Genomic selection of seed weight based on low-density SCAR markers in soybean

Y.J. Shu[1]*, D.S. Yu[1]*, D. Wang[1], X. Bai[2], Y.M. Zhu[2] and C.H. Guo[1]

[1]Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin, Heilongjiang Province, China
[2]College of Life Science, Northeast Agricultural University, Harbin, Heilongjiang Province, China

*These authors contributed equally to this study.
Corresponding authors: C.H. Guo / Y.M. Zhu
E-mail: kaku_2008@163.com / ymzhu2001@neau.edu.cn

ABSTRACT. With the development of molecular marker technology, crop breeding has been accelerated by marker-assisted selection for the improvement of quantitative traits. However, due to the traits' polygenic nature, traditional marker-assisted selection methods are ill-suited for identification of quantitative trait loci. Genomic selection (GS) was introduced into crop breeding to achieve more accurate predictions by considering all genes or markers simultaneously. We used dozens of sequence-characterized amplified region (SCAR) markers for genotyping soybean varieties, and we identified markers associated with hundred-seed weight. The best linear unbiased predictor and Bayesian liner regression methods were used to construct GS models to predict the hundred-seed weight trait based upon genotype information for trait selection. Both GS models showed good prediction performance in soybean, as the correlation coefficient between genomic estimated breeding values and true breeding values was as high as 0.904. This indicated that GS was performed effectively based on dozens of SCAR markers in soybean; these markers were of low density but easily detectable. Therefore, the

combination of GS modeling and highly effective molecular marker technology involving SCAR markers can facilitate genetic breeding in soybean. This approach may also be suitable for genetic selection in other crops, such as wheat, maize, and rice.

Key words: Genomic selection; SCAR markers; Hundred-seed weight; Soybean

INTRODUCTION

Over the past 2 decades, molecular marker technology has advanced rapidly and been successfully applied to the genetic breeding domain of animals and crops (Varshney et al., 2005). Molecular markers allow for the detection of stable genetic elements in individuals, and are more reliable than phenotype traits, which are variable under different environments. Therefore, molecular markers have gained popularity among genetic breeders aiming to improve specific desirable traits of animals or plants. This approach is known as marker-assisted selection (MAS), and involves screening of major genes, or genetic loci known widely as quantitative trait loci (QTL), that control development of some traits. Breeders screen groups of individuals with the trait of interest to identify the related genetic loci, which can then be used for genetic improvement by backcrossing the associated genomic fragments into target individuals. Meanwhile, the QTL information is analyzed to identify the precise genes or other genetic elements that determine the trait's development. This method is known as map-based cloning.

MAS is currently used in genetic breeding of many animal and plant species, especially for crop genetic improvement, from mapping QTL by making bi-parental populations to cloning and applying the genes. However, 3 flaws that are inherent to the MAS approach have limited its widespread adoption. First, the bi-parental mapping populations used in most QTL studies, for which representative genetic elements are inadequate, do not readily translate to breeding applications. Second, the QTL analysis is based on many assumptions, including no interaction between genotype and environment, no epistatic effect, and no pleiotropism. It assumes that the genes or genetic elements controlling the traits are unbiased, which is likely not the case since it is known that complex genetic-regulatory networks exist between most of the genetic elements and traits studied to date. Third, QTL analysis is used to determine the major effective loci of a quantitative trait, which often identifies a genomic region that contains many genes that may act in concert or contribute to a trait to a different extent. However, many traits of crops are not determined by single genes mediating large effects. Instead, they are controlled by several small-effect genes. When QTL are used to estimate these effects, the results will be biased, and the small-effect QTL will be missed entirely as a result of using stringent significance thresholds (Gross and Olsen, 2010).

