLUPs estimated from Equation 2, divided by the square root of h^2 in each population or maturity group. All prediction accuracies were estimated with package "rrBLUP" in R 3.0.2 (R Development Core Team, 2010). Accuracy was tested against zero through t-Test: $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is the accuracy, n is the number of lines in each population.

Training Population Design

Selecting an appropriate set of lines and a sufficient number of markers to include in training population to ensure reliable prediction accuracy for targeted populations is referred as training population design in GS study. To investigate effect of training

population design on GS prediction accuracy, we constructed two types of GS models: one is GS_a trained only with all common breeding parents in the training set; the other is GS_b trained with both common breeding parents in the training set and current breeding lines except the population to be predicted in the validation set. Accuracy was tested against zero through t-Test: $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is the accuracy, n is the number of lines in each population. We also determined the effect of marker numbers on GS accuracy by including random samples of 96, 192, 288, 384, and 768 SNPs from the full marker set. This was repeated 100 times to avoid sampling bias of markers and the average of 100 replications was used to represent the prediction accuracy. Similarly, random samples of 48, 96, 144, 192, and 240 lines from the training set were used to determine the effect of training population sizes on GS accuracy. Four biparental populations in the validation set with moderate GS accuracy were selected to assess the effect of marker numbers and training population sizes on GS accuracy: two populations (M08-359 and M08-326) for yield, one population (M08-271) for protein, and one population (M08-208) for oil. All prediction accuracies were estimated with R package "rrBLUP" in R 3.0.2 (R Development Core Team, 2010).

Evaluation of Selection Response

A public soybean breeder selected top 20% of breeding lines in the validation set based on a comprehensive consideration including the GEBV of yield, protein, and oil, specific breeding goals of lines, and his general breeding experience. A total of 47 breeding lines were selected, and further advanced in new experimental line yield trials (NEL) as part of

routine breeding program in 2013. Still the breeding lines differed in maturity, so lines were assigned to specific experiments in different locations across MN based on their 2012 maturity evaluations. Test locations across MN included Becker, Danvers, Rosemount for MG0; Lamberton, Waseca, Westbrook for MGI. Each line was evaluated for yield at three locations with two replications at each location. Seed protein and oil were only evaluated at one location with two replications for each line. The full dataset for each trait is unbalanced across locations due to not all lines were evaluated in the same locations. A mean of two replication for each line was used in the subsequent analysis. At least four check cultivars were set up in each experiment in each test location. There are two shared check cultivars between 2012 and 2013: Sheyenne and MN1410.

Selection response was estimated as $R = (\mu_1 - \mu_0) - (Check_{2012} - Check_{2013})$ (Equation 3), where R is response, μ_1 is population mean of performance in 2013, and μ_0 is the mean of population performance in 2012, $Check_{2012}$ is the mean of the shared check cultivars in 2012, and $Check_{2013}$ is the mean of the shared check cultivars in 2013. Standard error (SE) of the response is: $\sqrt{\frac{V_{gl2013}}{47 \times 3} + \frac{V_{gl2012}}{273 \times 2} + \frac{V_{ck2013}}{2 \times 3} + \frac{V_{ck2012}}{2 \times 2}}$, where V_{gl2013} is genotype x location variance for 47 selected lines evaluated in 2013, V_{gl2012} is genotype x location variance for 273 lines evaluated in 2012, V_{ck2013} is genotype x location variance for the two shared check cultivars evaluated in 2013, and V_{ck2012} is genotype x location variance for the two shared check cultivars evaluated in 2012. A Z-test (z = R/SE) was conducted to determine the significance of response.

Since a severe drought occurred in Minnesota in 2012, it was possible that the drought stress had influenced the plant growth and field management dramatically leading to biased estimation of prediction accuracy. We also assessed the prediction accuracy for the 47 lines evaluated in 2013 following the same methods as previously described.

Results

Marker Effect

Genomic prediction model for each trait was developed by fitting the phenotypic BLUPs as response variables and all genome-wide SNP markers as random variables in the RRBLUP model. Marker effects of all SNPs were estimated simultaneously by solving the model. As expected, the distribution of marker effect for yield was more centered on zero than that of protein and oil (Fig. 12). This indicated that the genetic variance of yield was explained by markers with small effects, while the genetic variance of protein and oil was more likely explained by markers with small to moderate effects, which was consistent with the degrees of complexity of genetic architectures among the three traits.

Heritability and Prediction Accuracy within Populations in 2012

In the validation set, there were 273 F_6-stage breeding lines from 13 recent crosses from the ongoing University breeding program. The number of breeding lines in each population ranged from 13 to 25 with a mean of 20 (Table 8). We first estimated the entry-mean based heritability (h^2) within each population. Seed protein and oil were only evaluated at one location for each line in some populations, so the heritability of protein

and oil was not estimated in those populations. The average heritability was 0.51 with a range of 0.22-0.73 for yield, 0.57 with a range of 0.21-0.79 for protein, 0.59 with a range of 0.27-0.88 for oil, indicating the considerable interactive effect of genetic by environment on these three agronomic traits performance for breeding candidates (Table 8). The t-Test suggested that the heritability was significantly different from zero in ten out of 13 populations for yield and in three out of four populations for both protein and oil (Table 8).

Among the 13 crosses, seven crosses were made between MN varieties (MN x MN), four crosses were made between a MN variety and a variety from other state (MN x Other), one cross was made between a MN variety and Plant Introduction line (MN x PI), and one cross made between a MN variety and an ancestral cultivar (MN x Ancestor). The GS model developed based on a training set composed of common breeding parents of Minnesota program predicted yield with low to moderate accuracy (0.27-0.48) in five out of seven MN x MN crosses as well as the MN x Ancestor cross (Table 8). The predictive ability of GS for yield in MN x Other crosses and MN x PI crosses was poor (Table 8). The t-Test suggested that the accuracy for yield was significantly greater than zero in only two MN x MN crosses (Table 8). With respect to protein, four out of seven MN x MN crosses had low to moderate prediction accuracies (0.16-0.36), while the accuracies were erratic in all other crosses (Table 8). The t-Test suggested that the accuracy for protein was significantly greater than zero in only one MN x other cross (Table 8). Accuracies were low to moderate (0.15-0.34) for oil in MN x Other crosses and a MN x PI cross, and low to negative in the MN x MN crosses (Table 8). None of the accuracies for oil was significantly greater than zero in t-Test (Table 8).

Heritability and Prediction Accuracy within Maturity Group in 2012

We also assessed the prediction accuracy across populations within each maturity group. Because there are three major test regions in Minnesota (North, Central, and South) for testing soybean cultivars corresponding to three maturity groups (MG00, MG0, and MGI), the breeding lines evaluated in each test regions were pooled. The yield heritability was highest 0.82 in North (MG00), and lowest 0.29 in South (MGI) (Table 9). Seed protein and oil were only evaluated at one location for each line in some populations, so the heritability of protein and oil was not estimated in all test regions. For soybean lines grown in South (MGI), the prediction accuracy of GS was 0.26 for yield, 0.12 for protein, and 0.36 for oil (Table 9). For soybean lines grown in Central (MG0), the prediction accuracy was 0.28 for yield and 0.34 for protein, but only -0.12 for oil (Table 9). The t-Test suggested that the accuracy for both yield and protein was significantly greater than zero in central test regions (Table 9). For soybean lines grown in North (MG00), however, the prediction accuracy was negative for all three traits (Table 9).

Heritability and Prediction Accuracy in 2013

Compared to 2012, weather and soil moisture conditions were typical in Minnesota in 2013. We assessed the heritability and prediction accuracy for the 47 lines evaluated in 2013. The heritability was low as 0.12 for yield, while was moderate as 0.53 and 0.47 for protein and oil, respectively (Table 10). The prediction accuracies for all three traits were significant greater than zero: 0.88 for yield, 0.52 for protein, and 0.77 for oil (Table 10).

Training Population Design

By pooling all available lines including the training set and the validation set other than the population to be predicted, the GS model performed slightly better for yield in MN x MN crosses with all accuracies as positive (Table 8). The t-Test suggested that the accuracy for yield was significantly greater than zero in three MN x MN crosses (Table 8). However, prediction accuracies were still erratic for protein and oil in crosses with the exception of moderate accuracies in five out seven MN x MN crosses for oil (Table 8). The accuracy for oil in one MN x MN cross was tested as significantly greater than zero (Table 8).

We selected four individual biparental populations with moderate prediction accuracy as examples to determine how the number of markers and lines in training set affects GS accuracy. Among the four biparental populations, there were two populations (M08-359 and M08-326) for yield, one population (M08-271) for protein, and one population (M08-208) for oil. Within each of the four populations, the number of SNP markers that passed the quality filter and were segregating in both training set and validation population ranged from 610 to 943 (Table 8). Random samples of 96, 192, 288, 384, and 768 SNPs from the full marker set were included in GS modeling. In all four selected populations, increasing the number of markers from 96 to 768 improved prediction accuracy for all traits (Fig. 13). Overall, the effect of marker numbers was more substantial from 96 to 192, and from 384 to 768 (Fig. 13). In case of M08-326, the prediction accuracy increased from 0.14 for yield using 96 SNPs to 0.24 using 192 SNPs

(Table 8; Fig. 13). By contrast, only a slight improvement in prediction accuracy using 288 SNPs versus 384 SNPs was observed in M08-326 (Table 8; Fig. 13).

Similarly, random samples of 48, 96, 144, 192, and 240 lines from the training set were included in GS modeling. With more lines included in the training set, the prediction accuracy increased for all traits (Fig. 14). For example, the prediction accuracy was only 0.08 for yield with 48 lines in the training set compared to 0.48 using all available lines in M08-326 (Table 8; Fig. 14).

Selection Response

Genomic selection was conducted by selecting the top 20% of breeding lines with best predicted agronomic performance in the validation set. A total of 47 breeding lines were selected, and further evaluated in 2013. After adjusted by the shared check cultivars, the realized response to GS was -0.1 Mg ha^{-1} for yield, -0.2% for protein, and 0.1% for oil (Table 11). However, the responses were not significant according to the Z-test (Table 11). As expected, the responses of oil content in soybean seeds were negatively correlated with that of yield and protein (Table 11).

Discussion

Historical Trial Data for Training Set

Choosing optimal phenotypic data from a breeding program to include in the training set is critical but remains as a challenge for GS model training. Single year or single location data for GS training can lead to biased or inaccurate estimation of marker effect when the G x E effect plays a crucial role for quantitative traits. For an instance, we should expect

a low predictive ability of a GS model trained with phenotypic data collected from field trials in an unusual year with extreme weather conditions affecting plant growth and management dramatically. Multi-location and -year trials data are capable of capturing G x E effect to increase the stability of genomic prediction across locations and years. However, one must filter the specific environments to include in the model, because the prediction accuracy could be erratic when the test environments are distinct from training environments. The preliminary results from our study indicate that the genomic prediction for yield in MN x MN crosses was more consistent by using the marker effects estimated in target environments (the six test locations for the validation set), rather than across all available environments, including unrelated locations. In summary, to achieve consistent prediction in future implementation of GS, developing a GS model using highly relevant data from test regions is advised.

