RESEARCH ARTICLE

Evaluation of random forest regression for prediction of breeding value from genomewide SNPs

RUPAM KUMAR SARKAR[1], A. R. RAO[1]*, PRABINA KUMAR MEHER[1], T. NEPOLEAN[2] and T. MOHAPATRA[3]

[1]*ICAR-Indian Agricultural Statistics Research Institute, and* [2]*ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India*
[3]*ICAR-Central Rice Research Institute, Cuttack 753 006, India*

Abstract

Genomic prediction is meant for estimating the breeding value using molecular marker data which has turned out to be a powerful tool for efficient utilization of germplasm resources and rapid improvement of cultivars. Model-based techniques have been widely used for prediction of breeding values of genotypes from genomewide association studies. However, application of the random forest (RF), a model-free ensemble learning method, is not widely used for prediction. In this study, the optimum values of tuning parameters of RF have been identified and applied to predict the breeding value of genotypes based on genomewide single-nucleotide polymorphisms (SNPs), where the number of SNPs (P variables) is much higher than the number of genotypes (n observations) ($P >> n$). Further, a comparison was made with the model-based genomic prediction methods, namely, least absolute shrinkage and selection operator (LASSO), ridge regression (RR) and elastic net (EN) under $P >> n$. It was found that the correlations between the predicted and observed trait response were 0.591, 0.539, 0.431 and 0.587 for RF, LASSO, RR and EN, respectively, which implies superiority of the RF over the model-based techniques in genomic prediction. Hence, we suggest that the RF methodology can be used as an alternative to the model-based techniques for the prediction of breeding value at genome level with higher accuracy.

[Sarkar R. K., Rao A. R., Meher P. K., Nepolean T. and Mohapatra T. 2015 Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. *J. Genet.* **94**, 187–192]

Introduction

The Mendelian rules of genetic inheritance are most obvious, when the traits are controlled by a single gene. However, most of the agronomically important traits are complex in nature and are influenced by many genes with small effects, by nongenetic or environmental factors (Hill 2010). With the dramatic drop in the cost of generating DNA marker information, it is expected to get immediate benefits, such as, rapid delivery of varieties/breeds with improved yield, quality and tolerance to biotic and abiotic stresses. But due to complex nature of these traits, even the traditional marker-assisted selection has proved to be only partially effective (Jannink *et al*. 2010).

SNP-based markers have rapidly gained importance in molecular genetics due to their amenability for high-throughput detection (Mammadov *et al.* 2012) and dense coverage in the genome. The introduction of genomic selection (GS) method, where genomewide dense markers/SNPs are used to estimate the breeding value of the individuals, has shifted the paradigm. The GS has been successfully used in dairy cattle breeding (Lillehammer *et al.* 2011). Also, the accuracies obtained from GS method are sufficient to generate rapid gains in early selection cycles (Jannink *et al*. 2010).

GS typically consists of two steps: (i) estimation of the marker effects in a dataset with marker genotyped and phenotyped individuals (called the training set); and (ii) use the estimated marker effects to predict the breeding value of individuals that are not phenotyped (called the test set). In GS, since the number of individuals (n) genotyped is lesser than the number of markers/SNPs (p), the multiple linear regression models have failed due to the problem of over fitting. In the past, several statistical techniques such as parametric (Bayesian least absolute shrinkage and selection operator, best linear unbiased predictor, Bayesian ridge regression, ridge regression, elastic net), nonparametric (support vector machine, neural network) and semiparametric (reproducing kernel Hilbert space) have been used for the

*For correspondence. E-mail: arrao@iasri.res.in; rao.cshl.work@gmail.com.

Keywords. genomewide SNPs; penalized regression; prediction of breeding value; machine learning methods.

estimation of SNP effects and genomic prediction (Goddard *et al*. 2009; Crossa *et al.* 2010; Heslot *et al*. 2012).

