Assessing Predictive Properties of Genome-Wide Selection in Soybeans

Alencar Xavier,* William M. Muir,† and Katy Martin Rainey*,[1]
*Department of Agronomy and †Department of Animal Science, Purdue University, West Lafayette, Indiana 47907

ABSTRACT Many economically important traits in plant breeding have low heritability or are difficult to measure. For these traits, genomic selection has attractive features and may boost genetic gains. Our goal was to evaluate alternative scenarios to implement genomic selection for yield components in soybean (*Glycine max* L. merr). We used a nested association panel with cross validation to evaluate the impacts of training population size, genotyping density, and prediction model on the accuracy of genomic prediction. Our results indicate that training population size was the factor most relevant to improvement in genome-wide prediction, with greatest improvement observed in training sets up to 2000 individuals. We discuss assumptions that influence the choice of the prediction model. Although alternative models had minor impacts on prediction accuracy, the most robust prediction model was the combination of reproducing kernel Hilbert space regression and BayesB. Higher genotyping density marginally improved accuracy. Our study finds that breeding programs seeking efficient genomic selection in soybeans would best allocate resources by investing in a representative training set.

KEYWORDS
bayesian methods
genomic selection
machine learning
SoyNAM

Soybean is a major crop used for human and animal consumption due to its ability to fix nitrogen and its unique seed composition (Board and Kahlon 2011; Chan *et al.* 2012). Genetic improvements in plant yield and quality can partially address increasing global demands for food quantity and quality. Unfortunately, genetic gains in soybean are often limited by its complex genomic properties (Hyten *et al.* 2007), resulting in low trait heritabilities. For such traits, genomic selection may outperform conventional breeding methods (Muir 2007) as well as having other promising and attractive features (Heffner *et al.* 2009; Jannink *et al.* 2010; Nakaya and Isobe 2012). Yet realistically, when resources are limited, many factors must be taken into account to optimize genetic gains (Meuwissen *et al.* 2001; Henryon *et al.* 2014; Poland 2015). Among the most important of these factors are: 1) training population size, 2) density of markers, and 3) prediction model. However, for several reasons, the genetic architecture of the population and the traits under consideration also affect these factors: linkage disequilibrium (LD) is population dependent (Hyten *et al.* 2007), traits differ in terms of heritability, and models differ in their assumptions and may not be effectively realized for some traits. As such, it is possible to determine these factors only by evaluating collected data pertaining to the populations and traits of interest.

Few studies have investigated genomic prediction in soybean (Jarquín *et al.* 2014; Xavier *et al.* 2016), and very little is known about the impacts on accuracy that these three factors have in this crop. We evaluated these factors using data collected from soybeans that were part of a nested association mapping (NAM) population. NAM is a next-generation experimental population, which is the result of crosses among single or multiple parent inbred lines followed by formation of recombinant inbred lines (Morrell *et al.* 2012). Guo *et al.* (2012) performed the first study in a NAM for genome-wide prediction (GWP), but focused on within-family prediction. However, a NAM is a complex structured population (Hamblin *et al.* 2011, Jannink *et al.* 2010) that can also be used for across-family prediction (Endelman *et al.* 2014), which is ideal for our objectives to determine the importance of: 1) training population size, 2) density of markers, and 3) prediction model on the accuracy of GWP in soybean.

MATERIALS AND METHODS

Genetic material

The data used in this research came from SoyNAM, a soybean nested-association panel. The SoyNAM population (soynam.org) contains

doi: 10.1534/g3.116.032268
Manuscript received March 4, 2016; accepted for publication June 16, 2016; published Early Online June 17, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.032268/-/DC1
[1] Corresponding author: Department of Agronomy, Purdue University, 915 W. State St., West Lafayette, IN 47907. E-mail: krainey@purdue.edu

