Comparison of methods used to identify superior individuals in genomic selection in plant breeding

L.L. Bhering[1], V.S. Junqueira[1], L.A. Peixoto[1], C.D. Cruz[1] and B.G. Laviola[2]

[1]Departamento de Biologia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Embrapa Agroenergia, Parque Estação Biológica, Brasília, DF, Brasil

Corresponding author: L.L. Bhering
E-mail: leonardo.bhering@ufv.br

Genet. Mol. Res. 14 (3): 10888-10896 (2015)
Received March 25, 2015
Accepted July 1, 2015
Published September 9, 2015
DOI http://dx.doi.org/10.4238/2015.September.9.26

ABSTRACT. The aim of this study was to evaluate different methods used in genomic selection, and to verify those that select a higher proportion of individuals with superior genotypes. Thus, F2 populations of different sizes were simulated (100, 200, 500, and 1000 individuals) with 10 replications each. These consisted of 10 linkage groups (LG) of 100 cM each, containing 100 equally spaced markers per linkage group, of which 200 controlled the characteristics, defined as the 20 initials of each LG. Genetic and phenotypic values were simulated assuming binomial distribution of effects for each LG, and the absence of dominance. For phenotypic values, heritabilities of 20, 50, and 80% were considered. To compare methodologies, the analysis processing time, coefficient of coincidence (selection of 5, 10, and 20% of superior individuals), and Spearman correlation between true genetic values, and the genomic values predicted by each methodology were determined. Considering the processing time, the three methodologies were statistically different, rrBLUP was the fastest, and Bayesian LASSO was the slowest. Spearman correlation revealed that the rrBLUP and GBLUP methodologies were equivalent, and Bayesian LASSO provided the lowest correlation values.

Similar results were obtained in coincidence variables among the individuals selected, in which Bayesian LASSO differed statistically and presented a lower value than the other methodologies. Therefore, for the scenarios evaluated, rrBLUP is the best methodology for the selection of genetically superior individuals.

Key words: Statistics; Biometric; SNPs; Selection

INTRODUCTION

One of the most important stages of plant breeding is selection. In order to improve selection accuracy, different methods have been used, such as recurrent selection (Ordas et al., 2012), combined selection (Bhering et al., 2013), and selection based on best linear unbiased prediction (BLUP) (Viana et al., 2011).

However, the biggest breakthrough in selection accuracy during plant breeding came with the advent of molecular markers (Heffner et al., 2009). The first technique used was marker assisted selection (MAS), which used data from molecular markers (Kumar et al., 2011) and microsatellites (Wang et al., 2011). This technique has been successfully used to select for monogenic or oligogenic characteristics; however, their use for quantitative characteristics is limited since this method cannot locate genes of lower effect (Xu and Crouch, 2008).

MAS has two main limitations: the mapping populations used for studies of quantitative trait loci (QTL) are not easily obtained, and the statistical methods used to identify QTLs are unsuitable for characteristics ruled by genes of low effect (Heffner et al., 2009). Thus, Meuwissen et al. (2001) proposed genomic selection methods using high-density markers, with no need for genetic maps. In addition, this uses appropriate statistical methods to identify genes of low effect. Genomic selection has attracted attention since the advent of single nucleotide polymorphism (SNP) markers, because SNPs are abundant in the genome (Gunderson et al., 2005).

The main aim of genomic selection is to estimate the genomic genetic value (GEBV) of the individual that was genotyped only, using data from a mapping population in which individuals were phenotyped and genotyped (Meuwissen et al., 2001). The training population is used to estimate the parameters of the model that will be used to estimate the GEBV of each individual in the validation population. GEBV values will be used to select the best individuals during the selection cycles. Thus, individuals with only genotypic information are selected (Heffner et al., 2009).