The random forest (RF) (Breiman 2001) is one of the most successful machine learning methods that has been used widely and successfully in the classification of binary response, based on genomewide SNP data (Strobl *et al.* 2009). However, application of RF regression for genomic prediction in plants is yet to be fully explored. Moreover, predictive performance of the model-based methods as compared to RF regression in the event of SNPs is higher in number as compared to the genotypes, i.e., $p >> n$ need to be assessed. In this study, we assessed the predictive ability of RF regression method for prediction of breeding value of kernel length of maize germplasm by using a trained RF model with optimal values of parameters. Hence, a comparison was made between the RF regression and commonly used model-based techniques, namely, LASSO (Tibshirani 1996), elastic net (EN) (Zou and Hastie 2005) and ridge regression (RR) (Hoerl and Kennard 1968).

Materials and methods

Dataset

The whole genome SNP database of 269 maize accessions, spanning over 10 chromosomes, SNPs in each chromosome being 7914, 5913, 5744, 5677, 5611, 4187, 4199, 4387, 3741 and 3704, respectively were used as the predictor variables. SNPs were coded as 0 and 2 for the dominant and recessive homozygote, respectively and 1 for the heterozygote. Continuous response variable used in this study was maize kernel length. The SNP dataset was collected from the open source maize database available at http://panzea.org/db/gateway?file_id=Cook_etal_2012_MaizeSNP50_genos_282panel (Cook *et al.* 2012). The phenotypic data were collected through the phenotypic search option available at http://panzea.org/db/searches/webform/phenotype_search.

Penalized regression

For the regression model $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i$; $i = 1, 2, 3, ..., n$, the estimates of βs' using RR are obtained by minimizing the L_2 panelized residual sum of square,

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s,$$

where s (≥ 0) is the tuning parameter that controls the model complexity. Choice of an appropriate value of s is important to get an appropriate estimate. RR performs well when the subsets of true coefficients are small or zero, but it does not perform well when all the true coefficients are moderately large. However, it can still outperform linear regression over a pretty narrow range of small s values.

For the same model, LASSO estimate of βs' are obtained by minimizing residual sum of square subject to the constraint of L_1-type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool. In other words, the LASSO estimate is solution to the constrained OLS minimization problem,

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t,$$

where $t \geq 0$ is the tuning parameter. The tuning parameter $t \geq 0$ controls the amount of shrinkage that is applied to the estimates.

EN is a more generalized model that combines both the RR and LASSO penalty. In other words, the EN estimate of βs' are obtained by minimizing the residual sum of square subject to the constraints of both L_1 and L_2-type penalty on the regression coefficients i.e.,

$$\min \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left\{ (1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right\} \right]$$

for $\alpha = 0$, it is a RR problem and for $\alpha = 1$ it is a LASSO problem. For any other value of α i.e., $0 < \alpha < 1$, it is an EN problem.

RF regression

Let L= $\{(\mathbf{X}_i, Y_i), i=1, 2, ..., n\}$, be the learning dataset of n observations and p explanatory variables, where $\{Y_i\}$ is the continuous response for the vector of explanatory variable $\mathbf{X}_i$. Then, RF consists of B classification trees, where each tree is grown on a bootstrap sample drawn from the original learning dataset L. In each sample, ~37% of observations of the learning set do not play any role in the construction of tree and are called as out of bag (OOB) observations. These observations are treated as test set for the validation. In other words, each such observation is predicted by those classifiers, where it is not drawn and the final prediction is made on the basis of average rule. However, in case of test set, the target response of each instance is predicted by every classifier in the forest and the final response is made by taking the average over all the classifiers.

SNP selection

Since the prediction using RR, LASSO and EN are based on selected number of SNPs (SNPs with nonzero regression coefficients) rather than all SNPs, prediction using RF was also done using certain number of SNPs that are selected on the basis of variable importance. The variable importance of a given variable j was computed by randomly permuting the OOB values on the j^{th} variable. Mathematically,

$importance\,(j) = \frac{1}{B}\sum_{b=1}^{B} E_b\left(OOB_j - E_b\left(OOB_j\right)\right)$. Using the *importance* (j), the SNPs having higher values were identified as significant SNPs. Finally, the SNPs selected under respective models were used for the prediction of trait response.

