# ACCEPTED ARTICLE:

# Comparing machine learning algorithms and linear model for detecting significant SNPs for genomic evaluation of growth traits in F2 chickens

**Hossein Bani Saadat[a], Rasoul Vaez Torshizi[a]*, Ghader Manafiazar[b], Ali Akbar Masoudi[a], Alireza Ehsani[a], Saleh Shahinfar[c]**

[a] Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran

[b] Department of Animal Science and Aquaculture, Dalhousie University, Truro, NS, Canada.

[c] Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

*Corresponding author:

Rasoul Vaez Torshizi,

Department of Animal Science, Tarbiat Modares

University, Tehran, Iran, P.O. Box: 14115-336. Tel: +98 21 48292003. E-mail address:

rasoult@modares.ac.ir

## ABSTRACT

High-density single nucleotide polymorphisms (SNP) panels are expensive, especially in developing countries, but methods have been developed to detect critical SNPs from these panels and design low-density chips for genomic evaluation at lower cost. This study aimed to determine the efficiency of random forest (RF) and gradient boosting machine (GBM) algorithms, and Linear Model (LM) in identification of SNPs subsets to predict genomic estimated breeding values (GEBVs) for body weights at 6 (BW6) and 9 (BW9) weeks in broiler chickens and compare the predicted GEBVs with those obtained by the 60k SNP panel. The data were collected on 312 $F_2$ chickens that genotyped with 60K Illumina SNP BeadChip. After applying quality control, the remaining 45,512 SNPs were ranked based on p-values, mean square error percentage, and relative influence, obtained by LM, RF and GBM methods, respectively. Then subsets of top 400, 1000, 3000 and 5000 SNPs, selected by each method, employed to construct genomic relationship matrices for the prediction of GEBVs with genomic best linear unbiased prediction model. Results

indicated that predicted accuracies by RF and GBM were generally higher than LM. A Subset of 1000 SNPs selected by RF and GBM algorithms compared to the total SNPs increased accuracy from 0.38 to 0.64 and 0.66 for BW6, and from 0.42 to 0.60 and 0.66 for BW9, respectively. The findings of the present study provide that machine learning methods, especially GBM, can perform better than LM in selecting important SNPs and increase the accuracy of genomic prediction in broiler chickens.

**Keywords:** genomic evaluation, body weight, broilers, machine learning.

## INTRODUCTION

Single nucleotide polymorphisms (SNPs) have been widely utilized in biological research, cancer research, parentage testing, mapping of quantitative trait loci, and evaluation of genomic selection due to their effectiveness as genetic markers. High-density (HD) SNP panels are now accessible for many species due to advancements in high-throughput sequencing technology (Unterseer et al., 2014). One of the important factors in using high-density SNPs is the cost, which is a big limiting factor in utilizing it, especially in developing countries (Mrode et al., 2018). High-density SNP panels used for genomic evaluations have a large number of SNPs that have little to no effect on the traits and could decrease prediction accuracy (Ye et al., 2019). Therefore, various strategies have been performed to select SNPs with large effect from high-density SNP chips, such as selecting SNP evenly spaced across the genome (Habier et al ., 2009) and based on allelic frequency (Abdollahi et al., 2014).

It has been reported that detected subset of SNPs through conventional genome-wide association study (GWAS) increased the accuracy of genomic selection (Liu et al., 2020). On the contrary, Lu et al. (2020) indicated that pre-selecting SNPs based on estimates of variance contributed using weighted single-step genomic best linear unbiased prediction (ssGBLUP) or p-values using single-SNP GWAS did not increase accuracy of genomic predictions substantially in Japanese flounders. In conventional GWAS, a univariate phenotype is regressed on each SNP independently, due to small number of observations and large number of SNPs and LD between SNPs is not considered. Since SNPs are often correlated via linkage disequilibrium (LD), the most significant individual SNPs selected by linear regression may not be an optimal set for creating low-density chips. The undesirable statistical properties of the least squares prediction method for selection of SNPs has also been proposed by Wray et al. (2013).