After the pioneering study of Meuwissen et al. (2001), several researchers began to evaluate different genomic selection techniques in bovine (Taylor et al., 2012), sheep (Van der Werf, 2009), maize (Bernardo and Yu, 2007), wheat (Poland et al., 2012), and eucalyptus (Resende et al., 2012). Many studies have also been carried out to compare the accuracy of prediction of genomic selection methods its (Muir, 2007; Zhong et al., 2009; Daetwyler et al., 2010). However, most studies have only focused on the following features: comparison between methods; a statistical approach involving each method; effects that will be estimated by the model; if the model will use only one variance for all markers; whether it will use one variance for each marker; or whether the Bayesian method will make the measurements more accurate. However, these lack information that would be more useful when methods for genetic improvement are used, such as the superior individuals selected by each method, and how each method is able to identify superior individuals within the population. Obtaining this information is as important as discussing the method, and should be the focus of applied studies.

Therefore, the aim of this study was to evaluate different methods used in genomic selection, and to verify those that select a higher proportion of individuals with superior genotypes.

MATERIAL AND METHODS

Data simulation

To generate data, the simulation module of the GENES software was used (Cruz, 2013). Samples of 100, 200, 500, and 1000 individuals were generated with 10 linkage groups (LG) each. These population sizes are the most commonly used in breeding programs.

Genome simulation

A genome of 10 LG was simulated, which is similar to a diploid species 2n = 10, with a 100 cM size, considering the existence of 100 molecular markers for the linkage group, equally spaced; thus, a total of 1000 molecular markers were considered.

Parental simulation

Contrasting homozygous parents were simulated for generating the F1 generation; thus, parent 1 (AA) was coded with 1 for all markers, and parent 2 (aa) was coded with 0, for all markers.

Simulation of the mapping population

For each population size, 10 replications were simulated. An F2 population was generated, and for that, each F1 individual produced 5000 gametes. There was random fecundation, which generated F2 individuals. This process was repeated until all individuals were formed and all replications were carried out. This population was coded 0, 1, and 2, with 0 corresponding to the homozygous recessive individuals (aa) for the locus, 1 for the heterozygous individual (Aa), and 2 for the dominant homozygous individuals (AA) for the considered locus.

Simulation of genomic selection population

Genotypic and phenotypic variables were simulated. The genotypic variables contain the actual genetic value of each individual. The phenotypic value was simulated considering three different heritabilities: 20, 50, and 80%. Of the 1000 markers previously simulated, 200 controlled the characteristics, and the first 20 molecular markers in each LG were taken into account. Therefore, at 10 LG, there were 200 loci.

The binomial distribution of effects for each characteristic in each LG was used. The additive gene action of all loci was used, i.e., the dominance effect was considered null. To establish the phenotypic and genotypic values, these were added up to a constant equal of 100, thus preventing that any of the individuals for each of these variables presenting negative values for any variable.

Data analysis

After data were obtained, three methods of analysis were used: rrBLUP, GBLUP, and

Bayesian LASSO. We ran the Bayesian LASSO using Monte Karlo Markov Chain method (MCMC) with 10,000 iterations, burn-in of 1000 iterations, and thin of 20 iterations. To perform data analysis in each method, R package (R Development Core Team, 2012). BLR (Pérez et al., 2010), rrBLUP (Endelman, 2011), and MASS were used. We used a DELL 12th generation server, Intel Xeon E5-26 processor 3.30 GHz, 64 GB RAM, and a 1024 GB hard drive.

Evaluation of populations analyzed

The mapping process was performed after data were generated, starting by the segregation of individual loci. Chi-square tests were used (χ^2), at 5% probability, to verify the result of segregation for each marker in the populations generated. In addition, the restoration of all LGs was verified, with size, distance, and marker order, to conclude whether the populations simulated were F2 and contained the desired simulation properties.

Analyses comparison

To compare the analyses, variance analysis was carried out, followed by the Tukey test at 5% probability. The variables used were: processing time (in seconds), Spearman correlation, and the percentage of coincidence at 5, 10, and 20% of superior individuals of the simulated genetic values, and genetic genomic values were generated after the use of each method.

RESULTS

All the tested scenarios contained 1000 markers, the size of the population and heritability were varied, and subsequent analyses were performed by different methods. To better visualize the simulated populations, Figures 1 and 2 show the dispersion of markers in the most divergent populations (sizes: 100 and 1000; and heritability: 20 and 80%).