The development of molecular detection technology in recent years has yielded significantly cost-effective methods for detecting genotypes. Due to the large amount of molecular markers identified to date, geneticists have developed association mapping (AM) for use in natural populations. AM has been proposed to mitigate the lack of relevance of bi-parental populations in QTL identification; therefore, many QTL of crops have been mapped in this way (Zhu et al., 2008). For example, the use of AM in wheat, maize, and soybean has greatly promoted crop genetic breeding. Nevertheless, AM retains the disadvantage of biased effect

estimates, making it a poor predictor of line performance. AM is also better suited for manipulating QTL controlled by a few large-effect genes than those controlled by many small-effect genes. Unfortunately, complex polygenic traits are always determined by complex combinations of small-effect genes, which are crucial for the formation of new crop varieties. Therefore, AM is not well suited for many alleles of small-effect segregates, and substantial, reliable effects determined by many small-effect genes cannot be identified.

To minimize the limitations of traditional MAS, a new model has been proposed for genetic breeding, known as genomic selection (GS) (Heffner et al., 2009; Jannink et al., 2010). GS is a form of MAS that simultaneously estimates all marker effects (or genetic loci) across the whole genome to calculate genomic estimated breeding values (GEBVs) (Meuwissen et al., 2001). The GS does not define or identify subsets of significant markers, which could be used for MAS in QTL analysis or AM. Instead, it analyzes all markers of a population in total, to explain the total genetic variance with dense genome-wide markers, and sums all markers' effects to predict the breeding value of individuals. GS has been successfully applied in genetic breeding of both animals and plants, including bovine (e.g., Meuwissen et al., 2001; de Roos et al., 2007; Hayes et al., 2010), mouse (Legarra et al., 2008), wheat (e.g., Crossa et al., 2007; Pérez et al., 2010; Ober et al., 2011), and maize (Messina et al., 2011). In addition, many statistical models have been developed to analyze the different types of genetic and trait data from GS, such as best linear unbiased predictor (BLUP), stepwise regression, ridge regression (RR), and Bayesian estimation (Bayes A or B) (e.g., Messina et al., 2001; Lee et al., 2008; de los Campos et al., 2009; Luan et al., 2009; Hayes et al., 2010; Pérez et al., 2010; Crossa et al., 2010; Macciotta et al., 2010; Schulz-Streeck and Piepho, 2010; Zhang et al., 2010; Ober et al., 2011). These statistical models are used for training the GS models with a training population, and predicting which GEBVs will have significantly improved polygenic traits controlled by many loci of small effect (e.g., Solberg et al., 2008; Rafalski, 2010).

Soybean, one of the most important oil and vegetable protein sources for animals and humans, is an ancient tetraploid with a complicated genome and large numbers of repeated regions. Therefore, efforts in genetic breeding of soybean are complicated and advancements have been slow. Genomic duplications have resulted in many small-effect genes, which hinders QTL identification. In addition, since parental genomic features limit QTL identification by the population mapping method, its use is also restricted in soybean.

Therefore, in this study, sequence-characterized amplified region (SCAR) markers were selected for genotyping soybean varieties. By combining these data with hundred-seed weight (HSW) phenotype data, association analysis was performed to screen large-effect markers for MAS in soybean. In addition, the genotype and phenotype data from the various soybean varieties were used as a training population to construct genomic selection models by RR-BLUP and Bayesian linear regression (BLR), and the accuracy of GEBV predictions in GS was estimated.

MATERIAL AND METHODS

Plant materials

A total of 288 soybean varieties were used for association mapping and genomic selection, all of which were obtained from the Chinese Crop Germplasm Information System

(CGRIS; http://icgr.caas.net.cn/cgris_english.html), and phenotypic data about HSW of these soybean varieties were also collected from CGRIS websites (seen in Table S1). DNA was extracted from fresh leaf samples of all soybean varieties by the cetyltrimethylammonium bromide procedure (Murray and Thompson, 1980), and the products were dissolved in 50 µL distilled water. Quality of DNA was checked by resolution on 0.8% agarose gels, and DNA concentration was determined by spectrophotometric measurements (UV-2550; Shimadzu).