Machine learning (ML) techniques have been used in GWASs (Mokry et al., 2013). In the context of genome-enabled prediction of phenotypes, ML classification procedure was used by Long et al. (2007) in selection of SNPs for prediction of mortality traits in poultry. Random Forest (RF) (Breiman, 2001) has been applied to GWASs to identify SNP associated with phenotypes and to map QTL on the genomic regions (Minozzi et al., 2014). Gradient Boosting Machine (GBM) (Friedman, 2001) is another popular method of ML algorithm that has gained attention recently. Piles et al. (2021) showed that compared to parametric methods, the best prediction quality in terms of accuracy and stability was obtained with the GBM method for selecting SNPs in order to create low-density SNP chips. The RF and GBM algorithms are suitable alternative to other methods used for genomic evaluations at the expense of lower interpretability of results (González-Recio et al., 2010) and are the most appealing alternatives to analyze complex traits using dense genomic markers information (González-Recio and Forni, 2011).

Several ML algorithms have been used to detect subsets of important SNPs from high-density SNP chips in pig breeds (Schiavo et al., 2020), tropical Brahman cattle (Li et al., 2018) and purebred and commercial Korean native chickens (Seo et al., 2021). Different results have been reported in these studies either in the size of subsets of SNPs or in the outcomes of the methods. To best of our knowledge this approach has not been demonstrated in broiler chickens yet and will serve poultry industry with better insight on utilization of ML techniques in preselection of SNPs to enhance the accuracy of genomic selection. Therefore, the present study aimed to evaluate the efficiency of two ML algorithms, namely RF and GBM, in identifying a subset of SNPs affecting growth traits using a crossbreed chicken population for the genomic selection purpose. The accuracy of genomic breeding values predicted by subsets of SNPs selected by ML algorithms were compared with conventional GWAS and all available SNP set.

## MATERIALS AND METHODS

### Experimental population, phenotypic and genotypic data

A population of $F_2$ crosses between the fast-growing Arian line (AA) and the slow-growing Urmia Iranian indigenous chickens (NN) was used in this study. The $F_1$ birds were generated from the mating of AA ♂ × NN ♀ and NN ♂ × AA ♀ birds and reared for 12 weeks in poultry research farm of Tarbiat Modares University, Tehran, Iran. Then $F_1$ males from each reciprocal cross were mated each to 4–8 females from other families and $F_2$ chickens were produced. Chickens of $F_2$ generation were raised individually in cages equipped with water nipples and feeders for 12 weeks

under the same environmental conditions and ration. Individual weekly weight was collected throughout the growing period. A total of 312 birds from six different hatches were available. For the present study, body weights recorded at 6 (BW6) and 9 (BW9) weeks were used. More information about these traits can be found in Emrani et al. (2017). Before implication of ML, a multiple linear regression of observations on sex and hatch was used to adjust the body weight data (Brown and Reverter, 2002).

Genomic DNA was extracted from 312 blood samples using salting out method and stored at -20°C. After extraction, spectrophotometry and agarose gel electrophoresis methods were used to determine the quantity and quality of DNA. These DNA samples were genotyped with the Illumina Chicken 60K SNP BeadChip, in cooperation with Cobb-Vantress Inc., and the Aarhus University, Denmark. Quality control steps were applied to the original data with PLINK 1.9 software (Purcell et al., 2007). SNPs with call rate of <95%, minor allele frequency of <5%, a Hardy– Weinberg equilibrium test p-value $<1\times 10^{-6}$ were deleted (Emrani et al., 2017). After quality control, 45512 of SNPs for twenty-eight autosome chromosomes and 300 birds remained for final analysis.

**Methods for selecting markers**

The linear model for conventional GWAS was as follows:

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{Z}\mathbf{q} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of corrected phenotypic values for BW6 and BW9, $\mathbf{1}$ is an n-vector of ones, $\boldsymbol{\mu}$ is the population mean, $\mathbf{q}$ is the effect of the marker in the model, which is treated as a fixed regression of observation on genotype, $\mathbf{Z}$ is a vector containing genotypes of the marker with 0, 1 and 2 for $A_1A_1$, $A_1A_2$ and $A_2A_2$, respectively, and $\mathbf{e}$ is a vector of random residual effects, assuming $\mathbf{e}\sim N(0, \mathbf{I}\sigma_e^2)$, where $\sigma_e^2$ is the residual variance and $\mathbf{I}$ is the identity matrix. The genetic association tests were conducted using the '--Linear' command in PLINK v1.9 (Purcell et al., 2007). The SNPs were selected based on the p-values from GWAS results.