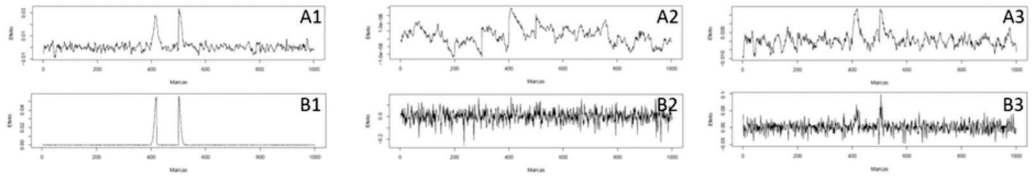

Figure 1. Effect of each marker in populations of 100 (**A**) and 1000 (**B**) individuals; Variables evaluated: genotypic value (**A1** and **B1**), phenotypic value with 20% heritability (**A2** and **B2**), and phenotypic value with 80% heritability (**A3** and **B3**).

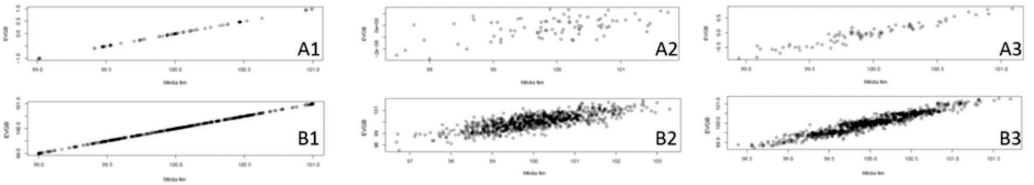

Figure 2. Scatter plots of predicted genotypic value for populations of 100 (**A**) and 1000 (**B**) individuals. Variables evaluated: genotypic value (**A1** and **B1**), phenotypic value with 20% heritability (**A2** and **B2**), and phenotypic value with 80% heritability (**A3** and **B3**).

Figure 1 shows the dispersion of all markers and their effects, considering the variable constituted only by the simulated genetic value (A1 and B1), and A refers to the population of 100 individuals, and B to the population of 1000 individuals. The effect of heritability in the simulated phenotypic variables is noted, and therefore A2 and A3 are the populations with 100 individuals with 20 and 80% heritability, respectively. As heritability increases (A3 and B3), the images become closer to that simulated with the genetic value only (A1 and B1).

Figure 2 shows the dispersion of predicted genotypic and phenotypic values. The observed and predicted genotypic values are the same and form a perfect line, with no outlying points; thus, the correlation between the predicted value is 1 (A1 and B1). For phenotypic values, as the heritability increases, the environmental variation decreases, and therefore, the predicted value becomes increasingly close to the actual value, tending to form a line (A3 and B3). Figure 2 shows that increasing the population interferes with the continuity of the sampled points.

Genetic maps were generated for all populations in order to evaluate the quality of simulated data. Starting with the segregation analysis of individual loci, chi-square tests (χ^2) were applied to check the segregation ratio of all populations generated. There was no segregation distortion, i.e., all markers typically segregate as a codominant F2 (1:2:1). All linkage groups were restored according to the parameters used, both in terms of the total size (100 cM), the main distance between markers, and the order of the markers that constitute the linkage group; thus, it was concluded that the simulated population had characteristics of an F2 population, and would be appropriate for use in this study. Therefore, it was important to carry out the aforementioned analyses to ensure that an F2 population was being used, and from this, it is possible to infer the genetic parameters set by Falconer and Mackay (1996) for this type of population, such as genetic, phenotypic, and environmental variance, all of which are estimated by RRBLUP, GBLUP, and Bayesian LASSO methods, and then used for to calculate heritability.