Unbalanced Breeding Data for Validation Set

The breeding data for the validation set used in our GS study were unbalanced mainly due to specific photoperiod requirements associated with soybean lines in different maturity groups. Soybean has been bred for adaptation to relatively narrow maturity ranges because it is a short-day crop and its development is largely determined by variety-specific day length requirements that initiate floral development. Soybean breeding lines derived from a cross involving parents with different maturities can have different maturity ratings, so that they need to be evaluated in corresponding latitude regions to achieve their maximum potential of agronomic performance. The resulting phenotypic data for any individual population will be unbalanced because not all lines in

the population are tested in the same set of environments. To account for the heterogeneous environmental effects within each population, we estimated the BLUPs from the unbalanced dataset using a mixed linear model, and subsequently fitted the BLUPs as the response variables in GS models. The BLUPs could be seen as shrunken magnitude of merit of the lines without effect on the rank of lines. We considered the BLUPs were more conservative estimations of genotypic value of individual lines compared to mean of phenotypic observations across environments.

Factors Affecting Prediction Accuracy

A consistent pattern of moderate prediction accuracy that was significantly greater than zero was only identified in the MN x MN crosses, while accuracies were erratic in the other crosses in 2012 (Table 8). This clearly suggests that the genetic relationship between training set and validation set was the primary driver of accurate GS prediction. Although breeders regularly introduced exotic germplasm into Minnesota breeding programs in the past, the average identity-by-state (IBS) value among 282 common breeding parents was approximately 0.7 (Fig. S7), indicating that soybean lines in the breeding germplasm collection were fairly closely related. In this situation, by assembling a large training population consisting of influential breeding parents from the breeding program, GS had the capacity to predict the performance of progeny lines in the ongoing breeding populations that share common ancestors only one or few generations in the past.

Given a moderate/low-density SNP array for GS, there is little reason to expect the SNPs to be causative variants. The phase and extent of LD between SNPs and

causative variants, therefore, are some of most important factors influencing SNP-array-based GS accuracy. The low or negative prediction accuracies may have been caused by differences in LD phases between SNPs and causative polymorphisms in the training and validation sets. In other words, a large portion of SNPs may have had different effects associated with training versus validation populations. Substantial SNP x population interaction effects were reported in elite maize breeding germplasm (Liu et al., 2011; Zhao et al., 2011). Previous studies in maize suggested the prediction accuracy in one specific population could be low and sometimes negative when the marker effects were estimated with only single cross data (Massman et al., 2013b). Responses of a GS model trained with random inbreds were significantly lower than that with PS for agronomic traits in maize (Jacobson et al., 2014). To achieve reliable genomic prediction accuracy in a population derived from a cross involving an unadapted parent, a possible strategy might be to train and apply a population-specific GS model within the specific cross.

To make the implementation of GS more cost-effective in a high throughput breeding, fewer marker data points providing comparable prediction accuracy would be desirable. The number of markers needed depends on the extent of LD between markers and causative variants as well as the genetic architecture of targeted traits. In cultivated soybean, Lam et al. (2011) observed an extensive average LD (~150 kb) based on the whole–genome sequence data. Within a biparental population, LD should be even more extensive and require comparatively few markers to capture causative variants. In a biparental GS study in wheat, 256 SNP markers were sufficient to predict grain quality traits with moderate accuracy (Heffner et al., 2011b). By contrast, our results showed that

the GS model using a set of 192 or 288 markers still failed to achieve the maximum prediction accuracy possible with all available markers (Fig. 13).

In this study, we observed moderate accuracies for yield but much lower accuracies for seed composition traits (protein and oil) in the MN x MN crosses (Table 8). The differences in GS prediction accuracies may reflect the degrees of complexity of genetic architectures between the traits. Specifically, yield is highly quantitative trait with more complex genetic architecture than that of seed composition (Fig. 12). Any differences in LD phases between SNPs and causative polymorphisms in training and validation sets should have a much larger effect on protein and oil than yield. If true, the rank of predictions would be changed dramatically, potentially leading to negative accuracies in cases where key associations between SNPs and causative protein or oil loci are out of phase in the corresponding validation populations. By contrast, prediction accuracy for yield is less sensitive to the changes in a few LD phases because the genetic variance of yield is likely controlled by hundreds or even thousands of genes with small effects (Fig. 12). Again, our results indicate GS is an appropriate genomic prediction approach for highly quantitative traits, while for less quantitative or qualitative traits, QTL mapping and conventional MAS are expected to be more effective.

Implementation of Genomic Selection

Despite the moderate to low accuracy for yield in this study, GS can be useful in the early stages of line development when phenotyping is typically not performed. In soybean breeding, selection among nearly-homogeneous F_4 or F_5 plants is typically based on plant type, disease resistance or other traits that can be reliably scored on a single-plant or

single-row basis. In a GS scheme, the genotypic potential of F_4 or F_5 plants can be predicted by genotyping the individuals with a set of genome-wide SNP markers based on a model like the one we developed in this study. This would allow progress from selection during these early stages in a conventional breeding program when selection for yield and other important quantitative traits is typically not performed.

Considering genetic gain per unit cost, GS is more favorable than PS because a maximum of two cycles of recurrent selection per year can be done with intensive use of year-round nursery involving a GS scheme for soybean. In the current PS scheme, the average accuracy of PS for yield is approximately 0.7 (square root of average heritability of 0.48) among MN x MN crosses (Table 8). By comparison, the average accuracy of GS for yield among MN x MN crosses is 0.35 per cycle (Table 8). Following the genetic gain equation: $\Delta G = h^2 \times \sigma_p \times r \times i \div L$ (Moose and Mumm, 2006), where ΔG is genetic gain, h^2 is narrow sense heritability, σ_p is phenotypic standard deviation, r is prediction accuracy, i is selection intensity, and L is generation interval, the genetic gain for GS is equivalent to that for PS when other factors in the equation are constant. Given rapid advances in genomic marker technology, genotyping cost becomes even lower than the cost of conducting multi-location and multi-year yield trials, GS has the potential to become a surrogate for PS in a high throughput breeding program.

Conclusion

The present study represents the first empirical assessment of prediction accuracy and responses to GS for yield, seed protein, and oil in a typical public soybean breeding program. We developed a program-specific GS model using a training set composed of

common breeding parents from the University of Minnesota breeding program. Our results indicate this program-specific GS model can predict yield with a significantly positive accuracy in only two MN x MN crosses in 2012, while the prediction accuracies for yield, protein, and oil were all significantly positive in 2013. Moreover, one generation of GS did not improve the population mean of yield, protein, and oil. Overall, we conclude that the program-specific GS has erratic usefulness for predicting agronomic traits in the soybean breeding program. To ensure reliable and consistent genomic prediction in future investigations, GS model should be developed with lines or populations closely related to the test populations, and using phenotypic data highly relevant to the test environments.

Table 8. Heritablity and genomic prediction accuracy in bi-parental populations in 2012

					Yield			Protein			Oil		
Category†	Population	Cross	Line‡	N_M§	h^2¶	r_a#	r_b††	h^2	r_a	r_b	h^2	r_a	r_b
MN x MN	M08-359	M02-391112 X MN1701CN	18	821	0.40	0.38	0.63**	NA	0.32	0.11	NA	0.21	0.27
	M08-395	MN0091 X M01-269046	24	1315	0.46*	0.01	0.16	NA	-0.25	-0.07	NA	-0.09	0.08
	M08-305	MN0907 X MN0071	21	745	0.54*	0.31	0.27	0.63**	-0.09	-0.20	0.58**	-0.16	-0.04
	M08-326	M03-177-3004 X MN0096SP	16	692	0.34	0.48*	0.50*	0.79**	0.19	0.36	0.88**	-0.03	0.15
	M08-434	M02-333013 X M02-328023	22	1331	0.50*	0.42*	0.71**	NA	0.36	-0.18	NA	0.03	0.46*
	M08-332	M02-375002 X MN0807SP	18	664	0.53*	0.30	0.03	NA	0.16	-0.06	NA	-0.06	0.25
	M08-369	M99-327049 X MN0304	17	684	0.63**	-0.20	0.18	0.65**	-0.33	-0.26	0.62**	0.05	-0.20
MN x Other	M08-271	M03-276016 X IA2064	25	610	0.73**	0.01	0.07	0.21	0.46*	0.00	0.27	0.15	0.33
	M08-208	MN0302 X LD05-16638	25	943	0.62**	-0.01	-0.08	NA	-0.20	0.08	NA	0.34	-0.08
	M08-344	LD00-2187 X MN0308CN	41	922	0.46**	0.10	0.21	NA	0.04	0.18	NA	0.17	-0.19
	M08-207	MN0107 X LD05-16413	13	943	0.22	-0.11	-0.77**	NA	-0.16	-0.17	NA	0.24	-0.47
MN x PI	M09-247	M98-297090 X PI603337A	15	1309	0.55*	-0.44	-0.66**	NA	-0.50*	-0.10	NA	0.19	0.20
MN x Ancestor	M08-391	SHEYENNE X M02-141020	20	684	0.67**	0.27	-0.56*	NA	-0.12	-0.07	NA	0.04	0.21

† MN x MN, MN x Other, MN x PI, and MN x Ancestor represent a cross made between a Minnesota variety and a Minnesota variety, a variety from other states, a Plant Introduction line, and an Ancestral line, respectively.

‡ Line, the number of breeding lines in each population.

§ N_M, the number of segregating SNP markers in each population.

¶ h^2, the entry mean based narrow-sense heritability of each population. NA, seed protein and oil were only evaluated at one location for

each line in 9 biparental populations, so that estimates of heritability were not assessable for these populations.

r_a, the genomic prediction accuracy adjusted by square root of the entry-mean heritability of each population with only historical lines and data in training.

†† r_b, the genomic prediction accuracy adjusted by square root of the entry-mean heritability of each population with both historical and biparental lines and data in training.

* The correlation was significantly different from zero based on t-Test. * $P \leq 0.05$, ** $P \leq 0.01$.