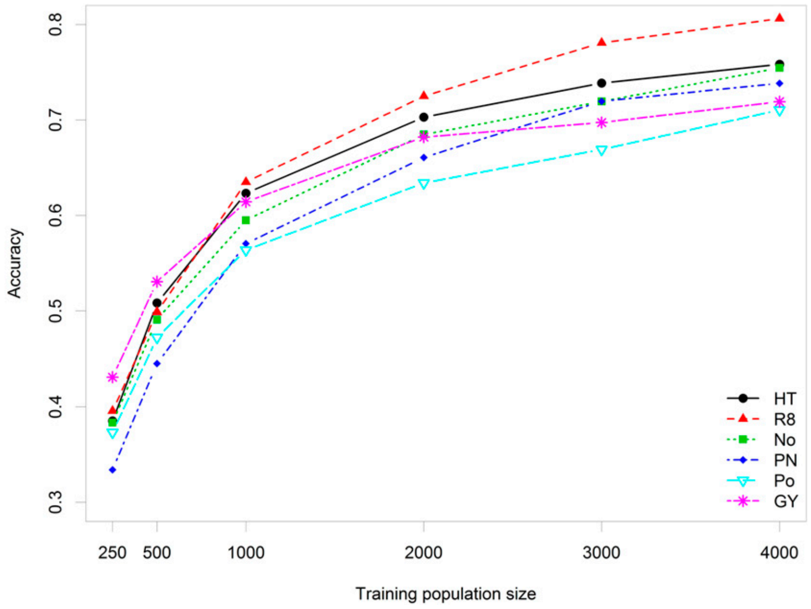

Figure 1 Effect of training population size on accuracy for six soybean traits. Plant height (HT), days to maturity (R8), number of reproductive nodes (No), pods per node (PN), number of pods (Po), and grain yield (GY).

5555 recombinant inbred lines (RIL) with maturity ranging from late maturity group II to early IV, derived from 40 biparental populations that share IA3023 as a common parent. Among the 40 founder parents, 17 lines are U.S. elite public germplasm, 15 have diverse ancestry, and eight are plant introductions. The genomic relationship among the lines is presented in Supplemental Material, File S1.

Lines were genotyped in the F_5 generation with a 5k single nucleotide polymorphism (SNP) chip. The chip was designed using SNPs discovered after complete sequencing of the DNA of all parental lines and, as such, is not biased by sampling issues associated with rare variants (Daetwyler *et al.* 2013; Heslot *et al.* 2013). The pairwise linkage disequilibrium between SNPs is shown in File S1. We removed nonsegregating SNPs and variants with a minor allele frequency (MAF) lower than 0.15 (Jarquín *et al.* 2014). We also removed redundant markers, such as markers in full disequilibrium (LD), so that the genotypic data would represent natural bins (Xu 2013b). We coded the remaining genotypes as 0, 1, and 2 (Strandén and Christensen 2011), and imputed missing SNPs using random forest implemented in the R package missForest (Stekhoven and Bühlmann 2012).

Field experimental design

Phenotypic data were collected from the SoyNAM population in 2013 and 2014 in West Lafayette, IN. We divided each of the 40 biparental families, with approximately 140 RILs each, into four blocks of 35 RILs each (40 families × 4 blocks = 160 subblocks). The 160 subblocks were randomly assigned into the field, and the 35 RILs within each subblock were also randomized. In both years, lines were planted during the third week of May in two-row plots (2.9 m × 0.76 m), at a density of approximately 36 plants/m^2.

Phenotypes

We evaluated six traits: grain yield, days to maturity, plant height, pod number, node number, and pods per node. Grain yield was measured in grams per plot adjusted to a standard moisture for soybeans seeds of 13%. We collected days to maturity three times a week, with back and forward scoring of plots that matured in the intervals. Using a barcode ruler, we measured plant height in three plants per plot. We also counted the number of reproductive nodes and pods in the main stem during the reproductive stages R7–R8 (Fehr *et al.* 1971), measuring three and six plants per plot for 2013 and 2014, respectively, with the count of pods per node (P/N) being the ratio of these data points.

Factors evaluated

Training population size: We sampled subsets of 250, 500, 1000, 2000, 3000, and 4000 RILs at random as a training set to predict a validation set of 500 RILs that were not included in the training set.

Density of markers: We tested subsets of the genotypic data as proposed by Meuwissen *et al.* (2001), using the whole panel, half panel, and quarter panel, corresponding to the 4077, 2039, and 1020 SNP markers, respectively. We formed the whole panel using all SNPs, the half panel by systematically choosing every other SNP, and the quarter panel by systematically choosing every fourth SNP.

