Application of Machine Learning-Based Classification to Genomic Selection and Performance Improvement

Zhixu Qiu[1,2], Qian Cheng[1,2], Jie Song[1,2], Yunjia Tang[1,2], and Chuang Ma[1,2](✉)

[1] State Kay Laboratory of Crop Stress Biology for Arid Areas, Northwest A&F University,
Yangling 712100, Shaanxi, China

[2] Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Yangling
712100, Shaanxi, China
cma@nwafu.edu.cn

Abstract. Genomic selection (GS) is a novel breeding strategy that selects individuals with high breeding value using computer programs. Although GS has long been practiced in the field of animal breeding, its application is still challenging in crops with high breeding efficiency, due to the limited training population size, the nature of genotype-environment interactions, and the complex interaction patterns between molecular markers. In this study, we developed a bioinformatics pipeline to perform machine learning (ML)-based classification for GS. We built a random forest-based ML classifier to produce an improved prediction performance, compared with four widely used GS prediction models on the maize GS dataset under study. We found that a reasonable ratio between positive and negative samples of training dataset is required in the ML-based GS classification system. Moreover, we recommended more careful selection of informative SNPs to build a ML-based GS model with high prediction performance.

Keywords: Genomic selection · Marker-assisted breeding · Relative efficiency · Machine learning · Random forest

1 Introduction

Genomic selection (GS) is a promising marker-assisted breeding paradigm that aims to improve breeding efficiency through computationally predicting the breeding value of individuals in a breeding population using information from genome-wide molecular markers (e.g., single nucleotide polymorphisms [SNPs]) [1]. During GS, a prediction model is firstly built with a training population for modeling relationships between high-throughput molecular markers and phenotype of individuals, and then employed to predict the breeding value of individuals in a testing (breeding) population, which are only genotyped but not phenotyped [2]. Individuals with higher prediction scores are finally selected for the breeding experiment. Although GS has been demonstrated to be effective in the breeding of dairy cattle [3], pig [4] and chicken [5], its application in crop breeding is still challenging, in term of high prediction performance, because of the deficiency of robust prediction models for the limited training population size [6],

D.-S. Huang et al. (Eds.): ICIC 2016, Part I, LNCS 9771, pp. 412–421, 2016.
DOI: 10.1007/978-3-319-42291-6_41

the nature of genotype-environment interactions [7], and the complex linkage disequilibrium and interaction patterns between molecular markers [8].

Numerical efforts have been made to develop GS prediction models with regression algorithms for predicting breeding values equal or close to real phenotypic values. Some of representative regression-based GS models are BayesA [1], BayesB [1], BayesC and ridge regression best linear unbiased prediction (rrBLUP) [10, 11]. However, in real breeding situation, it is not necessary to correctly predict phenotypic values of all individuals in a candidate population, because only individuals with high breeding value are selected for further breeding [12]. Therefore, GS has recently been regarded as a classification problem with two classes: individuals with higher phenotypic values and individuals with lower phenotypic values [13]. Some researchers even defined three classes: individuals with upper, middle and lower phenotypic values [14]. For this purpose, the classification-based GS started to be investigated with machine learning (ML) technologies, including random forest (RF) [13], support vector machine (SVM) [13] and probabilistic neural network (ANN) [14]. ML is a branch of artificial intelligence that employs various mathematical algorithms to allow computers "learn" from the experience and to perform prediction on new large datasets [15]. Instead of building a regression curve that fits all the training data, ML-based classification approaches estimate the probability of each individual that belongs to different classes. The superiority of ML-based classification over traditional regression-based approaches has been reported on several crop GS datasets [13]. Nevertheless, the application of ML in GS is still required to be explored, because very little is known about the direction toward the performance improvement of ML-based classification approaches.

Several factors may limit the performance of ML-based classification systems. One is the ratio between positive and negative samples (RPNS) in training dataset, which has been demonstrated in the ML-based prediction of mature miRNAs [16, 17], protein-protein interactions [18] and stress-related genes [19, 20]. For classification-based GS, the prediction model is required to be trained with positive and negative samples generated from the separation of training population according to phenotypes of individuals. However, the effect of RPNS on the prediction performance of ML-based GS classification approaches was rarely explored in the literature [14].