In the RF algorithm, which contains several decision trees, a bootstrap sample of original training data is used to grow each tree. The RF algorithm predicts the outcome by averaging the outputs obtained from all the trees in the forest (Breiman, 2001). When making bootstrap samples to grow each tree, approximately 34 percent of records will not be selected, which is called Out Of Bag (OOB) records. To calculate importance of each SNP, OOB error was calculated by predicting the outcome of OOB samples via the corresponding tree. Then the values of each predictor were

permuted (shuffled) and prediction error of OOB samples were calculated again. The mean square error percentage (MSEP) difference between permuted and non-permuted samples (averaged over all the trees in the forest) indicated the importance or predictive ability of that particular predictor. The 'randomForest' package was used to perform this analysis in R software (Breiman, 2013).

In the GBM algorithm, the basic functions are weak learners such as a decision trees. The purpose of the Boosting algorithm is to enhance ensemble of weak learners into a strong learner. In this method, a basic learner such as a decision trees are added sequentially to the residuals of the previous tree, and it is expected that by focusing on the incorrectly predicted data in the previous tree, error rate in the next tree will be lessened and as long as the error rate is decreasing, the boosting algorithm will continue (Friedman, 2001). In the present study, important markers in the GBM method are identified by relative influence (RI), which is the average of reduction in MSEP over all the trees when that particular SNP to split the data (Friedman, 2001). The 'gbm' package was used to perform this algorithm in R software (Greenwell et al., 2019). For GBM and RF methods, hyper-parameters tuning performed via nested grid search within a 3-fold cross-validation on the 75 percent randomly selected subset of the data.

**Genome-wide screening for top ranking SNPs**

All SNPs were ranked from the most important to the least important SNP by criteria values of RF (increase in MSEP), GBM (RI), and LM (p-value) using 'dplyr' package implemented in R (Wickham et al., 2023). For the 5000 number of important SNPs, obtained from LM, RF and GBM, venn diagrams were drawn by the 'VennDiagram' package (Chen and Boutros, 2011). Top 400, 1,000, 3000, and 5000 SNPs with the above-mentioned criteria were used to create genomic relationship matrices.

**Genomic estimated breeding value**

Genomic estimated breeding values (GEBV) were derived using genomic best linear unbiased prediction (GBLUP) model. The statistical model of GBLUP is written as follows (Gianola et al., 2006):

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ is an n-vector of corrected phenotypes, $\mathbf{1}$ is an n-vector of ones, $\boldsymbol{\mu}$ is the population mean, $\mathbf{g}$ is a vector of random additive genomic values with $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2$, where $\mathbf{G}$ is the additive genomic

relationship matrix between genotyped individuals and $\sigma_g^2$ is the additive genomic variance, and **e** is the vector of random residual effects with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$,, where $\sigma_e^2$ is the residual variance, and **I** is the identity matrix. The additive genomic relationship matrix (**G**) is constructed as $\frac{\mathbf{ZZ}'}{\mathbf{m}}$, where **Z** is the matrix of centered and standardized genotypes for all individuals and **m** is the number of markers. Kernel Hilbert space regression method was used to implement the GBLUP approach and the genomic heritability in the selected subsets and all markers was estimated using the Bayesian Generalized Linear Regression (BGLR) package (Pérez-Rodríguez and de Los Campos, 2022) in R software. The Gibbs sampler was run for 50,000 iterations, with a 10,000 burn-in period and a thinning interval of 5 iterations, i.e., 10,000 samples were used for inference.