Prediction models: We tested the prediction performance of four additive models (parametric), two kernel models (nonparametric), and each combination of additive and kernel model. Combining models is a strategy of ensemble learning that uses the kernel to account for background genetic effects and the additive term to capture markers with large effects (Kärkkäinen and Sillanpää 2012). This practice is frequently used to incorporate pedigree information into prediction models (Henderson 1986; Muir 2007, de los Campos *et al.* 2009, Heffner *et al.* 2009), but instead we used the molecular data to represent the relationship among genotypes.

The additive models we evaluated included BayesA, BayesB, BayesC, and Bayesian least absolute shrinkage and selection operator (BLASSO), and two kernel models, reproducing kernel Hilbert space (RKHS), and genomic best linear unbiased predictor (GBLUP). GBLUP was based on a single linear kernel (Xu 2013a) and RKHS was based on multiple Gaussian kernels (de los Campos *et al.* 2010). We fitted the models using the BGLR package (Pérez and de los Campos 2014). In-depth theoretical bases for the model building are described elsewhere (Sorensen and Gianola 2002; Kärkkäinen and Sillanpää 2012; Gianola 2013; de los Campos *et al.* 2013; Pérez and de los Campos 2014).

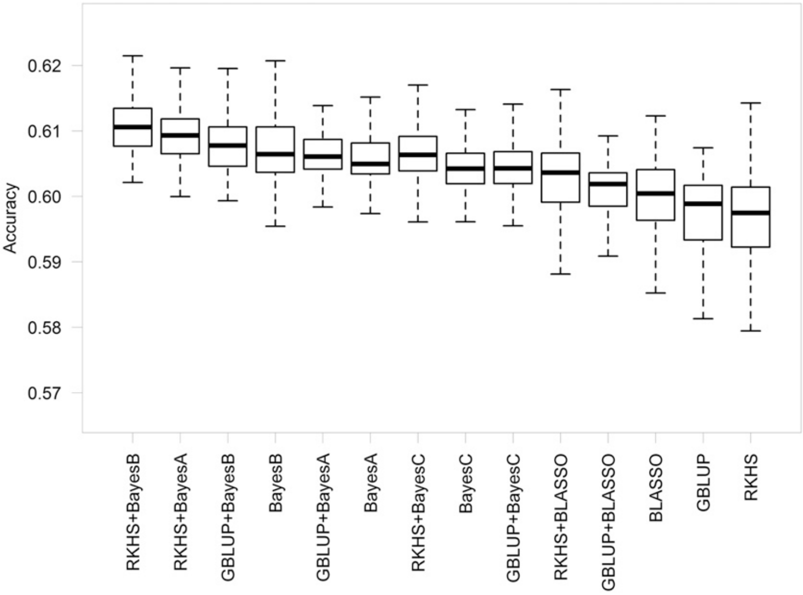

Figure 2 Boxplots of the accuracy of the genomic prediction models in soybeans tested in a variety of scenarios (*i.e.*, combinations of trait, number of SNPs, environment, and training population size). Whiskers represent the upper and lower limit, and the box represents the quartiles Q1 (25%), Q2 (median), and Q3 (75%). Models include additive methods (BayesA, BayesB, BayesC, and BLASSO), kernel methods (GBLUP and RKHS), and each combination of both. BLASSO, Bayesian least absolute shrinkage and selection operator; GBLUP, genomic best linear unbiased predictor; RKHS, reproducing kernel Hilbert space; SNP, single nucleotide polymorphism.

Method of evaluation

Predictive ability (PA) is defined as the correlations between observed (y) and predicted ($\hat{y}$) values computed through cross-validation (in which observations are not used to create the predictions). PA was based on 20 cross-validations for each combination of factors under evaluation. We estimated accuracy as PA divided by the square-root of heritability (Lehermeier *et al.* 2013).