Another factor that influences the prediction accuracy is the number of informative features used to build ML-based prediction systems. In GS, thousands of molecular markers are usually used as the input features of ML-based prediction systems. Due to the limited training population size in many crop GS experiments, it is difficult to model the complex relationships between genome-wide molecular markers and phenotypic values [21]. Known that not all molecular markers are contributed to the trait phenotype [22], selecting a subset of molecular markers that is informative and small enough to deduce prediction models has become an important step toward effective GS [23]. Although hands of feature selection algorithms have been developed for the ML-based classification problems in the research area of bioinformatics and computational biology [24], it is still not clear whether these feature selection algorithms work well in the selection of informative molecular markers for improving the performance of ML-based classification systems in GS programs.

In this study, we developed a bioinformatics pipeline to perform ML-based classification for GS. We employed the random forest (RF) algorithm to build a ML-based classifier named rfGS, and explored the performance of rfGS affected by different factors on a maize GS dataset. We found that an optimized ratio between training positive and negative samples is required for ML-based GS models. Moreover, we confirmed that the selection of molecular markers is an important way of performance improvement, while the rrBLUP (ridge regression best linear unbiased prediction)-based SNP selecting yields better results than mean decrease accuracy (MDA) and mean decrease Gini (MDG), which are widely used in RF-based classification problems.

2 Methods and Materials

2.1 GS Data Set

The GS data set used in this study comprises individuals from 242 maize lines with each individual phenotyped for the grain yield under drought stress. These individuals were genotyped using 46374 single-nucleotide polymorphism (SNP) markers (Illumina MaizeSNP50 array). This data set can be publicly downloaded at the CIMMYT (International Maize and Wheat Improvement Center) website (http://repository.cimmyt.org/xmlui/handle/10883/2976).

2.2 GS Prediction Models

We built GS prediction models with four widely used regression algorithms (ridge regression best linear unbiased prediction [rrBLUP], BayesA, BayesB and BayesC) and one representative ML algorithm random forest (RF). For regression algorithms, the relationships between SNPs and phenotypic values can be generally expressed as $y = \eta + X\beta + ZA\ + e$, where y is the vector of phenotypic values, η is a common intercept, X is a full-rank design matrix for the fixed effects in β, which indicates the factor (e.g., population structure) influences phenotypes, $Z = \sum_k z_k$ is the allelic state at the locus k, $A = \sum_k a_k$ is marker effect at the locus k, and $e \sim \mathrm{N}(0, \sigma_e^2)$ where e is the vector of random residual effects and σ_e^2 is the residual variance [9]. In Z, the allelic state of individuals can be encoded as a matrix of 0, 1 or 2 to a diploid genotype value of AA, AB, or BB, respectively [2].

For rrBLUP, $A \sim N\left(0, \lambda\sigma_a^2\right)$ is calculated as following formula:

$A = \left(Z^T Z + \lambda I_P\right)^{-1} Z^T y$, where $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$ is the ratio between the residual and marker variances. The rrBLUP algorithm was implemented using the "mixed.solve" function in R package rrBLUP (https://cran.r-project.org/web/packages/rrBLUP/index.html).

For the Bayesian regression analysis, the conditional distribution of A can be estimated using the user-given marker information and phenotypic values. The prior distribution can be estimated using different algorithms in the Bayesian framework. We selected BayesA (scaled-t prior), BayesB (two component mixture prior with a point of mass at zero and a scaled-t slab), BayesC (two component mixture prior with a point of mass at zero and a Gaussian slab), respectively. BayesA, BayesB and BayesC were implemented using the "BGLR" function in R package BGLR (https://cran.r-project.org/web/packages/BGLR/index.html).

Random forest, developed by Breiman [25], is a combination of random decision trees. Each tree in the forest is built using randomly selected samples and SNPs. RF outputs the probability of each sample to be the best class based on votes from all trees. RF is a powerful ML algorithm that has been widely applied in many classification problems [26, 27]. The RF algorithm was implemented using the R package randomforest (https://cran.r-project.org/web/packages/randomForest/index.html). The number of constructed decision trees (ntree) was set to be 500, other parameters were used default values.