From the heritability, it was possible to estimate the selection gain, select the best individuals, and then calculate the Spearman correlation and the coincidence. Analysis of processing time, Spearman correlation, and coincidence between the individuals selected based on phenotypic value and genotipic breeding value for different percentege of the individuals selected (5, 10, and 20%) was carried out using the analysis of variance for balanced data, followed by a 5% probability test. The analysis time differed significantly according to the population size and analysis method used, with rrBLUP having the shortest processing time, and Bayesian LASSO having the longest processing time (Table 1).

Table 1. Time taken for each analysis in seconds, considering populations of different sizes (100, 200, 500, and 1000 individuals) and different methods [Bayesian LASSO (BLASSO), GBLUP, and RRBLUP].

Methods	Population size			
	100	200	500	1000
BLASSO	39.57[a]	43.51[a]	57.59[a]	70.08[a]
GBLUP	4.33[b]	4.55[b]	5.95[b]	12.22[b]
RRBLUP	0.29[c]	0.51[c]	1.67[c]	6.55[c]

*Means followed by different letters differ statistically at 5% probability by the Tukey test.

Spearman correlation is a nonparametric correlation that measures the variation of the materials ranking. It is extremely important that it is as high as possible, since this would allow the methods to select the genetically superior individuals, and to maintain their order. For this variable,

the methods showed significant differences at heritabilities of 20 and 50%, and Bayesian LASSO presented the worst Spearman correlation (Table 2). At a heritability of 80%, the three methods were equivalent; however, numerically, Bayesian LASSO had the worst values (Table 2).

Table 2. Spearman correlation between breeding values and genomic values, considering different heritability (20, 50, and 80%) and methods [Bayesian LASSO (BLASSO), GBLUP, RRBLUP].

Methods	Herytability		
	20%	50%	80%
BLASSO	0.59[b]	0.78[b]	0.92[a]
GBLUP	0.80[a]	0.87[a]	0.95[a]
RRBLUP	0.80[a]	0.90[a]	0.95[a]

Means followed by different letters differ statistically at 5% probability by the Tukey test.

Considering the different population sizes, the results of the Spearman correlation did not differ statistically between the methods. However, Bayesian LASSO continued to provide the lowest values for this variable, which is not desirable for breeding, while the RRBLUP method presented higher values for all population sizes evaluated (Table 3).

Table 3. Spearman correlation between genetic and genomic values, considering different population sizes (100, 200, 500, and 1000 individuals) and methods [Bayesian LASSO (BLASSO), GBLUP, RRBLUP].

Methods	Population size			
	100	200	500	1000
BLASSO	0.67[a]	0.71[a]	0.80[a]	0.82[a]
GBLUP	0.79[a]	0.86[a]	0.88[a]	0.94[a]
RRBLUP	0.79[a]	0.86[a]	0.91[a]	0.94[a]

Means followed by different letters differ statistically at 5% probability by the Tukey test.

The percent coincidence (5, 10, and 20%) of superior individuals is important for breeding since, if selection of phenotypically superior individuals is carried out, it is desirable that they are genetically superior. Therefore, this provides key information of which method would be more useful for selection in breeding programs. For 200, 500, and 1000 individuals for a selection intensity of 5% of the superior individuals, Bayesian LASSO differed statistically from the other methods, and provided the lowest coincidence value between genetic and genomic genetic values predicted. For the other combinations of population size and percent coincidence, there were no statistical differences between the methods; however, Bayesian LASSO provided numerically lower values for this variable under all situations evaluated (Table 4).

On the other hand, when we compared the methods for each percentage of superior individuals selected and heritability was observed that Bayesian LASSO was statistically lower than RRBLUP and GBLUP, except when the heritability was 80% and the selection of the 5% superior individuals where do not have statistic difference between methods (Table 5).

Table 4. Percent coincidence between genetic values and genomic values, considering the percentage of superior individuals (5, 10, and 20%), different population sizes (100, 200, 500, and 1000 individuals), and methods [Bayesian LASSO (BLASSO), GBLUP, RRBLUP].