Table 9. Heritability and genomic prediction accuracy in three different test regions in 2012

		Yield		Protein		Oil	
Category†	Line‡	h^{2}§	r¶	h^2	r	h^2	r
North	42	0.82*	-0.05	NA	-0.49**	NA	-0.54**
Central	176	0.42*	0.28*	NA	0.34**	NA	-0.12
South	55	0.29*	0.26	NA	0.12	NA	0.36*

† North test region includes Crookston, Moorhead, and Shelly, MN; central test region includes Morris and Roseville, MN; and south test region includes Waseca and Lamberton, MN.

‡ Line, the number of breeding lines in each test region.

§ h^2, the entry-mean based heritability of each region. NA, seed protein and oil were only evaluated at one location for each line in 9 biparental populations, so that estimates of heritability were not assessable for protein and oil.

¶ r, the genoimc prediction accuracy adjusted by square root of the entry-mean heritability of each population.

* The correlation was significantly different from zero based on t-Test. * $P \leq 0.05$, ** $P \leq 0.01$.

Table 10. Heritablity and genomic prediction accuracy in 2013

Line†	Yield		Protein		Oil	
	$h^{2‡}$	$r^{§}$	h^2	r	h^2	r
47	0.12	0.88**	0.53**	0.52**	0.47**	0.77**

† Line, the number of breeding lines in each test region.

‡ h^2, the entry-mean based heritability.

§ r, the genoimc prediction accuracy adjusted by square root of the heritability.

* The correlation was significantly different from zero based on t-Test. * $P \leq 0.05$, ** $P \leq 0.01$.

Table 11. Population Mean and Realized Response to Genomic Selection

Trait	μ_0†	CK_{2012}‡	μ_{GS}§	CK_{2013}¶	R_{GS}#	P††
Yield	2.2	2.4	3.1	3.5	-0.1	0.46
Protein	35.3	33.5	36.7	35.1	-0.2	0.46
Oil	18.1	18.9	17.5	18.2	0.1	0.54

† The mean of base population in 2012. Yield, Mg ha^{-1}; protein and oil, %.

‡ The mean of two shared check cultivars in 2012.

§ The mean of selected lines based on genomic selection in 2013.

¶ The mean of two shared check cultivars in 2013.

The realized response ($\mu_{GS} - \mu_0$) - (CK_{2013}-CK_{2012}) to genomic selection.

†† P values of Z-test.

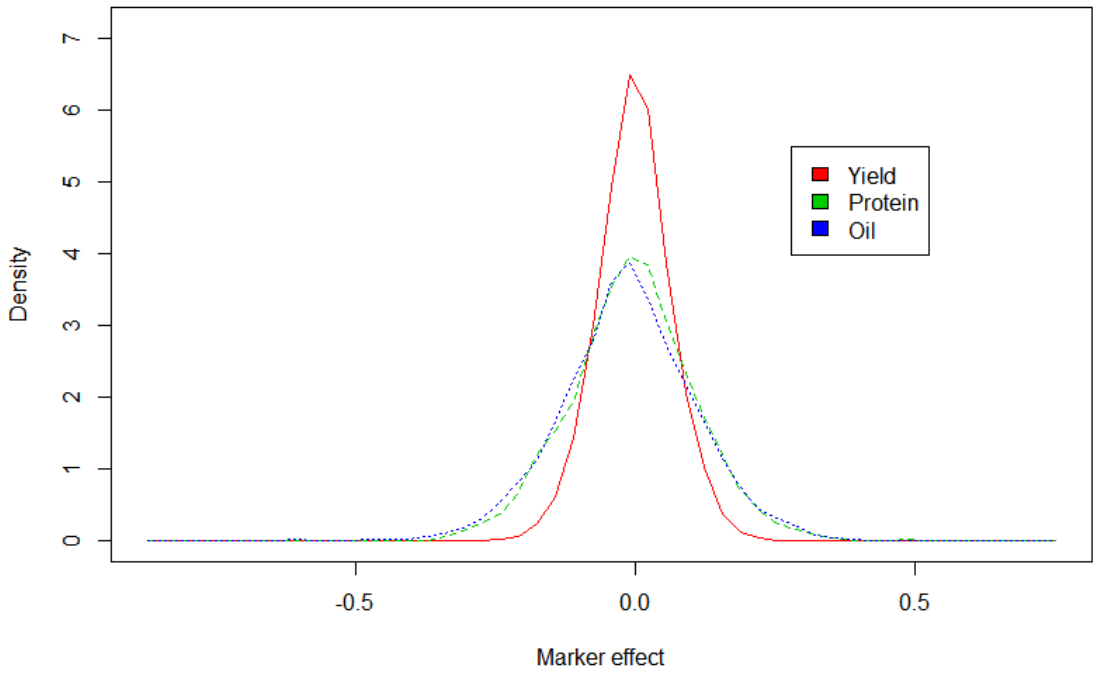

Figure 12. Density plot of marker effect estimated from ridge-regression best linear unbiased prediction (RRBLUP) for yield, protein, and oil.

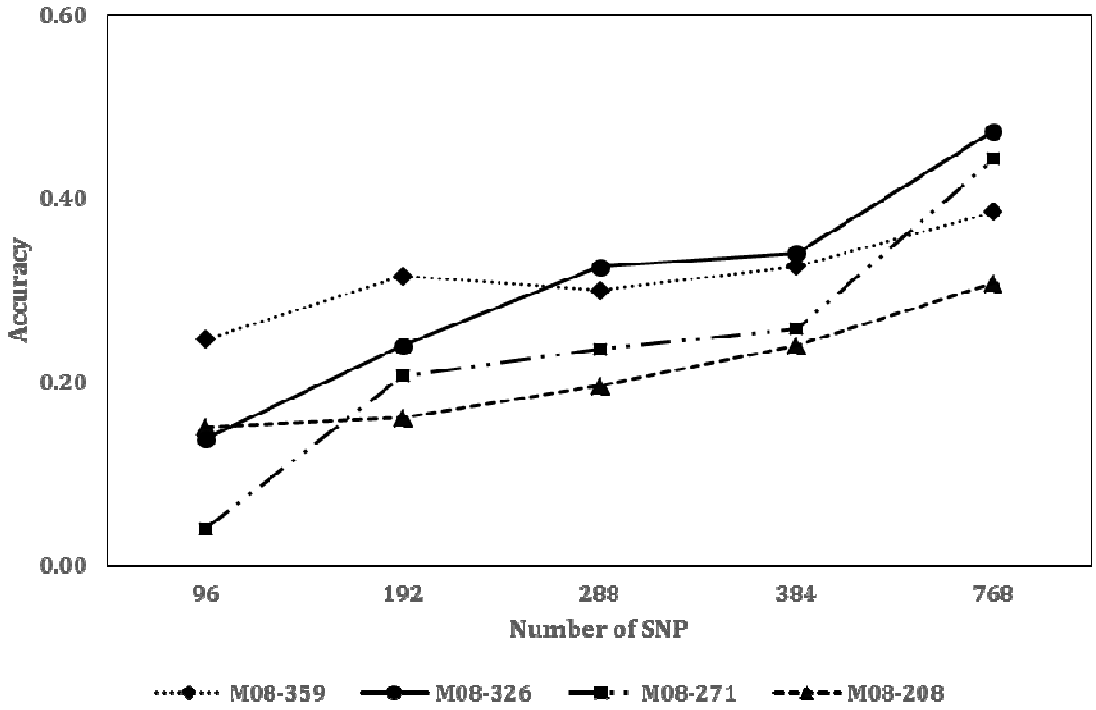

Figure 13. The mean of prediction accuracy with different number of SNP markers in genomic selection for yield, protein, and oil in four biparental populations. The prediction accuracy was the mean of 100 replications with random sampling of SNPs from the full marker set. The accuracy was assessed for yield in M08-359 and M08-326, for protein in M08-271, and oil in M08-208.

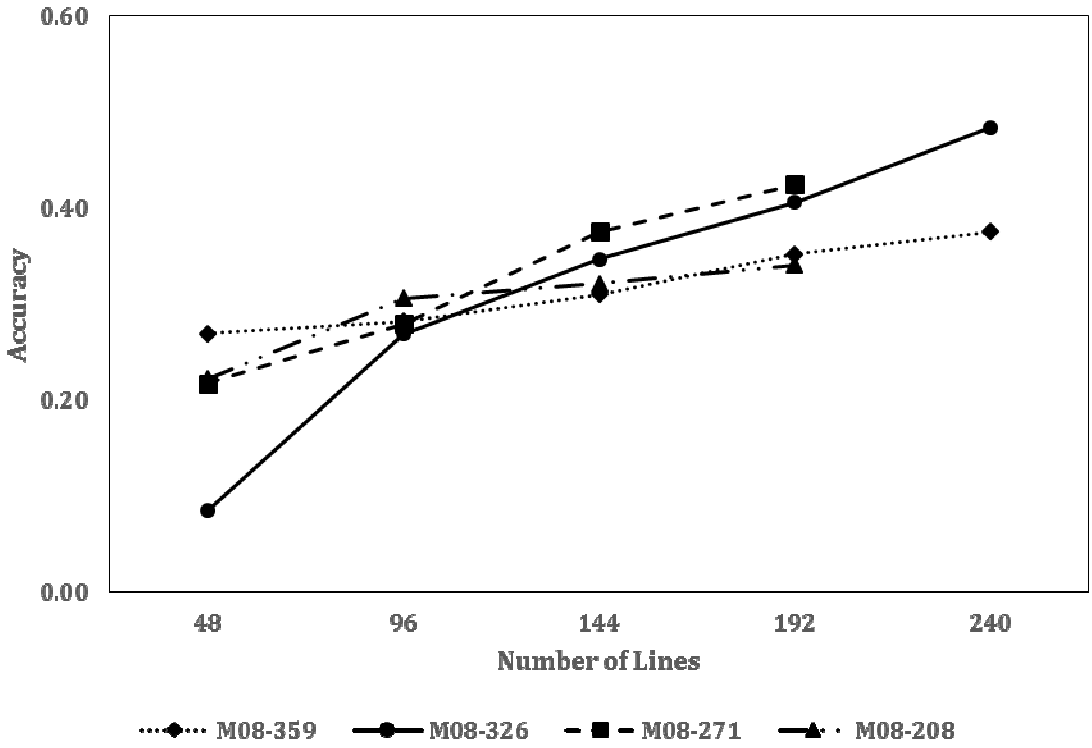

Figure 14. The mean of prediction accuracy with different number of training lines in genomic selection for yield, protein, and oil in four biparental populations. The prediction accuracy was the mean of 100 replications with random sampling of SNPs from the full marker set. The accuracy was assessed for yield in M08-359 and M08-326, for protein in M08-271, and oil in M08-208.

Bibliography

Afzal, A.J., A. Srour, N. Saini, N. Hemmati, H.A. El Shemy, and D.A. Lightfoot. 2012. Recombination suppression at the dominant *Rhg1/Rfs2* locus underlying soybean resistance to the cyst nematode. Theor. Appl. Genet. 124:1027-1039.