2.3 SNP Selection

RF-Based SNP Selection. RF provides two built-in measurements for estimating the importance of each feature: MDA and MDG [28]. For a given feature, the MDA quantifies the mean decrease of the predictor when the value of this feature is randomly permuted in the out-of-bag samples, while the MDG calculates the quality of a split for every node of a tree by means of the Gini index. The higher MDA or MDG value indicates the more importance of the feature in the prediction. Both MDA and MDG were calculated by the R package randomforest.

rrBLUP-Based SNP Selection. The rrBLUP model estimates the marker effect for reflecting the importance of each SNP in the prediction of the correlation between genotype and phenotypic values. We selected informative SNPs according to the absolute values of marker effects.

2.4 Performance Evaluation

As previously described [13], the relative efficiency (RE) measurement was used to evaluate the prediction performance of each GS prediction model. Of note, other measurements, such as sensitivity, specificity and area under receiver operating characteristic (ROC), may be also interesting in the GS program. The RE was defined as below:

$$R(\alpha) = \frac{\mu'_\alpha - \mu}{\mu_\alpha - \mu},$$

where μ represents the mean phenotypic value of the whole GS dataset, μ_α denotes the mean of real phenotypic values of the top α individuals with extreme phenotypic values, μ_α is the mean of the real values of extreme individuals (ranked by the predicted values) that have the top α. RE ranges from −1 to 1. A higher RE value indicates a high degree that extreme individuals can be predicted by the classifier. The possible α value ranged from 10 % to 50 % was considered in this study.

Leave-one-out cross-validation (LOOCV) test was used to evaluate the prediction performance and robustness (Fig. 1). In the LOOCV, each individual was picked out in turn as an independent test sample, and all the remaining individuals were used as training samples for building the GS prediction model with rrBLUP, BayesA, BayesB, BayesC or RF algorithm (Fig. 1A–C). This process was repeated until each individual was used as test data one time (Fig. 1A–C). Because sampling strategy was used in the three Bayesian-related regression models and RF-based ML classification model, the LOOCV test was repeated 10 times for calculating the average performance of all tested GS algorithms at each possible percentile value (α).

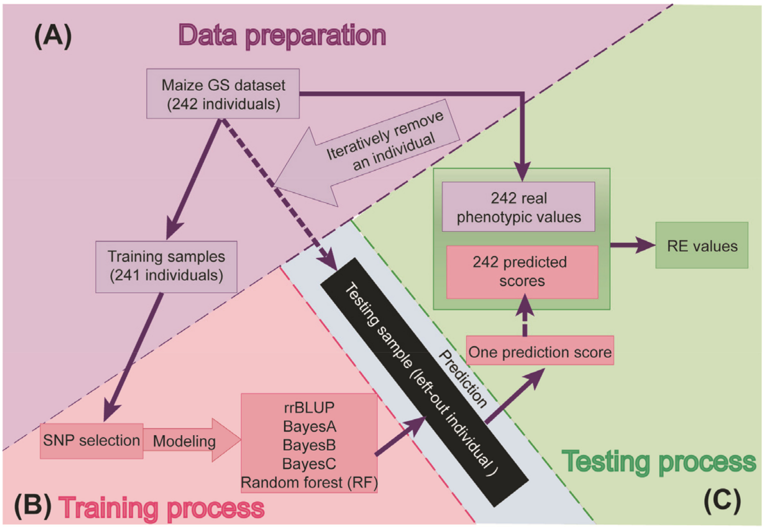

Fig. 1. Overview of LOO cross-validation test for performance evaluation of GS prediction models built with rrBLUP, BayesA, BayesB, BayesC and RF algorithms.

3 Results and Discussion

3.1 Performance Comparison Between rfGS and Four Representative GS Algorithms

The prediction performance of five algorithms (rrBLUP, BayesA, BayesB, BayesC and rfGS) was evaluated using the LOO cross validation test, which iteratively selected one individual as the testing sample and the other individuals as the training samples. The relative efficiency (RE) measurement was used to estimate the prediction accuracy of these algorithms for correctly selecting the best individuals at a given percentile value (α). As shown in Fig. 2, the RE of BayesA gradually decreases from 0.33 to 0.23, when α increases from 10 % to 50 %. Similar results are observed for BayesB and BayesC. Differently, the RE of rrBLUP remarkably decreases from 0.40 to 0.34 when α increases from 10 % to 15 %, but notably increases at higher percentile values (α = 18 %, 22 %, 27 %, 39 % and 47 %). rfGS shows a different pattern of RE compared to the other four algorithms, and reaches the highest RE value (0.53) when α is 14 %. These results indicate that the performance of all five algorithms is influenced by the percentile value.