	Population size/coincidence %											
	100			200			500			1000		
	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%
BLASSO	47.5[a]	46.66[a]	57.91[a]	45.0[b]	53.54[a]	60.10[a]	58.3[b]	58.75[a]	67.25[a]	53.8[b]	58.79[a]	68.39[a]
GBLUP	55.0[a]	61.25[a]	67.50[a]	64.58[a]	69.58[a]	71.14[a]	76.3[a]	71.75[a]	78.83[a]	80.6[a]	75.79[a]	79.43[a]
RRBLUP	56.6[a]	61.25[a]	67.50[a]	64.5[a]	69.58[a]	71.14[a]	76.3[a]	71.75[a]	78.87[a]	80.6[a]	75.62[a]	79.39[a]

Means followed by different letters statistically differ at 5% probability by the Tukey test.

Table 5. Percent coincidence between genetic and genomic values, considering the percentage of superior individuals (5, 10, and 20%), different heritabilities (20, 50, and 80%) and methods [Bayesian LASSO (BLASSO), GBLUP, RRBLUP].

	Populations size/coincidence %								
	20%			50%			80%		
	5%	10%	20%	5%	10%	20%	5%	10%	20%
LASSO	29.41[b]	36.87[b]	49.77[b]	50.66[b]	55.12[b]	66.02[b]	79.58[a]	72.20[b]	77.77[b]
GBLUP	54.50[a]	57.70[a]	65.81[a]	71.25[a]	71.29[a]	77.43[a]	86.25[a]	79.79[a]	82.52[a]
RRBLUP	54.50[a]	57.50[a]	65.81[a]	72.91[a]	71.37[a]	77.39[a]	86.25[a]	79.75[a]	82.56[a]

Means followed by different letters statistically differ at 5% probability by the Tukey test.

DISCUSSION

The analysis time differed significantly according to the population size and analysis method used, with rrBLUP having the shortest processing time, and Bayesian LASSO having the longest processing time (Table 1). This result was expected since the processing time of Bayesian methods, such as Bayesian LASSO, is related to the size of the Monte Carlo chain (MCMC) through the simulated Markov Chain (Meuwissen et al., 2001). Considering the processing time variable, the three methods did not differ between different heritabilities; thus, changing the heritability value did not change the efficiency of computer processing, and for this reason, the results are not shown. Pérez et al. (2010) showed that the Bayesian LASSO processing time was 11 s for 1000 interactions, and the data file consists of 599 individuals and 1279 markers, and twice the time was spent on analyses using the Bayesian Ridge Regression method (BRR). The longer time spent by Bayesian LASSO is justified since the Gibbs sampler, used in the Bayesian method of this study, belongs to the MCMC class of methods. This method is used to obtain the marginal distribution a posteriori of the model parameters; however, it requires more computational time. On the other hand, methods based on frequent statistics, such as RRBLUP and GBLUP, require less computational time since they do not need interaction to converge.

The accuracy with which the genomic genetic values are predicted is limited by two main factors: the linkage disequilibrium between the markers and the QTL, which may be incomplete, and thus the marker does not explain all the variance of the QTL; and sampling error in the estimation of marker effects. These errors increase with environmental variance (Meuwissen et al., 2001). The first factor explains why no Spearman correlation was observed close to 100%. Since the number of markers was small, these markers could not be in linkage disequilibrium with the QTL. The

second limiting factor explains why Spearman correlation was lower at 20% heritability, since at this magnitude, environmental variance is greater than for the other heritabilities analyzed in this study.

The fact that the Spearman correlation calculated in Bayesian LASSO was lower than that calculated by other methods may be due to two reasons: the small number of markers and thus marker density used, and the fact that not all markers may have been in linkage disequilibrium with the QTL.

This affects Bayesian LASSO more than the other methods because, according to Rolf et al. (2010), even with the low number of markers, it is possible to estimate the G matrix used in GBLUP precisely, meaning that this method estimates the genomic genetic value with high accuracy, even when using small number of SNP markers. Another factor may be the use of a more informative priori. Once there was no previous knowledge of the hyperparameters used by Bayesian LASSO (λ), flat priori values were used. When using a flat priori, the estimates of genetic values are based only on the likelihood function, making the results of Bayesian methods (Bayesian LASSO) close to the results found by frequentist methods (RRBLUP and GBLUP) (De Los Campos et al., 2009).