Albrecht, T., H. Auinger, V. Wimmer, J.O. Ogutu, C. Knaak, M. Ouzunova, H. Piepho, and C. Schön. 2014. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet.1-12.

Albrecht, T., V. Wimmer, H. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C. Schön. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123:339-350.

Aoki, T., K. O'Donnell, Y. Homma, and A.R. Lattanzi. 2003. Sudden-death syndrome of soybean is caused by two morphologically and phylogenetically distinct species within the *Fusarium solani* species complex--*F. virguliforme* in North America and *F. tucumaniae* in South America. Mycologia 95:660-684.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome 4:132-144.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, N.A. Tinker, and J. Jannink. 2013. Genomic, marker-assisted, and pedigree-BLUP selection methods for β-glucan concentration in elite oat. Crop Sci.

Bao Y, Vuong T, Meinhardt C, Tiffin P, Denny R, Chen S, Nguyen HT, Orf JH, Young ND (2014) Potential of Association Mapping and Genomic Selection to Explore PI88788 Derived Soybean Cyst Nematode Resistance. The Plant Genome doi:10.3835/plantgenome2013.11.0039

Barrett, J.C., B. Fry, J. Maller, and M. Daly. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Bernardo, R. 2010. Genomewide selection with minimal crossing in self-pollinated crops. Crop Sci. 50:624-627.

Bernardo, R. Genomewide selection when major genes are present.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082-1090.

Bernstein, E., Z. Atallah, N. Koval, B. Hudelson, and C. Grau. 2007. First report of sudden death syndrome of soybean in Wisconsin. Plant Dis. 91:1201-1201.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Brar, H.K., S. Swaminathan, and M.K. Bhattacharyya. 2011. The *Fusarium virguliforme* toxin FvTox1 causes foliar sudden death syndrome-like symptoms in soybean. Mol. Plant-Microbe Interact. 24:1179-1188.

Brown, S., G. Yeckel, R. Heinz, K. Clark, D. Sleper, and M.G. Mitchum. 2010. A high-throughput automated technique for counting females of *Heterodera glycines* using a fluorescence-based imaging system. J. Nematol. 42:201.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Sci. 52:707-719.

Chang, S., T. Doubler, V. Kilo, R. Suttner, J. Klein, M. Schmidt, P. Gibson, and D. Lightfoot. 1996. Two additional loci underlying durable field resistance to soybean sudden death syndrome (SDS). Crop Sci. 36:1684-1688.

Chawla, S., C.R. Bowen, T.L. Slaminko, H.A. Hobbs, and G.L. Hartman. 2013. A public program to evaluate commercial soybean cultivars for pathogen and pest resistance. Plant Dis. 97:568-578.

Chen, S., P.M. Porter, C.D. Reese, and W.C. Stienstra. 2001. Crop sequence effects on soybean cyst nematode and soybean and corn yields. Crop Sci. 41:1843-1849.

Chilvers, M., and D. Brown-Rytlewski. 2010. First report and confirmed distribution of soybean sudden death syndrome caused by *Fusarium virguliforme* in southern Michigan. Plant Dis. 94:1164-1164.

Combs, E., and R. Bernardo. 2013. Genomewide selection to introgress semidwarf maize germplasm into US corn belt inbreds. Crop Sci. 53:1427-1436.

Concibido, V.C., B.W. Diers, and P.R. Arelli. 2004. A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci. 44:1121-1131.

Cook, D.E., T.G. Lee, X. Guo, S. Melito, K. Wang, A.M. Bayless, J. Wang, T.J. Hughes, D.K. Willis, and T.E. Clemente. 2012. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science 338:1206-1209.

Crossa, J., G. de Los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, and J. Yan. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724.

Crossa, J., Y. Beyene, S. Kassa, P. Perez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueno, V.S. Windhausen, E. Buckler, J.L. Jannink, M.A. Lopez Cruz, and R. Babu. 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903-1926.

de Farias Neto, Austeclinio L, G.L. Hartman, W.L. Pedersen, S. Li, G.A. Bollero, and B.W. Diers. 2006. Irrigation and inoculation treatments that increase the severity of soybean sudden death syndrome in the field. Crop Sci. 46:2547-2554.

de Farias Neto, A.L., R. Hashmi, M. Schmidt, S.R. Carlson, G.L. Hartman, S. Li, R.L. Nelson, and B.W. Diers. 2007. Mapping and confirmation of a new sudden death syndrome resistance QTL on linkage group D2 from the soybean genotypes PI 567374 and 'Ripley'. Mol. Breed. 20:53-62.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375-385.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One 6:e19379.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4:250-255.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Gelin, J., P. Arelli, and G. Rojas-Cifuentes. 2006. Using independent culling to screen plant introductions for combined resistance to soybean cyst nematode and sudden death syndrome. Crop Sci. 46:2081-2083.

Gongora-Canul, C., and L. Leandro. 2011. Effect of soil temperature and plant age at time of inoculation on progress of root rot and foliar symptoms of soybean sudden death syndrome. Plant Dis. 95:436-440.

Gongora-Canul, C., and L. Leandro. 2011. Plant age affects root infection and development of foliar symptoms of soybean sudden death syndrome. Plant Dis. 95:242-247.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet. Sel. Evol. 43:7.

Guo, B., D. Sleper, P. Arelli, J. Shannon, and H. Nguyen. 2005. Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. Theor. Appl. Genet. 111:965-971.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., R.L. Fernando, and J.C. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Harren, J.E. 2013. Identification of QTL (s) Associated with Resistance to Sudden Death Syndrome (SDS) in Soybeans [dissertation]. University of Minnesota at Twin Cities.

Hartman, G., Y. Huang, R. Nelson, and G. Noel. 1997. Germplasm evaluation of glycine max for resistance to *Fusarium solani*, the causal organism of sudden death syndrome. Plant Dis. 81:515-518.

Heffner, E.L., J. Jannink, and M.E. Sorrells. 2011a. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4:65-75.

Heffner, E.L., J. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011b. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51:2597-2606.

Hnetkovsky, N., S. Chang, T. Doubler, P. Gibson, and D. Lightfott. 1996. Genetic mapping of loci underlying field resistance to soybean sudden death syndrome (SDS). Crop Sci. 36:393-400.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, and Z. Zhang. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961-967.

Hyten, D.L., I. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E. Hwang, L.K. Matukumalli, and P.B. Cregan. 2010. A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci. 50:960-968.

Iqbal, M., S. Yaegashi, R. Ahsan, K.L. Shopinski, and D.A. Lightfoot. 2005. Root response to fusarium solani f. sp. glycines: Temporal accumulation of transcripts in partially resistant and susceptible soybean. Theor. Appl. Genet. 110:1429-1438.

Iqbal, M., K. Meksem, V. Njiti, M.A. Kassem, and D. Lightfoot. 2001. Microsatellite markers identify three additional quantitative trait loci for resistance to soybean sudden-death syndrome (SDS) in Essex× Forrest RILs. Theor. Appl. Genet. 102:187-192.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo, 2014. General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54:895-905.

Jia, G., X. Huang, H. Zhi, Y. Zhao, Q. Zhao, W. Li, Y. Chai, L. Yang, K. Liu, and H. Lu. 2013. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nat. Genet. 45, 957-961.

Jia, Y., and J. Jannink. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513-1522.

Jin, H., G. Hartman, C. Nickell, and J. Widholm. 1996. Characterization and purification of a phytotoxin produced by *Fusarium solani*, the causal agent of soybean sudden death syndrome. Phytopathology 86:277-282.

Kassem, M., J. Shultz, K. Meksem, Y. Cho, A. Wood, M. Iqbal, and D. Lightfoot. 2006. An updated ‘Essex’ by ‘Forrest’ linkage map and first composite interval map of QTL underlying six soybean traits. Theor. Appl. Genet. 113:1015-1026.

Kazi, S., J. Shultz, J. Afzal, J. Johnson, V. Njiti, and D.A. Lightfoot. 2008. Separate loci underlie resistance to root infection and leaf scorch during soybean sudden death syndrome. Theor. Appl. Genet. 116:967-977.

Kim, M., D.L. Hyten, A.F. Bent, and B.W. Diers. 2010. Fine mapping of the SCN resistance locus from PI 88788. The Plant Genome 3:81-89.

Koenning, S.R., and J.A. Wrather. 2010. Suppression of soybean yield potential in the continental United States from plant diseases from 2006 to 2009. Plant Health Prog Http://dx.Doi.org/10.1094/PHP-2010-1122-01-RS.

Kurle, J., S. Gould, S. Lewandowski, S. Li, and X. Yang. 2003. First report of sudden death syndrome (*Fusarium solani* f. sp. *glycines*) of soybean in Minnesota. Plant Dis. 87:449-449.

Kurle, J., D. Malvick, C. Floyd and G. Anderson. 2010. Five years of monitoring foliar diseases of soybean in Minnesota. p. S66-S67. *In* Five years of monitoring foliar diseases of soybean in Minnesota. Phytopathology, 2010. AMER PHYTOPATHOLOGICAL SOC 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA.

Lam, H., X. Xu, X. Liu, W. Chen, G. Yang, F. Wong, M. Li, W. He, N. Qin, and B. Wang. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42:1053-1059.

Li, H., Z. Peng, X. Yang, W. Wang, J. Fu, J. Wang, Y. Han, Y. Chai, T. Guo, and N. Yang. 2012. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. 45: 43-50.

Lightfoot, D.A., P.T. Gibson, and K. Merkem. 2001. Soybean Sudden Death Syndrome Resistant Soybeans, Soybean Cyst Nematode Resistant Soybeans and Methods of Breeding and Identifying Resistant Plants. US Patent 6300541.

Liu, W., M. Gowda, J. Steinhoff, H.P. Maurer, T. Würschum, C.F.H. Longin, F. Cossic, and J.C. Reif. 2011. Association mapping in an elite maize breeding population. Theor. Appl. Genet. 123:847-858.

Long, N., D. Gianola, G.J. Rosa, and K.A. Weigel. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123:1065-1074.

Lorenz, A., K. Smith, and J. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52:1609-1621.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151-161.

Luckew, A.S., S.R. Cianzio, and L.F. Leandro. 2012. Screening method for distinguishing soybean resistance to in resistant × resistant crosses. Crop Sci. 52:2215-2223.

Luckew, A., L. Leandro, M. Bhattacharyya, D. Nordman, D. Lightfoot, and S. Cianzio. 2013. Usefulness of 10 genomic regions in soybean associated with sudden death syndrome resistance. Theor. Appl. Genet.1-13.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G. Dixon, P. Kulakow, and J. Jannink. 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. Crop Sci. 53:1312-1325.