Optimization of parameters

A set of 100 different values were considered for the tuning parameters i.e. s, t and λ under RR, LASSO and EN methods respectively. Ten-fold cross validation technique (Stone 1974) was followed for each value of tuning parameter and conducted independently for each method. The value of tuning parameter that gives minimum mean cross validation error was selected for building the final model in each case. To get the optimum value of tuning parameters in RF, different combination of *node size*, *mtry* and *ntree* were followed and that combination was taken for the final model building, where the OOB error rate was minimum and stabilized.

Implementation

All the required coding was performed in R statistical software. For the RF, *randomForest* (Liaw and Wiener 2002) and for LASSO, RR and EN, the *glmnet* (Friedman *et al.* 2010) package of R software were used.

Results and discussion

The RF regression methodology was applied on the whole genome SNPs data and the OOB errors are plotted against the number of trees for the final *node size* of 5, 10 and 15 (figure 1). The number of variables tried at each split i.e. *mtry* was taken at six different values, namely, 1, sqrt(p), 20% of p, p/3, 50% of p and p, where p is the number of SNPs in the dataset and the number of trees (*ntree*) to be grown in the forest was kept at 1000. Under each combination of *node size* and number of trees, it was observed that the OOB error was minimum for *mtry* = p/3 (figure 1). Further, with the increase in the *node size* from 5 to 10, the OOB error increased. However, the OOB errors did not change significantly with the