Genotyping soybean varieties by SCAR markers

A total of 79 SCAR markers from across the whole genome were selected according to previous research (e.g., Shu et al., 2010; Shu et al., 2011; the accession numbers of these markers are listed in Table S2) for use in our study. These markers were amplified for genotyping of the soybean varieties. Amplifications were performed in 25-µL reaction mixtures containing 30 ng template DNA, 0.2 µM each primer, 800 µM dNTPs, 1.5 mM $MgCl_2$, 1 U *rTaq* polymerase, and 2.5 µL 10X PCR buffer. All primer pairs were first tested with the touchdown PCR program: 10 min at 94°C; 30 cycles of 30 s at 94°C, 30 s at 60°C minus 0.5°C per cycle, and 40 s at 72°C; and 10 min at 72°C for a final extension. PCR products were resolved on a 2% agarose gel stained with ethidium bromide, and dominant and recessive homozygous results were recorded as "zero" and "one", respectively.

Association mapping in soybean

The genome-wide association mapping was performed using the mixed model described by Kang et al. (2010):

$$y = u + \beta x + \varepsilon \qquad \text{(Equation 1)}$$

where y is the phenotypic outcome (the HSW); the variable u models the genetic background of each variety as a random effect; β is a vector of fixed effects that can model both environmental factors and population structure; and ε is a model residual. The parameter β was estimated using the generalized least-squares approach, as described by Kang et al. (2010).

Statistical models

In parametric models for GS, the linear model was expressed in matrix notion as follows:

$$y = X\beta + \varepsilon \qquad \text{(Equation 2)}$$

where $y = \{y_i\}$, which is the phenotypic outcome (HSW); $\beta = \{\beta_i\}$ is a marker covariate matrix; and $X = \{x_{ij}\}$ is a weight matrix regressed on marker covariates, while $\varepsilon = \{\varepsilon_i\}$ is a model residual.

Since many parameters needed to be determined, and few samples had been collected, the estimation of marker effects via multiple regression by ordinary least squares was not feasible. Instead, penalized estimation methods, such as RR and the Bayesian Lasso (BL), were used to estimate marker effects. In the RR model, the above optimization problem was simplified as suggested by Meuwissen et al. (2001):

$$[X^{'}X + \lambda I]\hat{\beta} = X^{'}y \text{ , equivalently, } \hat{\beta} = [X^{'}X + \lambda I]^{-1}X^{'}y \qquad \text{(Equation 3)}$$

where RR adds a constant, λ, to the diagonal of the matrix of coefficients, making the solution unique and shrinking the estimate of marker effects towards zero, with the extent of shrinkage increasing as λ increased.

Since RR had reduced all marker effects equally, which was presumably an incorrect assumption that could lead to over-shrinking of large effects, Bayesian methods were adopted to relax this assumption and better model marker effects of differing sizes. The combination of the RR model with the BL, as described by Crossa et al. (2010) and Pérez et al. (2010), addressed this issue of the GS model with the following formula.

$$y = 1 * \mu + X_F \beta_F + X_R \beta_R + X_L \beta_L + \varepsilon \qquad \text{(Equation 4)}$$

where y is a vector of response values and μ is an intercept; and X_F, X_R, and X_L are incidence matrices for the vectors of effects β_F, β_R, and β_L, respectively. The vector of regression coefficients β_F, β_R, and β_L were marker effects assigned by "fixed" assignation, RR estimation, and BL, respectively.

Genomic selection in soybean

The phenotype data of these varieties was collected by CGRIS, which provided formal authorization for the genetic analysis described in this study. All the statistical analyses were performed on the open source statistics software, R, with an RR-BLUP and BLR package. Association mapping in soybean was performed using the command "genome-wide association" of RR-BLUP in the software. For GS, the soybean varieties were divided into 2 groups by the Gibbs sampler, one being the training population (N = 238 individuals) and the other being the validation population (N = 50). The genomic selections were trained using the RR-BLUP and BLR methods as previously described (Pérez et al., 2010). The correlation between predicting the GEBVs and the true HSW values in the validation population was calculated for evaluating prediction effects of GS. The GS analysis employed 1000 replicates, and the GS models for RR-BLUP and BLR were compared to evaluate the predictive effects of the two methods.