Malvick, D., and K. Bussey. 2008. Comparative analysis and characterization of the soybean sudden death syndrome pathogen *Fusarium virguliforme* in the Northern United States. Canadian Journal of Plant Pathology 30:467-476.

Mamidi, S., S. Chikara, R.J. Goos, D.L. Hyten, D. Annam, S.M. Moghaddam, R.K. Lee, P.B. Cregan, and P.E. McClean. 2011. Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. The Plant Genome 4:154-164.

Massman, J.M., H.G. Jung, and R. Bernardo. 2013a. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53:58-66.

Massman, J.M., A. Gordillo, R.E. Lorenzana, and R. Bernardo. 2013b. Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126:13-22.

Meksem, K., T. Doubler, K. Chancharoenchai, N. Nijti, S. Chang, A.R. Arelli, P. Cregan, L. Gray, P. Gibson, and D. Lightfoot. 1999. Clustering among loci underlying soybean resistance to *Fusarium solani*, SDS and SCN in near-isogenic lines. Theor. Appl. Genet. 99:1131-1142.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

Mitchum, M.G., J.A. Wrather, R.D. Heinz, J.G. Shannon, and G. Danekas. 2007. Variability in distribution and virulence phenotypes of *Heterodera glycines* in Missouri during 2005. Plant Dis. 91:1473-1476.

Moose, S.P., and R.H. Mumm. 2008. Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol. 147:969-977.

Mueller, D., R. Nelson, G. Hartman, and W. Pedersen. 2003. Response of commercially developed soybean cultivars and the ancestral soybean lines to *Fusarium solani* f. sp. *glycines*. Plant Dis. 87:827-831.

Mueller, D., G. Hartman, R. Nelson, and W. Pedersen. 2002. Evaluation of *Glycine max* germplasm for resistance to *Fusarium solani* f. sp. *glycines*. Plant Dis. 86:741-746.

Navi, S.S., and X. Yang. 2008. Foliar symptom expression in association with early infection and xylem colonization by *Fusarium virguliforme* (formerly *F. solani* f. sp. *glycines*), the causal agent of soybean sudden death syndrome. Plant Health Progress Doi 10.1094/PHP-2008-0222-01-RS.

Niblack, T., G.L. Tylka, P. Arelli, J. Bond, B. Diers, P. Donald, J. Faghihi, V. Ferris, K. Gallo, and R.D. Heinz. 2009. A standard greenhouse method for assessing soybean cyst nematode resistance in soybean: SCE08 (standardized cyst evaluation 2008). Plant Health Progress Doi 10.1094/PHP-2009-0513-01-RV.

Njiti, V., J. Johnson, T. Torto, L. Gray, and D. Lightfoot. 2001. Inoculum rate influences selection for field resistance to soybean sudden death syndrome in the greenhouse. Crop Sci. 41:1726-1731.

Njiti, V., T. Doubler, R.J. Suttner, L. Gray, P. Gibson, and D. Lightfoot. 1998. Resistance to soybean sudden death syndrome and root colonization by *Fusarium solani* f. sp. *glycine* in near-isogenic lines. Crop Sci. 38:472-477.

Njiti, V., K. Meksem, M. Iqbal, J. Johnson, M.A. Kassem, K. Zobrist, V. Kilo, and D. Lightfoot. 2002. Common loci underlie field resistance to soybean sudden death syndrome in Forrest, Pyramid, Essex, and Douglas. Theor. Appl. Genet. 104:294-300.

Park, T., and G. Casella. 2008. The Bayesian Lasso. Journal of the American Statistical Association 103:681-686.

Pérez, P., G. de Los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. The Plant Genome 3:106-116.

Prabhu, R., V. Njiti, B. Bell-Johnson, J. Johnson, M. Schmidt, J. Klein, and D. Lightfoot. 1999. Selecting soybean cultivars for dual resistance to soybean cyst nematode and sudden death syndrome using two DNA markers. Crop Sci. 39:982-987.

Pritchard, J.K., M. Stephens, N.A. Rosenberg, and P. Donnelly. 2000. Association mapping in structured populations. The American Journal of Human Genetics 67:170-181.

R Development Core Team. 2005. R: A Language and Environment for Statistical Computing.

Radwan, O., Y. Liu, and S.J. Clough. 2011. Transcriptional analysis of soybean root response to *Fusarium virguliforme*, the causal agent of sudden death syndrome. Mol. Plant-Microbe Interact. 24:958-972.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5:94-100.

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J. Jannink, and A.E. Melchinger. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493-503.

Roy, K., D. Hershman, J. Rupe, and T. Abney. 1997. Sudden death syndrome of soybean. Plant Dis. 81:1100-1111.

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J. Jannink, and M. Sorrells. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome 5:51-61.

Sanogo, S., and X. Yang. 2001. Relation of sand content, pH, and potassium and phosphorus nutrition to the development of sudden death syndrome in soybean. Canadian Journal of Plant Pathology 23:174-180.

Schmidt, M., R. Suttner, J. Klein, P. Gibson, D. Lightfoot, and O. Myers Jr. 1999. Registration of LS-G96 soybean germplasm resistant to soybean sudden death syndrome and soybean cyst nematode race 3. Crop Sci. 39:598.

Schmitt, D., and G. Shannon. 1992. Differentiating soybean responses to *Heterodera glycines* races. Crop Sci. 32:275-277.

Schulz-Streeck, T., J.O. Ogutu, and H. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. Theor. Appl. Genet. 126:69-82.

Smith, A., B.R. Cullis, and R. Thompson. 2005. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. The Journal of Agricultural Science 143:449-462.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, and P.B. Cregan. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PloS One 8:e54985.

Srour, A., A.J. Afzal, L. Blahut-Beatty, N. Hemmati, D.H. Simmonds, W. Li, M. Liu, C.D. Town, H. Sharma, and P. Arelli. 2012. The receptor like kinase at Rhg1-a/Rfs2 caused pleiotropic resistance to sudden death syndrome and soybean cyst nematode as a transgene by altering signaling responses. BMC Genomics 13:368.

Stanton-Geddes, J., T. Paape, B. Epstein, R. Briskine, J. Yoder, J. Mudge, A.K. Bharti, A.D. Farmer, P. Zhou, and R. Denny. 2013. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. PloS One 8:e65688.

Sukumaran, S., W. Xiang, S.R. Bean, J.F. Pedersen, S. Kresovich, M.R. Tuinstra, T.T. Tesso, M.T. Hamblin, and J. Yu. 2012. Association mapping for grain quality in a diverse sorghum collection. The Plant Genome 5:126-135.

Technow, F., A. Bürger, and A.E. Melchinger. 2013. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3: Genes| Genomes| Genetics 3:197-203.

Triwitayakorn, K., V. Njiti, M. Iqbal, S. Yaegashi, C. Town, and D. Lightfoot. 2005. Genomic analysis of a region encompassing QRfs1 and QRfs2: Genes that underlie soybean resistance to sudden death syndrome. Genome 48:125-138.

Vaghchhipawala, Z.E., J.A. Schlueter, R.C. Shoemaker, and S.A. Mackenzie. 2004. Soybean FGAM synthase promoters direct ectopic nematode feeding site activity. Genome 47:404-413.

van Berloo, R., and R.C. Hutten. 2005. Peditree: Pedigree database analysis and visualization for breeding and science. J. Hered. 96:465-468.

Vick, C., S. Chong, J. Bond, and J. Russin. 2003. Response of soybean sudden death syndrome to subsoil tillage. Plant Dis. 87:629-632.

Wang, H., K.P. Smith, E. Combs, T. Blake, R.D. Horsley, and G.J. Muehlbauer. 2012. Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. 124:111-124.

Wang, S., D. Dvorkin, and Y. Da. 2012. SNPEVG: A graphical tool for GWAS graphing with mouse clicks. BMC Bioinformatics 13:319.

Winter, S.M., B.J. Shelp, T.R. Anderson, T.W. Welacky, and I. Rajcan. 2007. QTL associated with horizontal resistance to soybean cyst nematode in *Glycine soja* PI464925B. Theor. Appl. Genet. 114:461-472.

Wrather, J., and S. Koenning. 2009. Effects of diseases on soybean yields in the United States 1996 to 2007. Plant Health Prog.Doi 10:.

Würschum, T., S. Abel, and Y. Zhao. 2014. Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. Plant Breeding 133:45-51.

Würschum, T., J.C. Reif, T. Kraft, G. Janssen, and Y. Zhao. 2013. Genomic selection in sugar beet breeding populations. BMC Genetics 14:85.

Xu, X., L. Zeng, Y. Tao, T. Vuong, J. Wan, R. Boerma, J. Noe, Z. Li, S. Finnerty, S.M. Pathan, J.G. Shannon, and H.T. Nguyen. 2013. Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. Proceedings of the National Academy of Sciences 110:13469-13474.

Young, N.D. 1999. A cautiously optimistic vision for marker-assisted breeding. Mol. Breed. 5:505-510.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, and J.B. Holland. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Zhang, Z., E. Ersoz, C. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, and J.M. Ordovas. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42:355-360.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, and J.C. Reif. 2012. Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124:769-776.

Zhou, H., B. Steffenson, G. Muehlbauer, R. Wanyera, P. Njau, and S. Ndeda. 2014. Association mapping of stem rust race TTKSK resistance in US barley breeding germplasm. Theor. Appl. Genet.1-12.

Zhou, H., and B. Steffenson. 2013. Genome-wide association mapping reveals genetic architecture of durable spot blotch resistance in US barley breeding germplasm. Mol. Breed. 32:139-154.

Appendix

Table S1. A list of 282 soybean accessions in the association mapping (AM) panel.