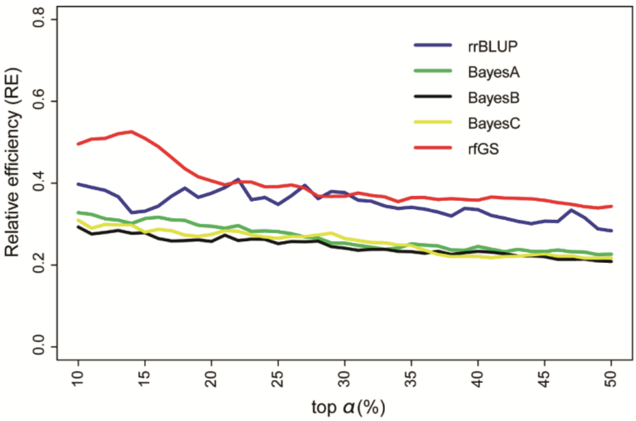

Fig. 2. The relative efficiency (RE) of five GS algorithms at different percentile values (α). (Color figure online)

Compared with BayesA, BayesB and BayesC, rrBLUP yields higher RE values at all tested percentile values. However, we found that the RE can be further improved by using rfGS for almost all tested percentile values. Our result suggests that compared with the widely used regression-based GS algorithms (BayesA, BayesB, BayesC and rrBLUP), RF-based ML classification system rfGS would be an alternative option for the GS program.

3.2 Performance of rfGS is Affected by the Ratio Between Training Positive and Negative Samples

We explored how the performance of rfGS changed with different ratios between positive and negative samples in the training dataset, by selecting the proportion of individuals in the best–worst classes to 20–80, 30–70, 40–60, 50–50 or 60–40.

In Fig. 3, it is shown that the RE of the setting 20–80 gradually decreases from 0.57 to 0.22, when α increases from 10 % to 50 %. Differently, the RE patterns under settings 30–70, 40–60, 50–50, have similar trend that the RE scores first increase when α increases from 10 % to 15 %, and then decrease when α increases from 15 % to 50 %. The RE under the setting 60–40, frequently fluctuates compared to other settings, and has a peak when α is 21 %. The different trends of these RE values under five proportion settings could be explained by the different ability of ML-based classifiers that identify the best individuals under the corresponding ratio. The setting 30–70 showed the best performance among the five different partitions evaluated. Overall, our findings show that the impacts of the ratio between training positive and negative samples on the performance of ML-based GS classifiers should not be neglected, and a reasonable proportion of best–worst classes in the training sets is important for GS program.

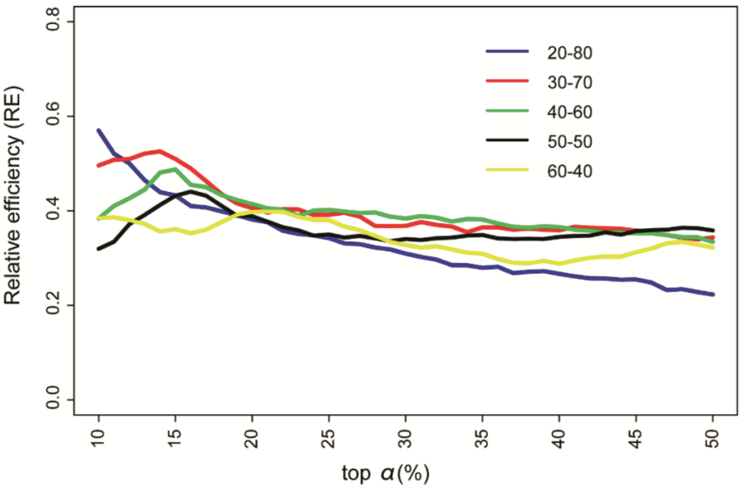

Fig. 3. The relative efficiency of rfGS is affected by the ratio between positive and negative samples in training dataset. (Color figure online)