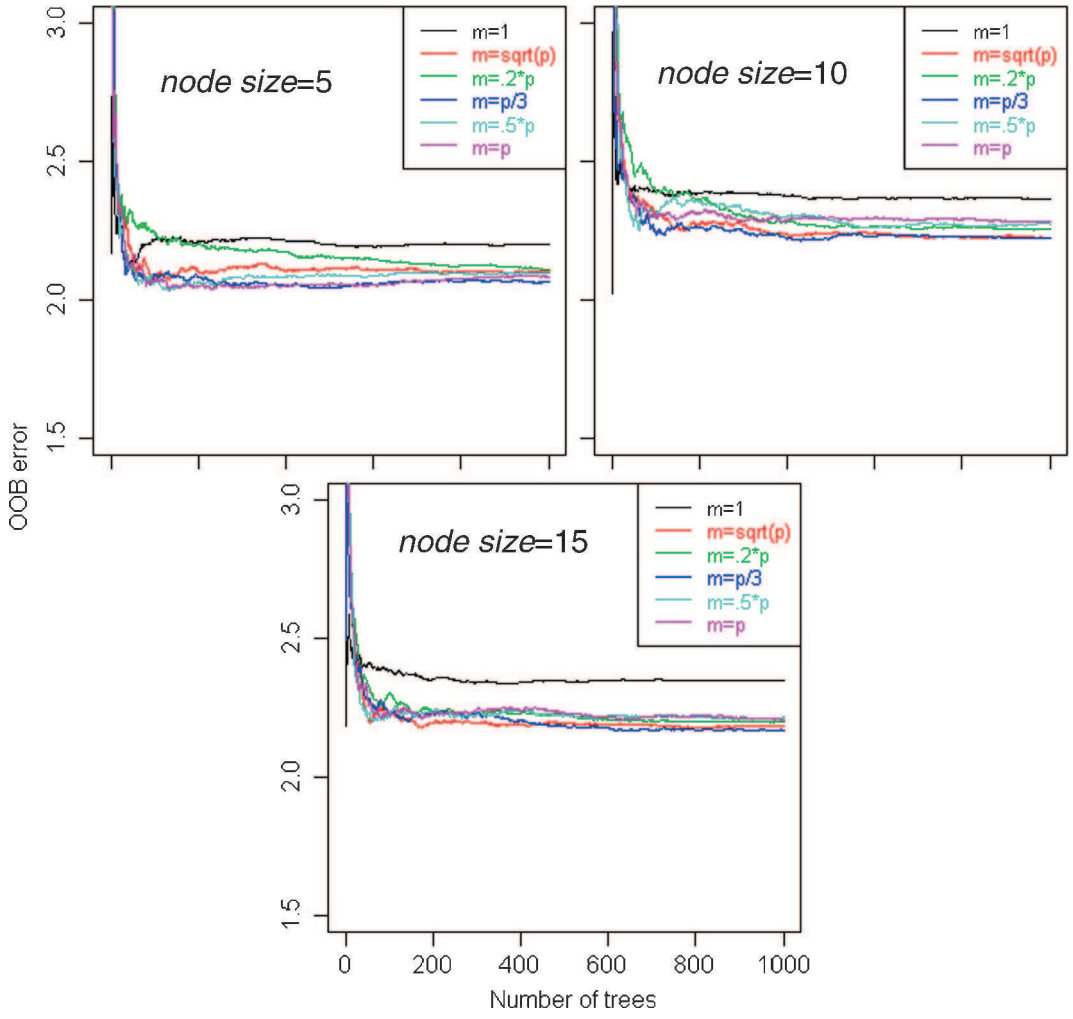

Figure 1. Plot of OOB errors against number of trees for different combination of *node size* and *mtry*.

increase in node size from 10 to 15. Hence, the node size 15 was not considered for further analysis. From the plot of the OOB error, it is also observed that the error stabilized on and around 500 trees.

For the purpose of prediction, the total number of observations were divided into two parts, one as training set and the other as test set. The allocation of observations in the two sets was done randomly to avoid any bias. Also, it was done in four different ways by keeping 10, 20, 30 and 40 per cent of the total observations in the test set, and the remaining in the training set. The RF methodology was applied on the training dataset with two different *node size*s 5 and 10 and the number of trees as 100, 200, 300, 400, 500 and 1000, keeping the *mtry* at *p/3*. The mean square errors (MSE) are computed from the respective test set based on the model that fitted the corresponding training set (table 1). It is observed that the increase in the size of test set has a negative impact on the prediction accuracy. In other words, irrespective of the *node size* and number of trees, the MSE in the test set increases with the reduction in the size of training set. This fact indicates that the RF requires a large chunk of dataset for efficient training of the model.

In comparison to the number of observations, the number of SNPs are very high in the dataset, but a fraction of SNPs may have been associated with the trait of interest. For this reason, instead of considering all the SNPs, a fraction of SNPs were selected based on the value of variable-importance measure of RF model with optimum parameter settings (*mtry* = p/3, *node size* = 5 and *ntree* = 1000) and further analysis was done with the selected SNPs. The percentages (%) of selected SNPs considered were 5, 2.5, 1, 0.5 and 0.25. The correlation between the observed and predicted trait response is observed to be highest as well as significant (at 1% level of significance) in case of 1% selected SNPs (table 2). On the other hand, though the correlation in the case of top 5% selected SNPs is highest, the MSE in both test and training set is higher as compared to 1% selected SNPs. The observed and estimated breeding values of genotypes under different models are given in table 3.

The minimum mean cross validation error for LASSO and RR were found at t=0.11468 and s=0.11468. Similarly, for EN with three different values of alpha, i.e. 0.25, 0.5 and 0.75, minimum values of mean cross validation errors were found at $\lambda = 0.48058$, 0.25173 and 0.17581, respectively. Under the optimum parameter settings, the percentage of SNPs used in RR, LASSO and EN with alpha being 0.25, 0.5 and 0.75 were 100, $\sim$ 0.25, $\sim$ 0.36, $\sim$ 0.25 and $\sim$ 0.25, respectively. With the optimum setting of parameters and selected SNPs, the model was fitted for predicting the continuous trait kernel length of maize genotype by taking 10% observations as test set (as evidenced from RF). The results obtained in terms of MSE and correlations revealed that the performance of RF is better than that of LASSO, RR and EN (table 4).

Prediction of the phenotype of the complex traits without phenotyping the genotypes is of great interest in order to shorten the breeding cycle while making cultivar improvement. The recent advances in high-throughput techniques permit the implementation of efficient genomewide SNP discovery (Mammadov *et al.* 2010). These advances in genotyping coupled with statistical techniques can be effectively used for prediction of breeding values at genome level. In this paper, we presented the effectiveness of the RF methodology for predicting the breeding value over LASSO, RR and EN.