ACME	M95-228092	M97-205062	M99-248011	MN1003SP	PI438265
AGASSIZ	M95-261096	M97-205082	M99-255012	MN1005	PI438445
AJMA	M95-265118	M97-205091	M99-255036	MN1006CN	PI438454
ALPHA	M95-273031	M97-205096	M99-274166	MN1007SP	PI445798
ALTONA	M95-273035	M97-205097	M99-278133	MN1008SP	PI445799
AMSOY	M95-274064	M97-206036	M99-278137	MN1009	PORTAGE
ANOKA	M95-274114	M97-207045	M99-278256	MN1010	PRIDE B216
ARCHER	M95-274129	M97-207052	M99-286047	MN1011CN	PROTO
BEESON	M95-274132	M97-209054	M99-286050	MN1101SP	SIBLEY
BELL	M95-275008	M97-209075	M99-286148	MN1104SP	SIMPSON
BERT	M95-278001	M97-251029	M99-286149	MN1105SP	STRIDE
BLACK KATO	M95-278007	M97-302004	M99-302003	MN1106CN	STURDY
BURLISON	M95-278022	M97-302128	M99-313054	MN1202SP	SURGE
CAPITAL	M95-279015	M97-304052	M99-316-1034	MN1203SP	SWIFT
CENTURY L2	M95-279022	M97-305077	M99-326040	MN1305SP	TOYOPRO
CHICO	M95-279028	M98-103039	M99-327049	MN1307	TRAILL
CHIPPEWA 64	M95-284082	M98-105090	M99-329038	MN1307SP	TRAVERSE
CLAY	M95-284113	M98-118006	M99-334034	MN1308SP	MINNATTO
COUNCIL	M95-287060	M98-134022	M99-334078	MN1406SP	UM3
DAKSOY	M95-287075	M98-210004	M99-337034	MN1407SP	VINTON
DWIGHT	M95-295100	M98-210060	M99-340047	MN1409SP	WALSH
EVANS	M95-295108	M98-211117	M99-341005	MN1410	WEBER
FREEBORN	M96-140012	M98-234042	M99-341028	MN1412SP	
GLENWOOD	M96-143031	M98-238010	M99-386097	MN1503SP	
GRANDE	M96-356055	M98-239080	MAPLE GLEN	MN1505SP	
GRANITE	M96-356062	M98-239263	MAPLE RIDGE	MN1603SP	
GRANT	M96-393101	M98-240104	MCCALL	MN1605SP	
HARK	M96-403029	M98-278072	MERIT	MN1606SP	
HAWKEYE	M96-412098	M98-279014	MN0081	MN1607SP	
HODGSON 78	M96-414071	M98-283034	MN0082SP	MN1802SP	
JACK	M96-414121	M98-283046	MN0091	MN1804CN	
JIM	M96-417038	M98-308007	MN0092	MN1805SP	
KASOTA	M96-417040	M98-308016	MN0095	MN1806SP	
KATO	M96-417149	M98-310021	MN0101	NORMAN	
LAMBERT	M96-452022	M98-310066	MN0102SP	OZZIE	
LESLIE	M96-452057	M98-310069	MN0103SP	PARKER	
M94-246028	M96-452059	M98-315056	MN0107	PELLA	
M94-275024	M96-452061	M98-324017	MN0201	PETERSON	
M94-278001	M96-473-3-1043	M98-324060	MN0203SP	PI180501	
M94-283002	M96-521075	M98-331009	MN0205SP	PI227565	
M95-101005	M96-745056	M98-332108	MN0207SP	PI257428	
M95-116024	M97-101025	M99-103172	MN0301	PI258385	
M95-118009	M97-101088	M99-113005	MN0402SP	PI297503	
M95-123006	M97-115063	M99-113168	MN0501SP	PI297532	
M95-123023	M97-120001	M99-118059	MN0502	PI347540C	
M95-123116	M97-121119	M99-121030	MN0603SP	PI347550B	
M95-202018	M97-121138	M99-137-1045	MN0606CN	PI372403A	
M95-206027	M97-158083	M99-204037	MN0804SP	PI437228	
M95-210133	M97-159146	M99-209070	MN0805SP	PI437267	
M95-211102	M97-164239	M99-215028	MN0901	PI437296	
M95-215050	M97-201070	M99-230063	MN0903SP	PI437610A	
M95-227016	M97-204114	M99-246068	MN0906SP	PI437994	

Table S2. A summary of SNP marker gaps across soybean genome in 1,536 SNP assay.

Dis[†]	Obs[‡]	Mean[§]	Median[§]	Min[§]	Max[§]	Freq[¶]
0-1	956	0.4	0.3	0.0	1.0	62.2
1-2	260	1.4	1.4	1.0	2.0	16.9
2-3	118	2.4	2.4	2.0	3.0	7.7
3-4	73	3.4	3.4	3.0	4.0	4.8
4-5	34	4.4	4.3	4.0	4.9	2.2
5-30	95	8.0	6.8	5.0	29.4	6.2
Total	1536	1.4	0.7	0.0	29.4	100.0

[†]Dis, distance range between adjacent SNP markers (cM).
[‡]Obs, number of SNP marker gaps in the specified range.
[§]Mean, median and minimum and maximum distance of adjacent SNP markers (cM) in the specific range.
[¶]Freq, frequency of SNP marker gaps of the specific range in all marker gaps (%).

Table S3. Extent of significant pair-wise intra-chromosomal LD across soybean genome.

Dis†	N‡	Mean§	Median§	Min§	Max§	Freq¶
0-5	1331	2.2	2.0	0.0	5.0	20.1
5-10	801	7.4	7.4	5.0	10.0	12.1
10-15	521	12.4	12.3	10.0	15.0	7.9
15-20	488	17.3	17.2	15.0	20.0	7.4
20-50	2100	33.5	32.7	20.0	50.0	31.8
50-150	1370	67.6	64.2	50.0	126.7	20.7

†Interval (cM) between pair-wise SNP markers in significant intra-chromosomal LD
‡Number of significant SNP marker pairs of the specified interval
§Mean, median and minimum and maximum distance (cM) of significant SNP marker pair interval
¶Frequency of SNP marker pairs of specific interval in all marker pairs

Table S4. Soybean lines exhibiting resistance to SDS for all four traits.

Name	Maternal[a]	Paternal[b]	RLS[c]	FSS[d]	RR[e]	DMR[f]
M99-121030	MN0301	MN1801	6.8	1.0	67.1	36.4
M99-204037	ND95-1215	M92-270029	6.0	1.0	73.4	44.5
M99-278133	MN0301	S-1990	4.0	1.0	89.4	14.8
M99-278256	MN0301	S-1990	6.5	1.0	64.0	32.9
MN0903SP	TOYOPRO	STURDY	6.8	1.5	63.7	41.7
MN1008SP	M91-202001	M91-151044	7.4	1.0	69.3	42.9
MN1009	M91-116124	MN1301	5.2	1.4	108.0	32.1
MN1409SP	PI592916	M93-402312E	6.3	1.0	80.8	33.3
MN1503SP	IA2011	KATO	5.8	1.0	63.7	47.9
MN1607SP	M90-764	M90-2144	4.6	2.6	64.9	48.9
PROTO	II-70-504	II-69-42	6.0	2.4	129.3	-3.5

a Maternal, female parent of soybean line.

b Paternal, male parent of soybean line.

c RLS, root lesion severity.

d FSS, foliar symptom severity.

e RR, root retention (%).

f DMR, dry matter reduction (%).

Table S5. Statistics of historical breeding data for 282 MN soybean lines used in the genomic selection training set.

Trait	Data points†	Lines‡	Year§	Loc¶
Yield	2228	280	30	6
Protein	2003	243	28	6
Oil	2002	243	28	6

† Total number of phenotypic records.

‡ Total number of soybean lines with available phenotypic records.

§ Total number of years with available phenotypic records.

¶ Total number of locations with available phenotypic records.

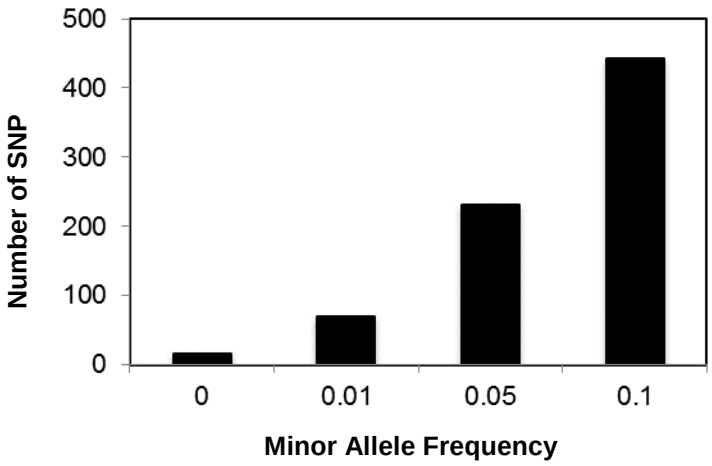

Figure S1. Distribution of minor allele frequency in 1,536 SNP array.

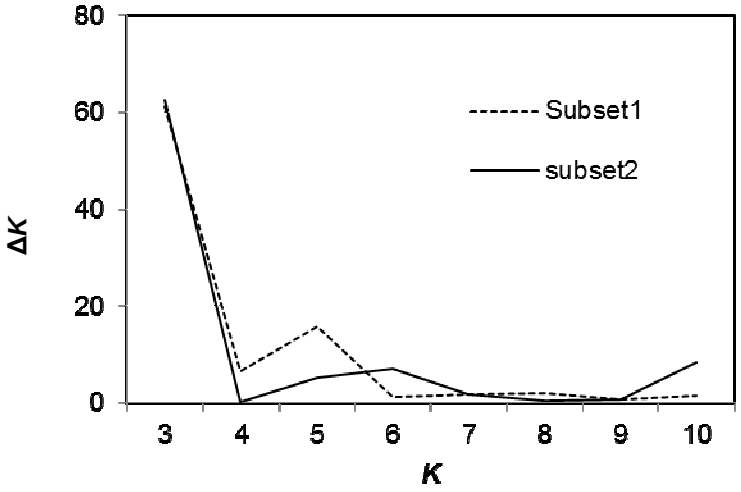

Figure S2. The natural logarithm probability difference between *k* and *k*-1 for two subsets of SNP markers based on ten independent runs in *STRUCTURE*. Posterior probability Pr(X|*k*) of each *k* was generated in *STRUCTURE*. Δ*k* represents the rate of lnPr(X|*k*) change from *k*-1 to *k*.