RESULTS

Genotyping the soybean varieties based on SCAR markers

Seventy-nine SCAR markers were genotyped in 288 soybean varieties. As shown in Figure S1, the longer bands were insertion sites, while the shorter ones were sites without insertion mutation. The 79 SCAR markers yielded more than 22,000 genotype data, shown in Table S1, and polymorphism index analysis showed that SCAR polymorphisms were lower than those of SSR markers, but consistent with single nucleotide polymorphism (SNP) markers. The results indicated that these SCAR markers had other similar characters as SNP markers; for example, both have a bivalent (bi-allelic) character, which makes the SCAR markers well-suited for association mapping and genomic selection. However, as compared to SNP

markers, SCARs are technically easier to detect. The PCR and agarose gel methods used with SCAR are low cost and simple.

AM in soybean varieties

Association analysis of the 79 SCAR markers in soybean varieties identified 6 markers that are associated with HSW of soybean (significance level: 0.05). As shown in Table 1, many of these 6 SCAR markers were associated with functional genes, with roles in soybean seed development; as such, they are expected to represent promising candidates in gene association mapping. The marker SCAR114 target gene, Glyma05g08110, encodes a protein similar to the LEUNIG protein in *Arabidopsis*. Since LEUNIG regulates flower development, the soybean gene Glyma05g08110 may also play an important role in soybean flower development, which ultimately affects HSW. The SCAR356 target gene, Glyma10g07280, encodes the poly(A)-binding protein 7, that regulates gene expression at the post-transcription level, but has yet to be directly correlated with soybean HSW. SCAR356 was located about 7 kb from a gene named Glyma10g07300, which encodes a product similar to the chloroplast outer membrane protein of *Arabidopsis*. Chloroplast outer membrane protein 1 is known to be expressed in plant pollen mother cells, where it functions in sepal differentiation, which would influence the soybean trait HSW.

Table 1. Significance of tests for association between soybean sequence-characterized amplified region (SCAR) markers and hundred-seed weight for the 288 soybean varieties using a mixed model.

Marker	P value	Target gene	Function annotation
SCAR114	0.0374	Glyma05g08110	Similar to LEUNIG of *Arabidopsis*
SCAR182	0.0500	Glyma03g25750	Lysyl tRNA synthetase
SCAR235	0.0366	Glyma01g32680	Putative mitogen-activated protein kinase kinase kinase
SCAR356	0.0417	Glyma10g07280	PolyA-binding protein 7
GmAu1-46	0.0332	No target gene	
GmAu1-474	0.0337	No target gene	

Genomic selection in soybean

The soybean genomic selection method was performed based on RR-BLUP and BLR methods, and the prediction results showed high correlation with true values of HSW in soybean (Figures 1 and 2). In RR-BLUP and BLR genomic selection, the correlation between GEBVs and true breeding values were 0.692 and 0.690, respectively. While the maximum relationship values were 0.854 and 0.904, both of which were higher than previously determined for wheat and maize (e.g., Bernardo and Yu, 2007; Crossa et al., 2010). The results indicated that the HSW trait was controlled by many small-effect genes, which was more suited to genomic selection than the traditional MAS methods. The RR-BLUP prediction standard deviation was relatively small, as were the range of the correlation coefficient of GEBVs and true breeding values. These results indicated that the prediction was stable, suggesting that the differences between sample populations had little effect on the prediction results (Figure 2). In contrast, the BLR prediction standard deviation and correlation coefficient were larger compared to RR-BLUP prediction (e.g., Pérez et al., 2010; Crossa et al., 2010; Schulz-Streeck and Piepho, 2010). The use of a different sampling population would thus be expected to pro-