Estimation of heritability (h^2) for each trait-year employed restricted maximum log-likelihood (REML) using the EMMA algorithm (Kang *et al.* 2008) implemented in the R package NAM (Xavier *et al.* 2015a) to solve a mixed model with an additive genomic covariance structure. Thus, avoiding the use of different whole-genome regression models to compute heritability (de los Campos *et al.* 2015). The mixed model was defined in probabilistic terms as $\mathbf{y} \sim \mathrm{N}(1\mu, \mathbf{ZGZ}'\sigma_a^2 + \mathbf{I}\sigma_e^2)$, where $\mathbf{y}$ is the vector of phenotypes of a given trait by year, μ is the overall mean, $\mathbf{Z}$ is the incidence matrix of genotypes, $\mathbf{G}$ is the genomic kinship matrix (VanRaden 2008), σ_a^2 is the additive genetic variance, and σ_e^2 is the residual variance. We computed heritabilities as $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$.

We limited the scope of the study to the impact of the defined factors upon accuracy. However, other measures of prediction properties have been suggested by Hastie *et al.* (2005) and Daetwyler *et al.* (2013), who used mean squared prediction error and prediction bias to identify problems with model fit.

Data availability

Phenotypes of yield, height, and maturity along with genotypes are available through the R package SoyNAM (Xavier *et al.* 2015b). Imputed genotypes and other phenotypes are available upon request.

RESULTS

Plant height was the most heritable trait, followed by grain yield and maturity, whereas the yield components appeared to be less heritable within the environment than yield itself. Figure 1 shows the effect of training population size on accuracy. Across traits, increasing the size of the training set from 250 to 4000 improved accuracy by 94.8% (from 0.384 to 0.747). Doubling the training population size increased accuracy by 9.1% on average. However, the improvements in accuracy decayed rapidly after 2000 individuals. Populations containing 1000–2000 individuals may represent the most cost-effective training sets, as gains become marginal for populations greater than 2000 individuals.

The number of markers was the least important factor for prediction in this study. Their effect on accuracy was marginal (1.33%). The use of all 4077 markers in comparison to 1020 increased accuracy from 0.600 to 0.607.

Figure 2 illustrates the performance of different prediction methods. Among the various prediction models, average accuracy ranges from 0.596 to 0.611. Further increases in accuracy were possible by selecting the optimum model for each scenario. The difference in accuracy among prediction models decreases as the training population size increases (File S1). The average improvement in accuracy achieved by selecting the optimal scenario model was 0.044 (3.84%). The best performing model for all traits was the combined model RKHS + BayesB (0.611).

We performed additional hypothesis testing using a *t*-test to compare between models to determine how specific model assumptions affect accuracy. Although differences between models were modest, models that included both an additive and kernel terms were significantly better than additive models alone ($P = 0.009$), and additive models were significantly better than kernel models ($P < 0.001$). Additive models combined with RKHS were significantly better than additive models combined with GBLUP ($P = 0.004$). Models with variable selection were significantly better than their all-included counterparts ($P < 0.001$).

DISCUSSION

Genomic estimated breeding values (GEBVs) generated through whole-genome regression (WGR) enable breeding programs to speed up the breeding process (Heffner *et al.* 2009; Endelman *et al.* 2014). Selection based on GEBVs is more reliable than that based on phenotypes alone or the traditional Quantitative Trait Loci (QTL) pyramiding (Nakaya and Isobe 2012), and it provides more genetic gains over the long term when compared to pedigree-based breeding values (Muir 2007). GEBVs are used to select unphenotyped material (Heffner *et al.* 2009), to perform more accurate selection of advanced lines (Endelman *et al.* 2014), to incorporate useful germplasm (Chung *et al.* 2014), to monitor the loss of genetic diversity (Henryon *et al.* 2014), and to select parental combinations for crosses (Mohammadi *et al.* 2015). Yet studies of GWP are important because the methodology for GEBV estimation is not fully understood (de los Campos *et al.* 2015), and the best prediction model varies among traits and from crop to crop (de los Campos *et al.* 2013). In addition, the contribution of genotyping density, population size, and phenotyping to genomic prediction is not clear when applied to real data (Wimmer *et al.* 2013). Prediction studies often provide conflicting results that vary according to the genetic basis of the population under evaluation (de los Campos *et al.* 2013).