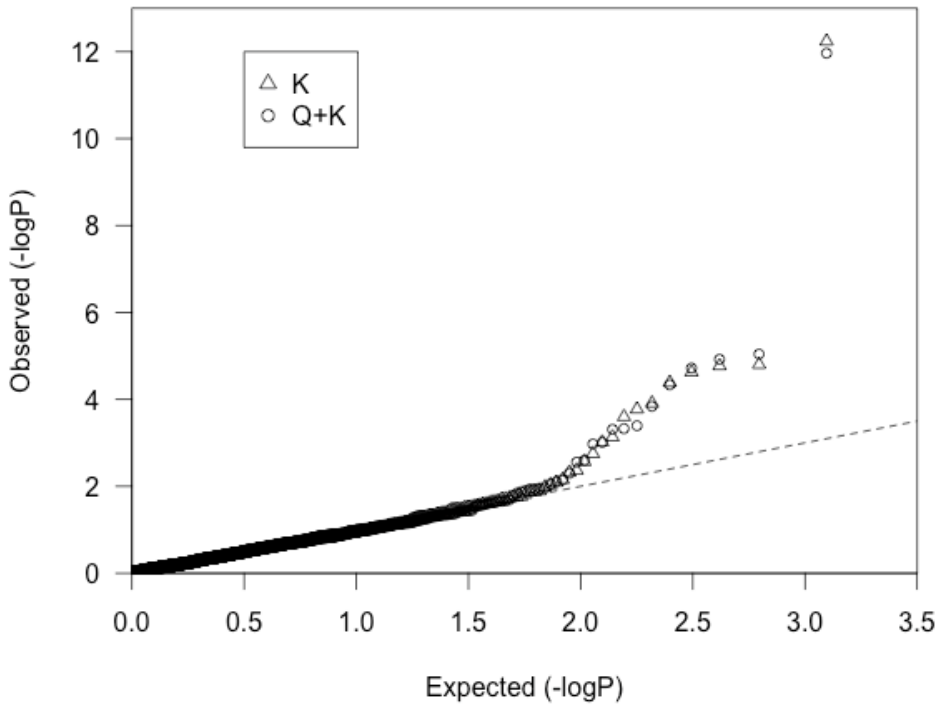

Figure S3. Quantile-Quantile plot with K and Q+K models. First 3 principal components were fixed in Q+K model.

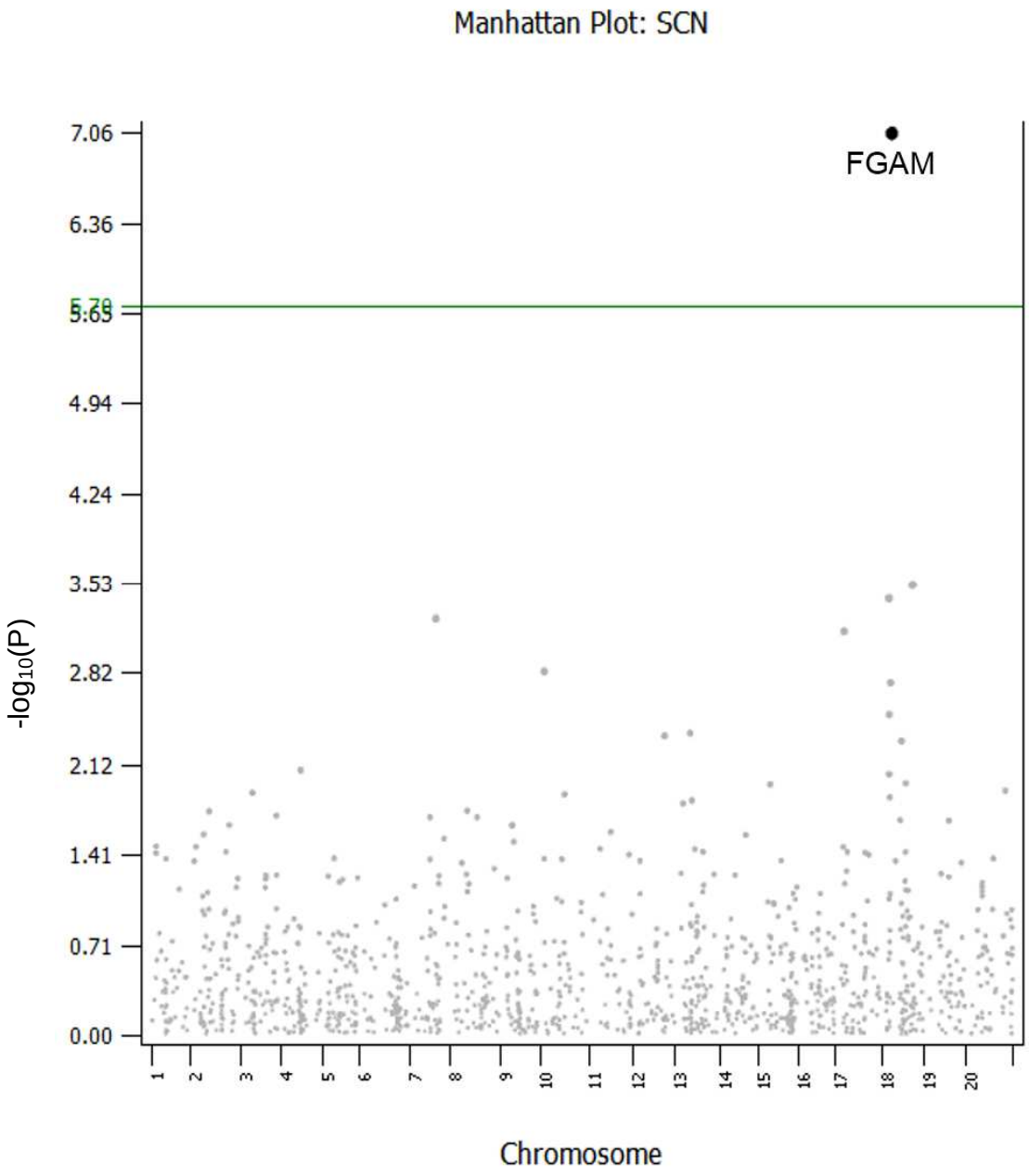

Figure S4. Manhattan plot of association mapping with Q+K model with *rhg1* locus as a fixed effect for SCN resistance. The green line represents the false discovery rate (FDR) of 5%.

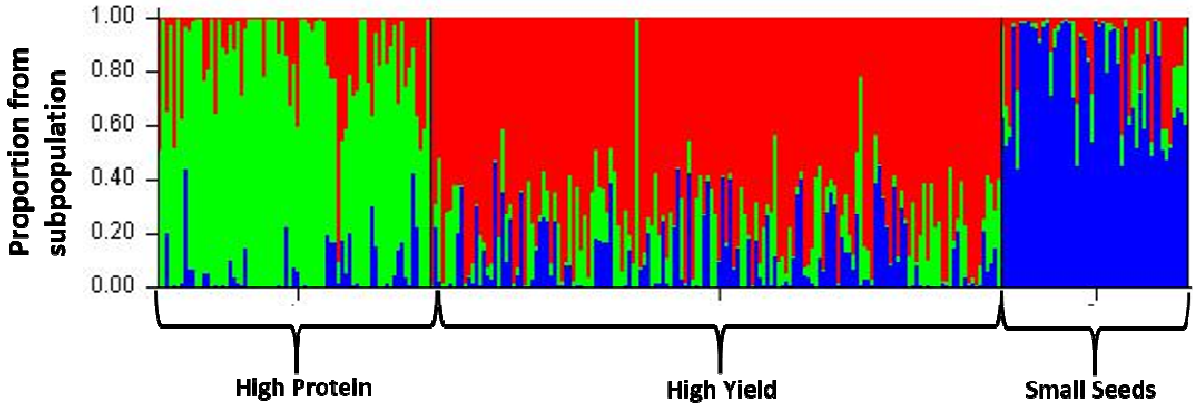

Figure S5 Population structure of 282 common breeding parents in the training set. Three subpopulations exist in the training set, representing high yield, high protein, and small seeds backgrounds. Subpopulation assignments were based on maximum membership probabilities for each soybean line. Membership probabilities were inferred from ten runs of STRUCTURE software using 227 single nucleotide polymorphism markers.

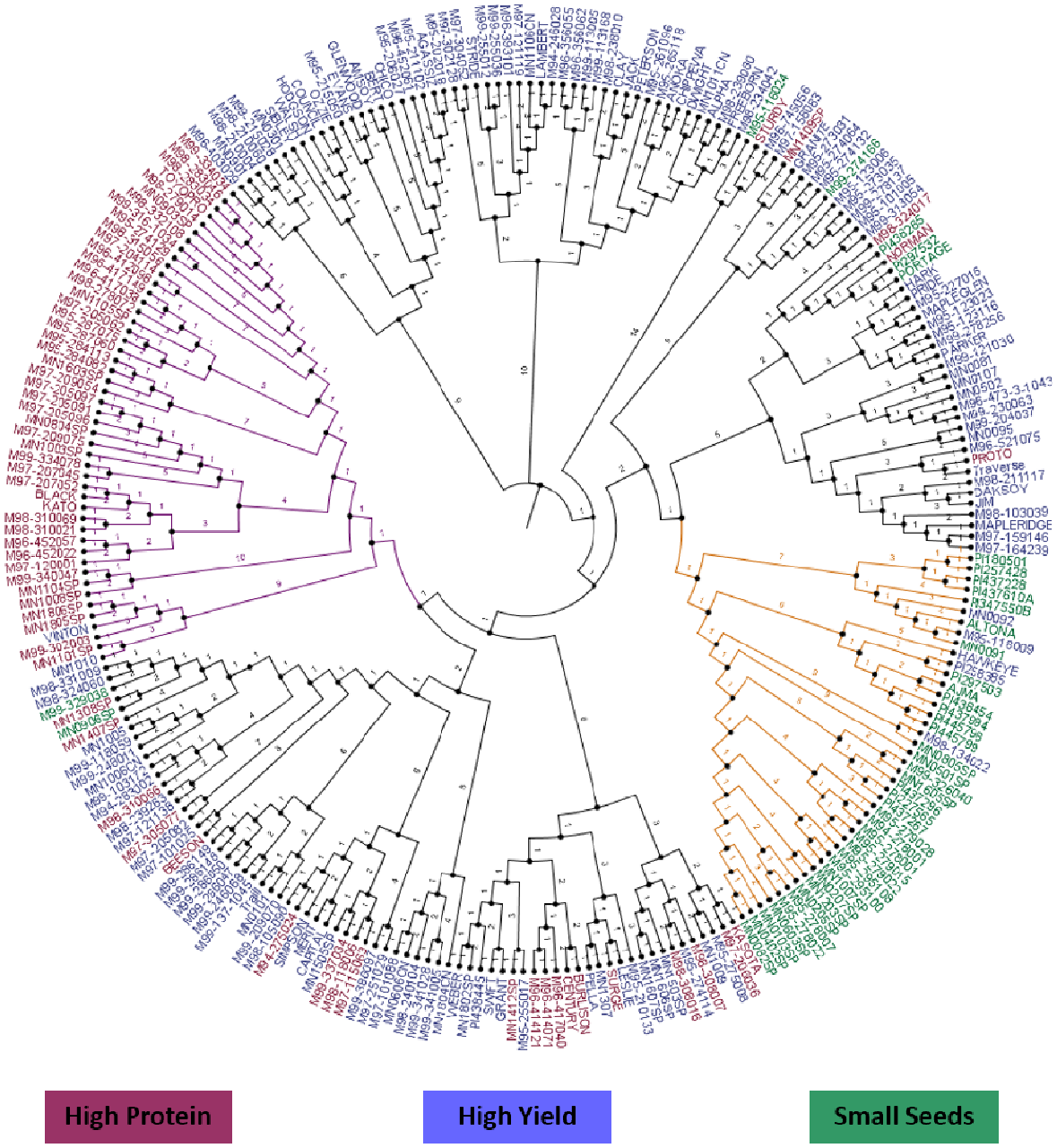

Figure S6 Cladogram based on SNP-based kinship matrix of common breeding parents in the training set. The SNP-based kinship matrix was generated by using Cladogram function implemented in TASSEL software. The lines in the group of “High Protein” are in purple, the lines in the group of “High Yield” are in blue, and lines in the group of “Small Seeds” are in green.