Table 1. MSE for various combination of *node size*, test set size, *ntree* and optimum *mtry*=p/3.

	node size = 5				*node size* =10			
	Proportion of observation as test set				Proportion of observation as test set			
ntree	10	20	30	40	10	20	30	40
100	0.906	1.433	1.689	1.997	0.970	1.411	1.703	2.002
200	0.893	1.405	1.701	1.978	0.970	1.419	1.838	1.995
300	0.948	1.358	1.699	1.953	0.945	1.393	1.753	1.951
400	0.973	1.353	1.701	1.999	0.939	1.355	1.775	1.983
500	0.965	1.353	1.689	1.989	0.946	1.390	1.661	1.959
1000	0.952	1.352	1.718	1.999	0.946	1.392	1.691	1.970

Table 2. Effect of selected SNPs based on variable importance on the MSE of training and test dataset.

Percentage of selected SNPs[a]	MSE in training set	MSE in test set	Correlation
5%	1.863	1.277	0.605**
2.50%	1.745	1.112	0.604**
1%	1.643	0.970	0.608**
0.50%	1.573	0.978	0.598**
0.25%	1.511	0.981	0.591**

**Significant at 1% level of significance. [a]SNPs having higher value of mean decrease in prediction accuracy.

Table 3. The observed and estimated breeding values of genotypes under different models.

Observed value	Predicted value					
	RR	LASSO	EN (α =0.25)	EN (α =0.50)	EN (α =0.75)	RF
6.73	8.34	8.94	8.74	8.86	8.93	8.05
10.32	8.91	9.22	9.37	9.33	9.28	8.81
10.32	9.02	9.47	9.40	9.42	9.42	8.87
8.63	9.05	8.55	8.62	8.57	8.60	8.99
8.62	9.45	9.06	9.05	8.98	8.93	8.90
10.41	8.82	9.50	9.43	9.45	9.47	8.91
8.62	9.44	8.00	8.32	8.12	8.12	8.96
9.64	9.22	9.35	9.33	9.36	9.39	8.79
8.05	6.58	6.86	7.03	7.01	6.98	6.64
9.56	8.94	9.21	9.07	9.17	9.24	8.71
7.17	9.07	8.60	8.64	8.65	8.65	8.59
8.23	8.85	9.07	8.92	9.00	9.00	8.43
8.79	9.76	9.05	9.05	9.00	8.96	8.90
9.49	8.81	9.02	8.90	8.99	8.99	8.79
9.46	10.40	9.03	9.13	9.05	9.04	9.15
10.72	9.99	9.58	9.53	9.57	9.59	9.57
9.35	8.87	8.86	8.85	8.85	8.89	8.88
7.59	6.65	7.46	7.39	7.46	7.47	6.95
10.14	9.63	9.81	9.74	9.80	9.81	9.21
10.25	6.99	8.76	8.77	8.86	8.93	8.58
8.90	8.94	9.45	9.57	9.56	9.56	8.96
10.99	9.15	8.85	9.02	8.94	8.87	8.88
10.28	9.82	8.98	8.95	8.96	8.95	9.17
7.62	7.98	8.26	8.30	8.27	8.29	8.12
10.36	9.95	8.64	8.97	8.88	8.81	8.88
7.82	8.02	8.74	8.70	8.74	8.81	8.45
7.49	8.64	8.39	8.50	8.47	8.42	8.53

Table 4. MSE in training and test sets and correlation between observed and predicted values.

Method	Percentage of SNPs used	MSE in training set	MSE in test set	Correlation b/n observed and predicted trait response
LASSO	0.25	2.286	1.059	0.539**
RR	100	2.286	1.377	0.431**
EN (α =0.25)	0.36	2.239	0.986	0.587**
EN (α =0.5)	0.25	2.336	0.989	0.582**
EN (α =0.75)	0.25	2.281	1.011	0.566**
RF	0.25	1.511	0.981	0.591**

** Significant at 1% level of significance.

All the methods are applied to predict the trait response 'kernel length' based on maize SNP genotyping and phenotyping data.

Optimal choice of the parameters plays an important role in predicting the response or breeding value owing to the fact that nonoptimal choice of these parameters may lead to over fitting or under fitting, and in turn results in a weak prediction. In case of genomewide SNP data, the numbers of SNPs are often larger than the number of observations and hence most of the assumptions of parametric model fail and not appropriate for the prediction. Besides, due to the large number of observations, there is always a chance of over/under fitting of the training model. Thus, the nonparametric method should be preferred in such study. Further, there is no thumb rule in choosing the number of observations for training and validation of the model. However, it has been observed that by taking larger proportion of observation for training often lead to better prediction. In this study, we have identified that the RF requires around 90% of total dataset as training set to capture the larger variability of the complex trait.