duce different prediction effects, as suggested by the fluctuating values obtained with different soybean populations (Figure 2). When the training population was used, the prediction was in good agreement with the true breeding values (Figure 1).
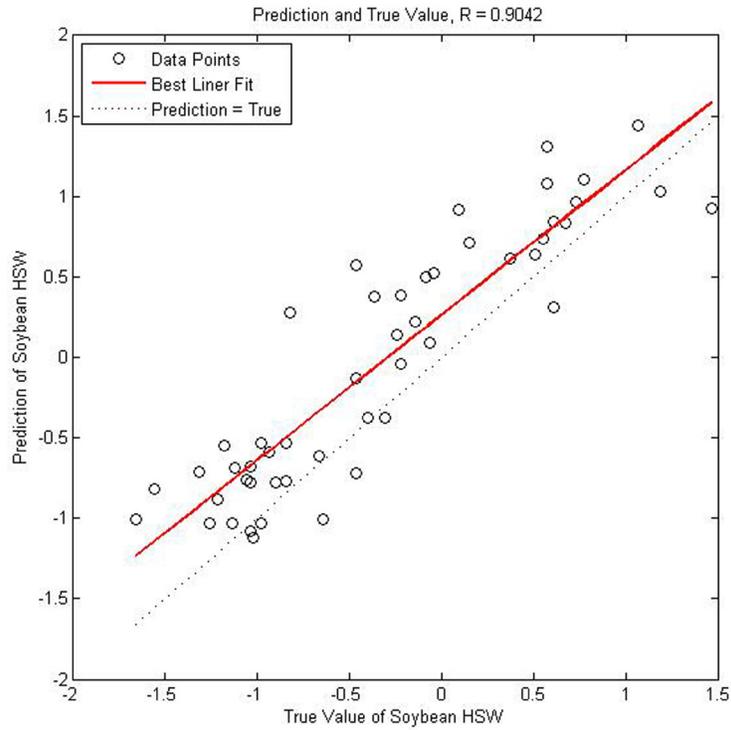

Figure 1. BLR correlation coefficient of genomic estimated breeding values and true breeding values. HSW = hundred seed weight.
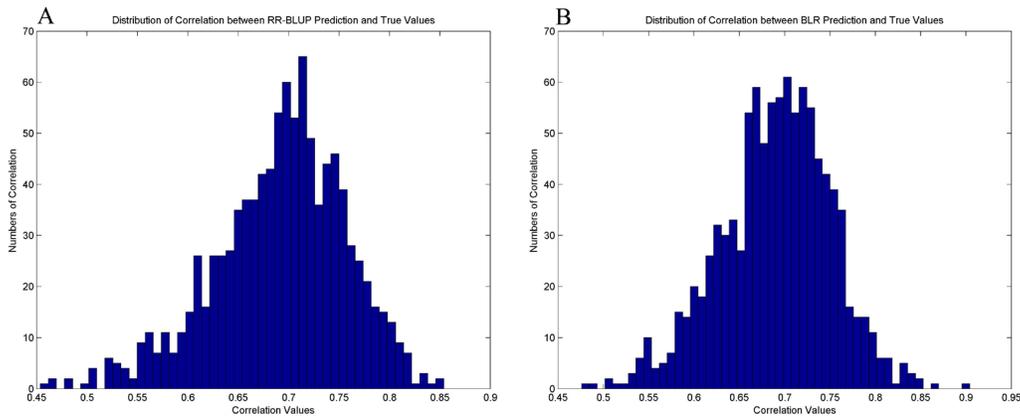

Figure 2. Distribution of correlations between predictions from RR-BLUP and BLR models and true breeding values.

DISCUSSION

Future of genomic selection in genetic breeding

To identify the genes or genetic loci that control agronomic traits of crops, genetic breeders have utilized mapping of QTLs or major genes (Varshney et al., 2005). Advancements in high-throughput genotyping molecular technologies have allowed geneticists to develop the MAS method of association mapping (Zhu et al., 2008), which allows for screening QTL in a natural population and has substantially reduced the time and labor required to map a population. However, both the QTL mapping and association mapping MAS methods are better suited to few large-effect genes, while many traits of animals and crops are determined by several small-effect genes; this feature limits the methods' ability to identify genes controlling traits in animals and crops.