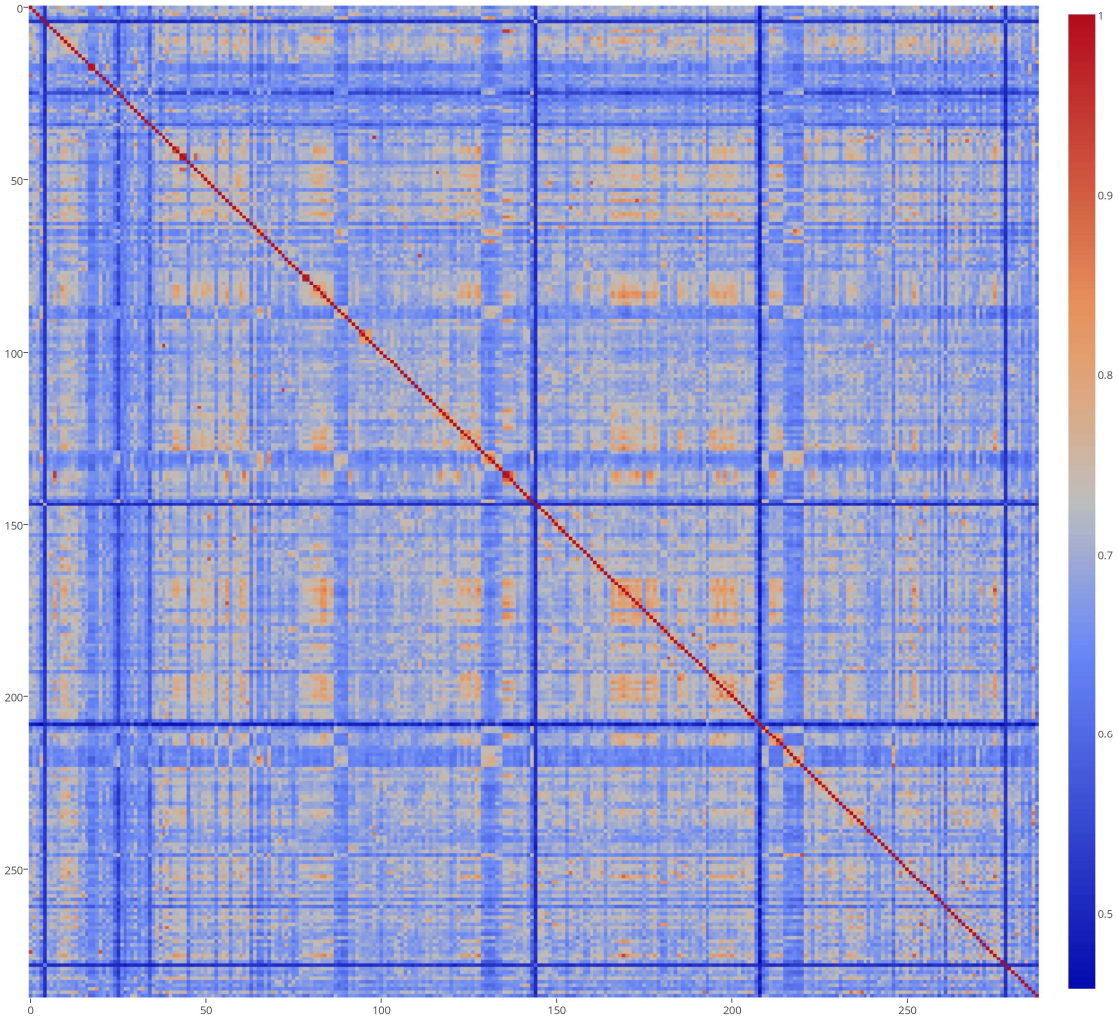

Figure S7 Heatmap based on pair-wise identity–by-state (IBS) values for soybean lines in the training set. IBS values for each pair of lines using a package “SNPRelate” in R. To avoid the spurious influence of SNP clusters in relatedness analysis, a LD-based pruned set of SNPs was used.

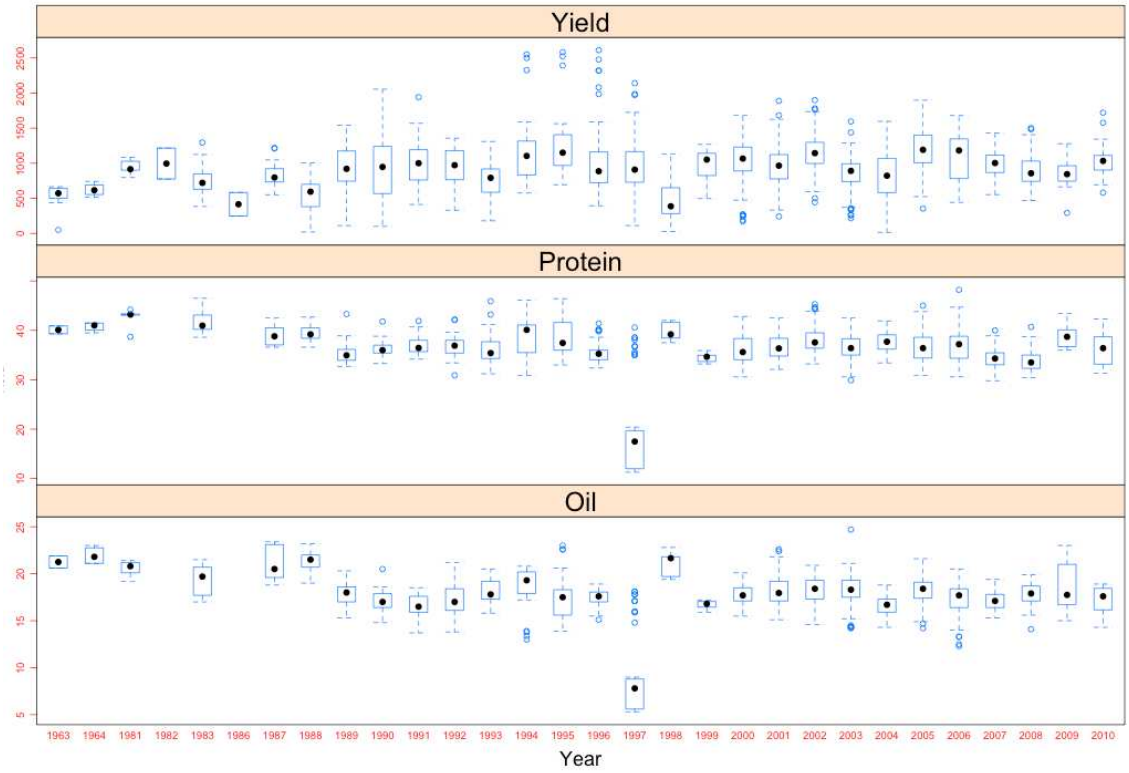

Figure S8 Variations of yield, protein, and oil for 282 common breeding parents across years. Yield, kg ha^{-1}; protein and oil, %.

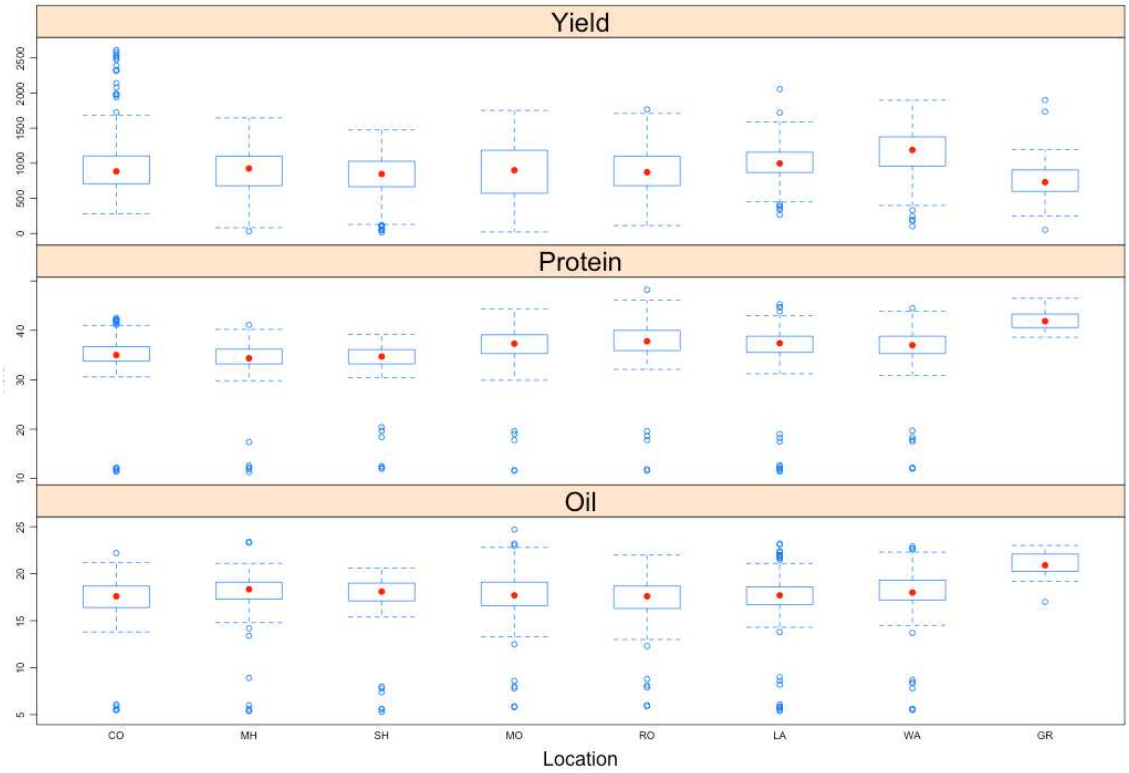

Figure S9 Variations of yield, protein, and oil for 282 common breeding parents across test locations. Yield, kg ha^{-1}; protein and oil, %. CO, Crookston, MH, Moorhead, SH, Shelly, MO, Morris, RO, Rosemount, LA, Lamberton, WA, Waseca, and GR, historical phenotypic data collected from Germplasm Resources Information Network (GRIN) website (www.ars-grin.gov).