Environmental factors

There are many strategies by which breeders can maximize genetic gains (Henryon *et al.* 2014). Robust breeding values rely on accurate phenotypic data collection and good environmental control by employing replications, checks, neighbor plot information, field plot techniques, and a well-planned field design (Heffner *et al.* 2009; Endelman *et al.* 2014). Similar results obtained in both the 2013 and 2014 environments (Table 1) indicate a stable level of genetic control across seasons. In this experiment, yield, height, and maturity were more heritable than yield components. We attribute the low heritability of yield components to their sensitivity to microenvironmental variation (Board and Kahlon 2011).

Replicated trials are not commonly used in GWP and mapping studies (Jannink *et al.* 2010). In this study, all cross-validations were performed within environment, which can affect the heritability and predictive ability in different ways (Endelman *et al.* 2014). In addition, genome-based heritability estimates in structured populations often provide lower values than pedigree-based estimates (Dekkers 2012). Nevertheless, results indicate that even low-heritable traits still provide reasonable accuracy. Muir (2007) pointed out that traits with low heritability have greater potential to be exploited. On the other hand, if accuracy is interpreted as the amount of genetic gains that genomic selection can exploit, less heritable traits with genomic data may provide accuracy comparable to more heritable traits.

Soybean yield components are commonly used as covariates in production systems to predict grain yield. The same approach should not be applied in genomic selection models targeting the genetic improvement of grain yield, because genetically correlated traits share additive genetic background (Valente *et al.* 2015). For breeding purposes, yield-component information is more suitable for enhancing grain yield, using multivariate models that accommodate the genetic relationship among traits (Rosa *et al.* 2011). In addition, low-heritable traits are favored by multivariate schemes (Sorensen and Gianola 2002).

Training population size

Training population size had the greatest impact on accuracy (Figure 1), which can determine the success of GWP. Two main properties of the training set are known to be critical to GWP: its relatedness to the validation set (Habier *et al.* 2007) and the population size (Nakaya and Isobe 2012). Good training sets must be representative of the germplasm under evaluation to capture the population structure and have a population size sufficient for an accurate estimation of allelic effects (Jannink *et al.* 2010). SoyNAM is a finite population with constrained structure, allowing the model calibration to become more accurate as the training set size increases. The remaining question is: what population size is sufficient for good prediction?

Quantitative traits are mostly controlled by alleles of small and medium effect, so that larger training sets will increase the signal-to-noise ratio (Muir 2007) and provide better learning properties (Okser *et al.* 2014). This potentially allows more accurate allelic effect estimates by minimizing the Beavis effect at the whole-genome level (Xu 2003). Besides the quantity of the training population, the quality also determines the success of prediction and long-term breeding (Bastiaansen *et al.* 2012). The quality of the training set with regard to its genetic variability depends on the effective population size, which is always smaller than the total number of genotypes. Soybean and other self-pollinated species often suffer from reduced effective population size because of their reproductive nature (Cowling *et al.* 2015; Hamblin *et al.* 2011) and narrow genetic basis restricted to elite germplasm (Hyten *et al.* 2006).

Table 1 Genomic heritability (h^2), average predictive ability [Cor(y, ŷ)], and accuracy in two environments (2013 and 2014) for six soybean traits

	h^2		Cor(y, ŷ)		Accuracy	
Trait	2013	2014	2013	2014	2013	2014
HT	0.522	0.478	0.459	0.418	0.635	0.604
R8	0.374	0.317	0.398	0.355	0.650	0.630
No	0.307	0.259	0.334	0.309	0.603	0.607
PN	0.238	0.189	0.275	0.258	0.563	0.593
Po	0.264	0.253	0.283	0.296	0.552	0.589
GY	0.494	0.409	0.423	0.399	0.602	0.623
Mean	0.366	0.317	0.362	0.339	0.601	0.608

h2, genomic heritability; Cor(y,ŷ), average predictive ability; HT, plant height; R8, days to maturity; No, number of reproductive nodes; PN, pods per node; Po, number of pods; GY, grain yield.

A sufficiently large training population size is also required when the ultimate goal is to perform selection of unphenotyped material (Heffner *et al.* 2009). When the training set is part of a breeding population being phenotyped and selected over generations, increasing the training population size is always beneficial to increase genetic gains (Bastiaansen *et al.* 2012; Hamblin *et al.* 2011; Muir 2007). In some cases, training population size is also critical for the choice of prediction model (Bastiaansen *et al.* 2012).