To minimize the limitations of QTL mapping and AM, Meuwissen et al. (2001) proposed a genomic selection method for MAS. The GS method does not rely on only the most significant markers for MAS, but considers all genetic effects of molecular markers across the entire genome to calculate GEBVs. GS uses a "training population" of individuals, which have been both genotyped and phenotyped, to develop a model for predicting GEBVs of a "candidate population" based on their genotypic information. These GEBVs cannot elucidate the mechanism of effective genes, but they are the ideal selection criterion in genetic breeding. In breeding cycles, GS is capable of selecting much earlier traits with genotypic information, and does so at lower cost and reduced labor time, which has revolutionized both animal and plant breeding (e.g., Wong and Bernardo, 2008; Heffner et al., 2009; Calus, 2010).

Advantage of SCAR markers in genetic breeding

SCAR markers facilitate detection of large fragments that have been inserted or deleted in the genome. SCAR features site-specificity, reproducibility, low cost, simplicity, and less time-consuming to results. Therefore, it has been widely used in genetic breeding strategies of plants and animals, such as genetic mapping, map-based cloning, and molecular-assisted breeding. Traditional SCAR markers were developed from the conversion of RAPD or AFLP markers; the initial development process, however, was complex, expensive, slow, and with low throughput, which limited the application and development of SCAR markers (Noguera et al., 2005). We previously developed a multitude of SCAR markers in soybean, which have since greatly decreased the cost of genotyping soybean (e.g., Shu et al., 2010; Shu et al., 2011). Furthermore, since the SCAR marker is bi-allelic, similar to the SNP markers, it is well-suited to any genetic research, including GS. Currently, the SSR and SNP markers are both widely used for GS in animals and crops; however, the SSR markers have many alleles per locus, which has added more potentially confounding parameters to the subsequent statistical analysis and limits its application in the GS model (Solberg et al., 2008). The SNPs are bi-allelic markers and detected by high-throughput genotyping. However, the instrumentation required for genotyping SNPs is extremely costly and the data analysis is very complicated, which has restricted their application in crop genetic breeding. However, SCAR markers, with their advantages of easy detection and bi-allelic nature, are ideal for genetic research in crops. In addition, most of the SCAR markers used in this study are gene-targeting markers, which

make them effective for genomic selection in soybean. The correlation coefficient between GEBVs and true breeding values was high (0.904), especially considering the low density of the SCAR genome-wide. Thus, SCARs are highly effective markers in genomic selection, as compared to SSRs or SNPs.

Application of GS in soybean genetic breeding

In soybean genetic breeding, many QTL have been identified by population mapping. Beneficial genes, such as those conferring abiotic- or disease-resistance and linked to QTL, have been mapped, identified, and manipulated to produce more robust soybean crops. However, QTL mapping is population-dependent, and the selection effect decreases in populations with different genomic structures. To identify more general molecular markers for MAS, association mapping in a natural population was developed for screening molecular markers, which are related to desired traits of crops. QTL for important traits in soybean have been identified by AM using SSR markers, including those related to protein content and iron deficiency (e.g., Wang et al., 2008; Jun et al., 2008). Importantly, however, AM cannot resolve problems of small-effect genes and multi-gene interaction, which has limited the application of AM in soybean.

Before genomic selection, the BLUP method should be used to predict yield performance of cross-based agronomic traits in soybean. This approach has previously guided the selection of high-performance parents for crossing to improve the yield trait of soybean plants (e.g., Panter and Allen, 1995a,b). However, the number of useful agronomic traits is small, and the agronomic traits are highly variable in different environments. Therefore, we constructed a genomic selection method based on genotype information for predicting the soybean trait HSW, which is stable in individuals. The results showed that GS was a more precise predictor than the selection-based phenotype data. Since GS in soybean considers the whole genome, it is more suitable for soybean genetic breeding than traditional MAS or selection based on other agronomic traits.

In summary, we used SCAR markers to genotype a population of different soybean varieties. RR-BLUP and BLR were used to perform the GS of HSW. The results showed that low-density SCAR markers sufficiently genotyped soybean individuals, and that the GEBVs from GS were highly correlated with the true breeding values. Therefore, GS would be suitable for estimation of crop breeding traits, not only in soybean, but also possibly in other economically important crops, such as wheat, maize, and rice. Finally, the collective data from our study indicates that GS may resolve the current limitations of the crop genetic breeding process.