According to Meuwissen et al. (2001), increasing the number of individuals analyzed, i.e., individuals genotyped and phenotyped, consequently increases the accuracy of the genomic selection method, and this occurs even when the number of individuals evaluated is already high. This was observed in the present study where the population of 1000 individuals had higher coincidence values when compared with the populations containing 100, 200, and 500 individuals.

Simulation studies may help breeders in decision-making, once it is possible to define the real genetic value of individuals. Therefore, this study showed that despite the methodological and statistical point of view being more interesting, Bayesian LASSO may not be the best method recommended for selection, while faster methods, such as RRBLUP, may be able to assist in selection carried out in plant breeding programs, allowing the breeder to select the most superior individuals for recombination or further steps of a breeding program.

In conclusion, Bayesian LASSO had the lowest Spearman correlation values and percentage of coincidence in the selection of superior individuals; it presented even greater computational requirements for processing, and was therefore considered to be the worst method for selection in this study.

RRBLUP and GBLUP obtained similar Spearman correlation results and percentage of coincidence, and both are suitable for use in selection.

Finally, RRBLUP had lower computational requirements for processing, and required less time for analysis; thus, this is the method that presented more satisfactory results in all of the variables.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Acknowledgments to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, which funded the scholarship, and FAPEMIG, CAPES and CNPq, Brazil, to financial support.

REFERENCES

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

Bhering LL, Barrera CF, Ortega D, Laviola BG, et al. (2013). Differential response of *Jatropha* genotypes to different selection methods indicates that combined selection is more suited than other methods for rapid improvement of the species. *Ind. Crops Prod.* 41: 260-265.

Cruz CD (2013). Genes: a software package for analysis in experimental statistics and quantitative genetics. *Acta Sci. Agron.* 35: 271-276.

Daetwyler HD, Pong-Wong R, Villanueva B and Woolliams JA (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185: 1021-1031.

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.

Endelman JB (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4: 250-255.

Falconer DS and Mackay TFC (1996). Introduction to Quantitative Genetics (4th edn). Harlow, Longman.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, et al. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37: 549-554.

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

Kumar J, Choudhary AK, Solanki RK and Pratap A (2011). Towards marker-assisted selection in pulses: a review. *Plant Breed.* 130: 297-313.

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

Muir W (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: 342-355.

Ordas B, Butrón A, Alvarez A, Revilla P, et al. (2012). Comparison of two methods of reciprocal recurrent selection in maize (*Zea mays* L.). *Theor. Appl. Genet.* 124: 1183-1191.

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.

Poland J, Endelman J, Dawson J, Rutkoski J, et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5: 103-113.

Resende MFR, Munoz P, Acosta JJ, Peter GF, et al. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol.* 193: 617-624.

R Development Core Team (2012). R: a language and environment for statistical computing. Vienna, Austria.

Rolf MM, Taylor JF, Schnabel RD, McKay SD, et al. (2010). Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genetics* 11: 24.

Taylor JF, McKay SD, Rolf MM, Ramey HR, et al. (2012). Genomic Selection in Beef Cattle. In: Bovine Genomics (Womack JE, eds). John Wiley & Sons, 211-233.

Van der Werf J (2009). Potential benefit of genomic selection in sheep. In: Proceedings of the Association for the Advancement of Animal Breeding and Genetics, 38-41.

Viana JMS, Faria VR, Silva FS and Resende MDV (2011). Best linear unbiased prediction and family selection in crop species. *Crop Sci.* 51: 2371-2381.

Wang C, Chen S and Yu S (2011). Functional markers developed from multiple loci in GS3 for fine marker-assisted selection of grain length in rice. *Theor. Appl. Genet.* 122: 905-913.

Xu Y and Crouch JH (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.* 48: 391-407.

Zhong S, Dekkers JC, Fernando RL and Jannink JL (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182: 355-364.