Prediction model

Varying the parameterizations of genomic information in prediction models to suit the particular genetic architecture of a trait can enhance prediction accuracy (Bastiaansen *et al.* 2012; Dekkers 2012; de los Campos *et al.* 2013). Pérez-Rodríguez *et al.* (2012) compared the performance of additive and kernel methods on two wheat traits across several environments, showing that BayesB better predicted yield grain while RKHS was the best model to predict days to heading. Similarly, Zhong *et al.* (2009) reported that GBLUP and BayesB are each better suited to different barley traits. Our results indicate that fitting parametric and semiparametric terms together provides a more robust prediction of soybean traits than either additive or kernel methods alone.

For the traits under evaluation, the combination of BayesB and RKHS provided the highest accuracy. Kärkkäinen and Sillanpää (2012) reported a synergy between BayesB and the semiparametric term, perhaps because kernels account for the relationship among individuals (Okser *et al.* 2014), while BayesB captures QTL in disequilibrium with markers in an additive fashion. The combination of a RKHS with BayesB includes flexible assumptions that account for different genetic interactions. RKHS enables the model to capture some level of epistasis (González-Camacho *et al.* 2012; Howard *et al.* 2014) with no assumptions about additive inheritance (de los Campos *et al.* 2009; Gianola *et al.* 2009), and BayesB allows markers to have large and/or null effect (Habier *et al.* 2011).

The decision to include kernels (pedigree or genomic) to account for the polygenic term in the prediction model depends on many factors, such as the marker density (Heffner *et al.* 2009), availability and complexity of pedigree data, and genetic architecture of the trait (de los Campos *et al.* 2013). Our results indicate that there is no advantage in utilizing kernel methods in this soybean population, in contrast to reports from simulations and studies with wheat and maize (González-Camacho *et al.* 2012; Pérez-Rodríguez *et al.* 2012; Howard *et al.* 2014).

In combined models, RHKS is a better complementary method than GBLUP. RKHS accounts for different levels of relationships among individuals due to the nonlinear nature of Gaussian kernels (de los Campos *et al.* 2010; González-Camacho *et al.* 2012). In kernel methods, markers are informative regardless of whether they are linked to any QTL (Habier *et al.* 2007), whereas null effect markers would harm any additive model incapable of performing efficient variable selection.

Our results indicate that models with a variable selection term provide better predictions. Efficient prediction often relies on consistent variable selection (Okser *et al.* 2014), especially in soybeans and other species that have a small genome, large LD blocks, and restricted diversity (Hyten *et al.* 2006, 2007; Chung *et al.* 2014), which together cause markers to present severe multicollinearity. Wimmer *et al.* (2013) showed that variable selection improves prediction in the presence of major effect genes and large populations in rice, wheat, and *Arabidopsis thaliana*.

Genotyping density

Higher genotyping density does not always increase accuracy (VanRaden *et al.* 2011), and subsets of the genotypic data sometimes outperform the entire dataset (Erbe *et al.* 2012). Xu (2013b) observed that artificial bins that compress genotypic information into fewer parameters could provide more accurate results than natural bins.

For the SoyNAM population, 1020 markers are enough to provide a consistent prediction, while higher density genotyping provides only marginal gains in PA. This result is likely associated with soybean's genomic properties, such as the existence of large disequilibrium blocks (Hyten *et al.* 2007) and the uneven distribution of SNPs in the soybean genome (Li *et al.* 2014).

The importance of larger SNP panels increases when the population structure is unknown, the number of selection cycles increases, and the LD between the QTL and marker decays (Bastiaansen *et al.* 2012; Daetwyler *et al.* 2013). In terms of allocating resources, our results support increasing population size over higher genotyping density, and using replicated trials when the number of genotypes in the training set is already sufficiently large (Lorenz 2013).

Conclusions

By comparing the accuracy associated with multiple factors, we showed that training population size is the main limiting factor for accuracy in soybeans. However, the rate of improvement decreased rapidly above 2000 individuals, suggesting that an optimal population size exists for the dataset in a study of between 1000 and 2000. The choice of prediction models was not unique for all scenarios. The best prediction model for this soybean population was the combination of RKHS and BayesB, which accommodates markers with large and null effect, also capturing some level of epistasis.