ACKNOWLEDGMENTS

Supplementary material

REFERENCES

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

Calus MPL (2010). Genomic breeding value prediction: methods and procedures. *Animal* 4: 157-164.

Crossa J, Burgueño J, Dreisigacker S, Vargas M, et al. (2007). Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177: 1889-1913.

Crossa J, de los Campos G, Perez P, Gianola D, et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186: 713-724.

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

de Roos AP, Schrooten C, Mullaart E, Calus MP, et al. (2007). Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *J. Dairy Sci.* 90: 4821-4829.

Gross BL and Olsen KM (2010). Genetic perspectives on crop domestication. *Trends Plant Sci.* 15: 529-537.

Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, et al. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.* 6: e1001139.

Heffner EL, Sorrells ME and Jannink JL (2009). Genomic selection for crop improvement. *Crop Sci.* 49: 1-12.

Jannink JL, Lorenz AJ and Iwata H (2010). Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genom.* 9: 166-177.

Jun TH, Van K, Kim MY, Lee SH, et al. (2008). Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* 162: 179-191.

Kang HM, Sul JH, Service SK, Zaitlen NA, et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42: 348-354.

Lee SH, van der Werf JH, Hayes BJ, Goddard ME, et al. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* 4: e1000231.

Legarra A, Robert-Granié C, Manfredi E and Elsen JM (2008). Performance of genomic selection in mice. *Genetics* 180: 611-618.

Luan T, Woolliams JA, Lien S, Kent M, et al. (2009). The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183: 1119-1126.

Macciotta NP, Gaspa G, Steri R, Nicolazzi EL, et al. (2010). Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *J. Dairy Sci.* 93: 2765-2774.

Messina CD, Podlich D, Dong Z, Samples M, et al. (2011). Yield-trait performance landscapes: from theory to application in breeding maize for drought tolerance. *J. Exp. Bot.* 62: 855-868.

Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.

Murray MG and Thompson WF (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8: 4321-4325.

Noguera FJ, Capel J, Alvarez JI and Lozano R (2005). Development and mapping of a codominant SCAR marker linked to the andromonoecious gene of melon. *Theor. Appl. Genet.* 110: 714-720.

Ober U, Erbe M, Long N, Porcu E, et al. (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics* 188: 695-708.

Panter DM and Allen FL (1995a). Using best linear unbiased predictions to enhance breeding for yield in soybean: I. choosing parents. *Crop Sci.* 35: 397-405.

Panter DM and Allen FL (1995b). Using best linear unbiased predictions to enhance breeding for yield in soybean: II. selection of superior crosses from a limited number of yield trials. *Crop Sci.* 35: 405-410.

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

Rafalski JA (2010). Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13: 174-180.

Schulz-Streeck T and Piepho HP (2010). Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proc.* 4 (Suppl 1): S8.

Shu Y, Li Y, Zhu Y, Zhu Z, et al. (2010). Genome-wide identification of intron fragment insertion mutations and their potential use as SCAR molecular markers in the soybean. *Theor. Appl. Genet.* 121: 1-8.

Shu Y, Li Y, Bai X, Cai H, et al. (2011). Identification and characterization of a new member of the SINE Au retroposon family (GmAu1) in the soybean, *Glycine max* (L.) Merr., genome and its potential application. *Plant Cell Rep.* 30: 2207-2213.

Solberg TR, Sonesson AK, Woolliams JA and Meuwissen TH (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86: 2447-2454.

Varshney RK, Graner A and Sorrells ME (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10: 621-630.

Wang J, McClean PE, Lee R, Goos RJ, et al. (2008). Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. *Theor. Appl. Genet.* 116: 777-787.

Wong CK and Bernardo R (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.* 116: 815-824.

Zhang Z, Liu J, Ding X, Bijma P, et al. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* 5: e12648.

Zhu C, Gore M, Buckler ES and Yu J (2008). Status and prospects of association mapping in plants. *Plant Genome* 1: 5-20.