**Cross-validation for the accuracy of genomic breeding values**

Accuracy of genomic prediction was calculated on 5-fold cross-validation base as follows (Li et al., 2018):

$$\text{Accuracy} = \frac{r_{GEBV,phen}}{\sqrt{h^2}}$$

Where $r_{GEBV,phen}$ is correlation coefficient between the predicted GEBVs of the birds in the test fold and the corrected phenotypes (phen) and $h^2$ is estimated heritability of the trait.

Unbiasedness of genomic prediction was calculated on 5-fold cross-validation base as follows:

$$b_{GEBV,phen} = r_{GEBV,phen}(S_{phen}/S_{GEBV})$$

Where $b_{EBV,phen}$ is regression coefficient of corrected phenotypes on GEBV that show unbiasedness of the GEBV, $r_{GEBV,phen}$ is correlation coefficient between the predicted GEBVs of the birds in the test fold and the corrected phenotypes, $S_{phen}$ is the standard deviation of corrected phenotypes and $S_{GEBV}$ is the standard deviation of predicted GEBVs. Finally, the Tukey HSD (Honestly Significant Difference) test was used to compare the significant differences between the best subsets of SNPs which had the highest increase in genomic prediction accuracy with each other and the all SNPs.

**RESULTS AND DISCUSSION**

The rank of SNPs from the most important to the least important for BW6 and BW9 are shown in Figure 1. Based on LM method, the 5000 pre-selected SNPs had a p-values range from $1.01 \times 10^{-5}$ to $7.60 \times 10^{-2}$ and $7.57 \times 10^{-6}$ to $8.09 \times 10^{-2}$ for BW6 and BW9, respectively. For RF method,

the importance of SNPs changes from positive to negative values. The highest positive value in RF indicates an increase in the MSEP when the SNP is randomly permuted compared to the prediction error before SNP permutation. In this model, 47%, 7%, and 46% of SNPs for BW6 and 47%, 9% and 44% of SNPs for BW9 had positive, zero, and negative effects, respectively. About 5% of SNPs for BW6 and 2.8% for BW9 (5000 pre-selected SNPs) had a MSEP increase more than 0.2, respectively. In the GBM method, 26% and 16% of SNPs had larger than zero effect for BW6 and BW9, respectively. In 5000 pre-selected SNPs with GBM method, none of the SNPs had a zero RI, however, 65.3% of SNPs for BW6 had a RI less than one and close to zero. For BW9, the amount of RI for last SNP of the 5000 pre-selected SNPs was 58.10%, and 52.52% of SNPs had a RI less than 1000. Based on this method, about 3.62% of SNPs for BW6 and 5.06% for BW9 (5000 pre-selected SNPs) had a RI more than 10000, respectively.

The total number of common SNPs between three methods are shown visually by Venn diagrams in Figure 2 for the top 5000 SNPs. A total of 924 and 1100 SNPs was common across three methods for BW6 and BW9, respectively. The results indicated that the similarity between RF and GBM method was higher than that observed between LM with RF and GBM. The estimates of genomic heritability for body weight traits using the genomic relationships matrix consisting of all or subsets of selected SNPs for three methods are presented in Figure 3. Genomic heritability for BW6 and BW9 consisting of all SNPs was estimated to be 0.28 and 0.30, respectively. These estimates were consistent with the studies of Demeure et al. (2013) and Abdollahi et al. (2014), who fitted all SNPs and reported a moderate estimates of 0.22 and 0.30, respectively, for growth traits in chickens. Genomic heritability estimation was increased when the matrix of genomic relationships was constructed by subsets of SNPs pre-selected by any of the three proposed methods in the present study. Pre-selected subsets of SNPs by GBM showed the highest rate of increase in genomic heritability in comparison with LM and RF.