The value of next-generation populations to exploit new genomic frontiers is not limited to genome-wide associations. Prediction experiments based on real data provide an important insight for resource allocation, planning, and decision making in soybean breeding programs that aim to optimize genetic gains through genomic selection.

Soybean is a crop of worldwide importance that has shown limited rates of genetic gains. The use of genomic information through prediction models is a possible solution for more effective genetic improvement. In this study, we showed how genomic prediction acts in complex soybean traits in a variety of scenarios. The study shows how different factors affect the estimation of genomic values within environments. We believe that future directions for genome-wide prediction studies in soybeans should evaluate predictions across environments and across generations, as well as the optimal prediction procedures for genetic panels in ongoing selection.

ACKNOWLEDGMENTS

We thanks United Soybean Board for funding the field experiment in 2013 and Dow AgroSciences for funding the data collection and field experiment in 2014. We also thank Gustavo de los Campos for providing an important insight to the manuscript.

LITERATURE CITED

Bastiaansen, J. W., A. Coster, M. P. Calus, J. A. van Arendonk, and H. Bovenhuis, 2012 Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44: 3.

Board, J. E., and C. S. Kahlon, 2011 *Soybean Yield Formation: What Controls It and How It Can Be Improved. Soybean Physiology and Biochemistry*, pp. 1–36 in Soybean Physiology and Biochemistry, edited by H. El-Shemy. Intech, Rijeka.

Chan, C., X. Qi, M. W. Li, F. L. Wong, and H. M. Lam, 2012 Recent developments of genomic research in soybean. J. Genet. Genomics 39(7): 317–324.

Chung, W. H., N. Jeong, J. Kim, W. K. Lee, Y. G. Lee *et al.*, 2014 Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 21(2): 153–167.

Cowling, W. A., K. T. Stefanova, C. P. Beeck, M. N. Nelson, B. L. Hargreaves *et al.* 2015 Using the Animal Model to Accelerate Response to Selection in a Self-Pollinating Crop. G3 (Bethesda) 5: 1419–1428.

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193(2): 347–365.

Dekkers, J. C., 2012 Application of genomics tools to animal breeding. Curr. Genomics 13(3): 207.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1): 375–385.

de los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92(04): 295–308.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

de los Campos, G., D. Sorensen, and D. Gianola, 2015 Genomic heritability: what is it? PLoS Genet. 11(5): e1005048.

Endelman, J. B., G. N. Atlin, Y. Beyene, K. Semagn, X. Zhang *et al.*, 2014 Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. 54(1): 48–59.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95(7): 4114–4129.

Fehr, W. R., C. E. Caviness, D. T. Burmood, and J. S. Pennington, 1971 Stage of development descriptions for soybeans, *Glycine max* (L.) Merrill. Crop Sci. 11(6): 929–931.

Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194(3): 573–596.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183(1): 347–363.

González-Camacho, J. M., G. de los Campos, P. Pérez, D. Gianola, J. E. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125(4): 759–771.

Guo, Z., D. M. Tucker, J. Lu, V. Kishore, and G. Gay, 2012 Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor. Appl. Genet. 124(2): 261–275.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12(1): 186.

Hamblin, M. T., E. S. Buckler, and J. L. Jannink, 2011 Population genetics of genomics-based crop improvement methods. Trends Genet. 27(3): 98–106.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin, 2005 The elements of statistical learning: data mining, inference and prediction. Math. Intell. 27(2): 83–85.

Heffner, E. L., M. E. Sorrells, and J. L. Jannink, 2009 Genomic selection for crop improvement. Crop Sci. 49(1): 1–12.

Henderson, C. R., 1986 Estimation of variances in animal model and reduced animal model for single traits and single records. J. Dairy Sci. 69(5): 1394–1402.

Henryon, M., P. Berg, and A. C. Sørensen, 2014 Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livest. Sci. 166: 38–47.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink, and M. E. Sorrells, 2013 Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9): e74612.

Howard, R., A. L. Carriquiry, and W. D. Beavis 2014 Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda) 4(6): 1027–1046.