3.3 Prediction Performance of rfGS Can Be Improved with SNP Selection Process

SNP selection is a process in which a subset of informative SNPs is selected for building GS prediction models. In ML-based classification, MDA and MDG are two powerful feature selection algorithms that are widely used in selected informative features from high-dimensional genomic data. In each round of LOOCV, we estimated the importance of each SNP using the MDA and MDG, respectively, and selected the top N (N = 50, 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000) to build GS prediction models (Fig. 4). We also performed the SNP selection based on the marker effects estimated by the rrBLUP algorithm (Fig. 4). Compared with using all 46374 SNPs, the proportion of predicted accuracy (RE) by selecting top 3000, 4500, 5000, 10000, 30000, 35000 SNPs using MDA increases from −4.48 % to 14.1 % (mean 5.59 % ± 3.47 %) when α increases from 28 % to 35 %. Meanwhile, the proportion is elevated from 0.46 % to 12.61 % (mean 6.29 % ± 2.68 %) by selecting top 4000, 5000, 15000, 20000, 35000 SNPs using MDG with α increasing from 24 % to 35 %. When top 100, 500, 3500, 15000 SNPs were selected, the proportion of prediction accuracy increases from −9.8 % to 34.79 % (mean 9.35 % ± 8.47 %) with a range of α from 20 % to 40 %. rfGS reaches the best performance when selecting the top 10000, 35000, 100 SNPs estimated with MDA, MDG, and rrBLUP algorithms, respectively. Compared to MDA and MDG, rrBLUP-based SNP selection requires the least SNPs to obtain the same prediction ability. It should be noted that, for the GS programs interested in the α ranged from 14 % to 16 % and from 40 % to 50 %, the predicted accuracy consistently decreases in all three SNP selection algorithms, suggesting that more powerful SNP algorithms are urgent to be developed.

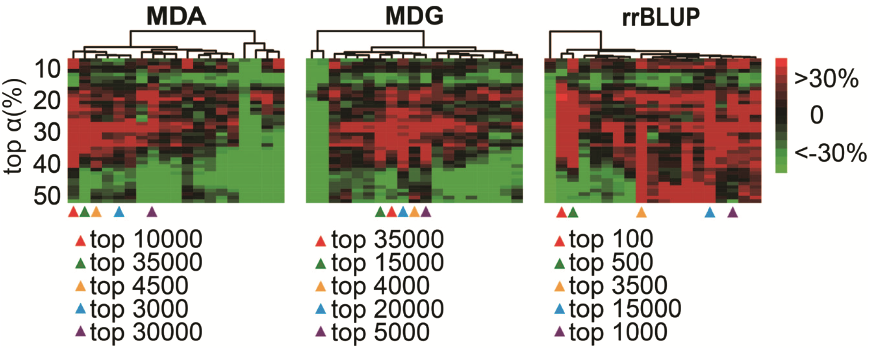

Fig. 4. The performance of mlDNA affected by different SNP selection algorithms. (Color figure online)

Overall, our result shows that the algorithms of selection important SNPs is effective for improving efficiency of GS, and rrBULP-based SNP selection is a promising approach.

4 Conclusions

In this study, we designed a bioinformatics pipeline to perform ML-based classification in GS, exemplified with the application of RF algorithm on a maize GS dataset. RF-based ML classification system rfGS outperforms the widely used regression-based GS algorithms (BayesA, BayesB, BayesC and rrBLUP) on the maize GS dataset under study. Some cautions are raised about the application of ML-based classification to GS. A reasonable proportion of training positive and negative samples is required to increase the prediction accuracy of ML-based GS model. Additionally, SNP selection is also viable to improve efficiency of GS, and rrBULP-based SNP selection is a promising algorithm. In the future, we will apply the graphics processing unit (GPU)-based acceleration technologies to perform the ML-based GS experiments with more complex ML algorithms (e.g., SVM, deep convolutional neural network) and more GS datasets.

Acknowledgement. This work was supported by the grants of the National Natural Science Foundation of China (No. 31570371), Agricultural Science and Technology Innovation and Research Project of Shaanxi Province, China (No. 2015NY011) and the Fund of Northwest A&F University (No. Z111021403 and Z109021514).