Considering that all the SNPs may not have associated with the trait response, SNPs having higher value of mean decrease in prediction accuracy were used to predict the response. Additionally, fitting of the training model with large number of SNPs could increase the cost of fitting model in terms of memory/storage allocation as well as

computational complexity. In RR, there is no option for fitting the model with selected variables but in case of LASSO and EN the models were fitted with a subset of variables, where the remaining variables were excluded due to small value of coefficients. While in RR, LASSO and EN, user has no choice to select significant variables. In RF, the user can select variables on the basis of variable importance for building the prediction model. Hence, the use of RF provided greater flexibility than RR, LASSO and EN. Moreover, RF is a nonparametric method that is robust to noise and occurrence of overfitting. In the present investigation, top 0.25% selected SNPs were used for the prediction under RF. On the other hand, for the prediction of breeding value, complete set of SNPs in RR, $\sim$ 0.25% of SNPs in LASSO, $\sim$ 0.36% of SNPs in EN with $\alpha = 0.25$, $\sim$ 0.25% of SNPs in EN with $\alpha = 0.5$ and $\sim$ 0.25% of SNPs in EN with $\alpha = 0.75$ were considered. The comparison was made in terms of MSE of training and test sets and the correlation between observed and predicted response. From the comparative analysis, the performance of RF was better than the penalized regression techniques. Also, all the estimated correlations between the observed and predicted trait response in the test set under LASSO, RR, EN and RF were significant at 1%. Thus, it is concluded that the RF can be efficiently used for predicting the breeding value from high-density marker data over model-based techniques in crop and animal breeding experiments.

Acknowledgement

R. K. Sarkar acknowledge the receipt of fellowship from PG School, IARI, New Delhi, during his Ph.D.

References

Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32.

Cook J. P., McMullen M. D., Holland J. B., Tian F., Bradbury P., Ross-Ibarra J. *et al.* 2012 Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* **158**, 824–834.

Crossa J., Campos G., Pérez P., Gianola D., Burgueño J., Araus J. L. *et al.* 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724.

Friedman J., Hastie T. and Tibshirani R. 2010 Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.* **33**, 1–22.

Goddard M. E., Wray N. R., Verbyla K. and Visscher P. M. 2009 Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* **24**, 517–529.

Heslot N., Yang H. P., Sorrells M. E. and Jannink J. L. 2012 Genomic selection in plant breeding: a comparison of models. *Crop Sci.* **52**, 146–160.

Hill W. G. 2010 Understanding and using quantitative genetic variation. *Philos. Trans. R. Soc. London, B Biol. Sci.* **365**, 73–85.

Hoerl A. E. and Kennard R. W. 1968 On regression analysis and biased estimation. *Technometrics* **10**, 422–423.

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

Liaw A. and Wiener M. 2002 Classification and regression by random forest. *R News* **2**, 18–22.

Lillehammer M., Meuwissen T. H. E. and Sonesson A. K. 2011 A comparison of dairy cattle breeding designs that use genomic selection. *J. Dairy Sci.* **94**, 493–500.

Mammadov J. A., Chen W., Ren R., Pai R., Marchione W., Yalçin F. *et al.* 2010 Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays* L.] genome for use in marker-assisted breeding. *Theor. Appl. Genet.* **121**, 577–588.

Mammadov J., Aggarwal R., Buyyarapu R. and Kumpatla S. 2012 SNP markers and their impact on plant breeding. *Int. J. Plant Genomics.* Article ID 728398.

Stone M. 1974 Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* **36**, 111–147.

Strobl C., Malley J. and Tutz G. 2009 An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **14**, 323–348.

Tibshirani R. 1996 Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288.

Zou H. and Hastie T. 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67**, 301–320.

Received 13 August 2014, in revised form 13 October 2014; accepted 21 October 2014
Unedited version published online: 29 October 2014
Final version published online: 8 May 2015