Hyten, D. L., Q. Song, Y. Zhu, I. Y. Choi, R. L. Nelson *et al.*, 2006 Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103(45): 16666–16671.

Hyten, D. L., I. Y. Choi, Q. Song, R. C. Shoemaker, R. L. Nelson *et al.*, 2007 Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175(4): 1937–1944.

Jannink, J. L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9(2): 166–177.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, and J. Jedlicka *et al.*, 2014 Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15(1): 740.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178(3): 1709–1723.

Kärkkäinen, H. P., and M. J. Sillanpää, 2012 Back to basics for Bayesian model building in genomic selection. Genetics 191(3): 969–987.

Lehermeier, C., V. Wimmer, T. Albrecht, H. J. Auinger, D. Gianola *et al.*, 2013 Sensitivity to prior specification in Bayesian genome-based prediction models. Stat. Appl. Genet. Mol. Biol. 12(3): 375–391.

Li, Y. H., Y. L. Liu, J. C. Reif, Z. X. Liu, B. Liu *et al.* 2014 Biparental resequencing coupled with SNP genotyping of a segregating population offers insights into the landscape of recombination and fixed genomic regions in elite soybean. G3 (Bethesda) 4(4): 553–560.

Lorenz, A. J. 2013 Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: A simulation experiment. G3 (Bethesda) 3(3):481–491.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829.

Mohammadi, M., T. Tiede, and K. P. Smith, 2015 PopVar: A genome-wide procedure for predicting genetic variance and correlated response in bi-parental breeding populations. Crop Sci. 55(5): 2068–2077.

Morrell, P. L., E. S. Buckler, and J. Ross-Ibarra, 2012 Crop genomics: advances and applications. Nat. Rev. Genet. 13(2): 85–96.

Muir, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124(6): 342–355.

Nakaya, A., and S. N. Isobe, 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. (Lond.) 110: 1303–1316.

Okser, S., T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti *et al.*, 2014 Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 12(11): e1004754.

Pérez, P., and G. de los Campos 2014 Genome-wide regression & prediction with the BGLR statistical package. Genetics 198: 483–495.

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, and J. Crossa, Y. Manès, *et al.* 2012 Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2(12): 1595–1605.

Poland, J., 2015 Breeding-assisted genomics. Curr. Opin. Plant Biol. 24: 119–124.

Rosa, G. J., B. D. Valente, G. de los Campos, and X. L. Wu, X. L., D. Gianola *et al.*, 2011 Inferring causal phenotype networks using structural equation models. Genet. Sel. Evol. 43: 6.

Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC methods in quantitative Genetics*. Springer-Verlag, New York.

Stekhoven, D. J., and P. Bühlmann, 2012 MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1): 112–118.

Strandén, I., and O. F. Christensen, 2011 Allele coding in genomic evaluation. Genet. Sel. Evol. 43: 25.

Valente, B. D., G. Morota, F. Peñagaricano, D. Gianola, K. Weigel *et al.*, 2015 The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. Genetics 200(2): 483–494.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–4423.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel, 2011 Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43: 10.

Wimmer, V., C. Lehermeier, T. Albrecht, H. J. Auinger, Y. Wang *et al.*, 2013 Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195(2): 573–587.

Xavier, A., S. Xu, W. M. Muir, and K. M. Rainey, 2015a NAM: Association Studies in Multiple Populations. Bioinformatics 31: 3862–3864.

Xavier, A., W. Beavis, J. Specht, B. Diers, R. Howard *et al.*, 2015b SoyNAM: Soybean Nested Association Mapping Dataset. CRAN, R package version 1.2. Available at: http://cran.mirrorcatalogs.com/web/packages/SoyNAM/index.html. Accessed: June 13, 2016.

Xavier, A., W. M. Muir, and K. M. Rainey, 2016 Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC Bioinformatics 17(1): 1.

Xu, S., 2003 Theoretical basis of the Beavis effect. Genetics 165(4): 2259–2268.

Xu, S., 2013a Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195(4): 1209–1222.

Xu, S., 2013b Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195(3): 1103–1115.

Zhong, S., J. C. Dekkers, R. L. Fernando, and J. L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182(1): 355–364.

Communicating editor: G. A. de los Campos