References

1. Meuwissen, T.H., Hayes, B.J., Goddard, M.E.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**, 1819–1829 (2001)
2. Desta, Z.A., Ortiz, R.: Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. **19**, 592–601 (2014)

3. Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E.: Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. **92**, 433–443 (2009)
4. Wellmann, R., Preuss, S., Tholen, E., Heinkel, J., Wimmers, K., Bennewitz, J.: Genomic selection using low density marker panels with application to a sire line in pigs. Genet. Sel. Evol. **45**, 28 (2013)
5. Wolc, A., Zhao, H.H., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., Stricker, C., Habier, D., Fernando, R.L., Garrick, D.J., Lamont, S.J., Dekkers, J.C.: Response and inbreeding from a genomic selection experiment in layer chickens. Genet. Sel. Evol. **47**, 59 (2015)
6. Isidro, J., Jannink, J.L., Akdemir, D., Poland, J., Heslot, N., Sorrells, M.E.: Training set optimization under population structure in genomic selection. Theoret. Appl. Genet. **128**, 145–158 (2015)
7. Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Ceron-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., Mathews, K.: Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity **112**, 48–60 (2014)
8. Brito, F.V., Neto, J.B., Sargolzaei, M., Cobuci, J.A., Schenkel, F.S.: Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. BMC Genet. **12**, 80 (2011)
9. Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J.: Extension of the Bayesian alphabet for genomic selection. BMC Bioinform. **12**, 186 (2011)
10. Endelman, J.B.: Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome **4**, 250–255 (2011)
11. de Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.: Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**, 327–345 (2013)
12. Blondel, M., Onogi, A., Iwata, H., Ueda, N.: A ranking approach to genomic selection. PLoS ONE **10**, 0128570 (2015)
13. Ornella, L., Perez, P., Tapia, E., Gonzalez-Camacho, J.M., Burgueno, J., Zhang, X., Singh, S., Vicente, F.S., Bonnett, D., Dreisigacker, S., Singh, R., Long, N., Crossa, J.: Genomic-enabled prediction with classification algorithms. Heredity **112**, 616–626 (2014)
14. Gonzalez-Camacho, J.M., Crossa, J., Perez-Rodriguez, P., Ornella, L., Gianola, D.: Genome-enabled prediction using probabilistic neural network classifiers. BMC Genom. **17**, 208 (2016)
15. Chen, X., Ishwaran, H.: Random forests for genomic data analysis. Genomics **99**, 323–329 (2012)
16. Sturm, M., Hackenberg, M., Langenberger, D., Frishman, D.: TargetSpy: a supervised machine learning approach for MicroRNA target prediction. BMC Bioinform. **11**, 292 (2010)
17. Cui, H., Zhai, J., Ma, C.: MiRLocator: machine learning-based prediction of mature MicroRNAs within plant pre-miRNA sequences. PLoS ONE **10**, e0142753 (2015)
18. Hamp, T., Rost, B.: More challenges for machine-learning protein interactions. Bioinformatics **31**, 1521–1525 (2015)
19. Shaik, R., Ramakrishna, W.: Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice. Plant Physiol. **164**, 481–595 (2014)
20. Ma, C., Xin, M., Feldmann, K.A., Wang, X.: Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in arabidopsis. Plant Cell **26**, 520–537 (2014)

21. Hickey, J.M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B.M., Grondona, M., Zambelli, A., Windhausen, V.S., Mathews, K., Gorjanc, G.: Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. **54**, 1476–1488 (2014)
22. Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P., Haley, C.S.: Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. **5**, 10312 (2015)
23. Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., Avendano, S.: Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. **124**, 377–389 (2007)
24. Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A.R., Gutig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Muller, V., Otto, T., Scholz, C., Schulz, W.A., Seifert, H.H., Schwope, I., Ziebarth, H., Berlin, K., Piepenbrock, C., Olek, A.: Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res. **30**, e21 (2002)
25. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
26. Lloyd, J.P., Seddon, A.E., Moghe, G.D., Simenc, M.C., Shiu, S.H.: Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. Plant Cell **27**, 2133–2147 (2015)
27. Panwar, B., Arora, A., Raghava, G.P.: Prediction and classification of NcRNAs using structural information. BMC Genom. **15**, 127 (2014)
28. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, S.A.: data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief. Bioinform. **14**, 315–326 (2013)