The highest estimates of heritability for BW6 with pre-selected SNPs by LM, RF, and GBM methods were, 0.42, 0.39 and 0.48, respectively, in subsets of 5000 SNPs (in LM) and 1000 SNPs (in RF and GBM). For BW9, the highest heritability was 0.43, 0.46 and 0.58 in subsets of 5000 SNPs (in LM), 1000 SNPs (in RF) and 3000 SNPs (in GBM), respectively. In comparison to all 45,512 SNPs, the heritability estimates were increased from 0.28 to 0.35 and from 0.30 to 0.46 through preselection of 5000 SNPs using LM model for BW6 and BW9, respectively. By using 1000 pre-selected SNPs, the increases in the heritability estimates were ranged from 0.28 to 0.37

for BW6 and 0.30 to 0.43 for BW9 with RF model and from 0.28 to 0.48 for BW6 and 0.30 to 0.54 for BW9 with GBM model. Ren et al. (2022) indicated that high-density SNP data provide more information for genomic evaluation compared to medium-density SNP data but they do not confer any advantage for heritability estimation. Literature studies reported that the estimates of genomic heritability were very sensitive to differences in LD between SNPs, suggesting that genomic heritability is overestimated in region with high LD and underestimated in region with low LD (Speed et al., 2012). The stronger LD of the remaining SNPs and the removal of the imperfect LD between the causal mutations may improve the genomic relationships between individuals and increase the heritability of the trait (Abdollahi et al., 2014; Ye et al., 2019). Abdollahi et al. (2014) estimated genomic heritability for body weight at 6 weeks in broilers chickens using the genomic relationship matrix consisting of all SNPs and a subset of selected SNPs and reported that the genomic heritability with selected SNP (0.59) is expected to be overestimated in comparison to all SNPs (0.30), however, the subsets of SNPs could increase the GEBV accuracy. The increase in the accuracy of GEBV has been reported by Luo et al, (2021) who proposed a strategy for genomic selection in aquaculture using a subset of markers selected by the p-value of GWAS and indicated that the prediction accuracy of a subset of top SNPs was higher than using total SNPs. Li et al. (2018) reported that ML methods can consider complex and nonlinear relationships. Therefore, they can produce a smaller error variance and increase genetic variance and heritability. These authors estimated the heritability of a subset of 3000 SNPs with GBM method to be higher than all of 38082 SNPs for body weight in Brahman cattle.

Figure 4 and Figure 5 show the mean accuracy and regression coefficient (as a measurement of unbiasedness) of genomic breeding value for BW6 and BW9 traits using SNP subsets in a 5-fold cross-validation scheme, respectively. Accuracy of genomic prediction for BW6 and BW9 using all SNPs was estimated to be 0.38 and 0.42, respectively, which was lower than the genomic prediction accuracy obtained from the subsets of selected SNPs (400, 1000, 3000 and 5000 selected with three methods). Average accuracy of genomic breeding value (± standard error) with top 400, 1000, 3000, and 5000 SNP subsets selected by LM, RF and GBM methods were estimated to be 0.56 (± 0.02), 0.61 (± 0.04), 0.52 (± 0.02) and 0.47 (± 0.02) for BW6, and 0.58 (±0.02), 0.61 (±0.02), 0.55 (±0.02) and 0.51 (±0.01) for BW9, respectively. Mean regression coefficient of genomic prediction on phenotype using total SNPs for BW6 and BW9 were estimated to be 0.76 and 0.90, respectively. With top 400, 1000, 3000, and 5000 SNP subsets selected by three methods,

mean regression coefficient (± standard error) were 0.94 (± 0.05), 0.97 (± 0.04), 0.94 (± 0.02) and 0.95 (± 0.01) for BW6, and 1.06 (± 0.01), 1.03 (± 0.02), 1.05 (± 0.04) and 1.06 (± 0.05) for BW9, respectively. The best average accuracy of genomic breeding value and regression coefficient provided by 1000 SNP subset was 0.61 (± 0.04) and 0.97 (± 0.04) for BW6 and 0.61 (±0.02) and 1.03 (± 0.02) for BW9, respectively. In the study of Liu et al. (2020), the highest accuracy of genomic breeding value by a subset of 817 SNPs selected from high-density SNP panels was 0.60 for body weight at the age of 12 weeks and by a subset of 354 SNPs was 0.45 for feed conversion ratio in broiler. Furthermore, several studies indicated a direct relationship between effective population size and the accuracy of GEBVs. The significant impact of smaller effective population size on the prediction accuracy of GBLUP has been revealed by Daetwyler et al. (2010), which is a reflection of strong linkage disequilibrium between variants due to close genetic relatedness between individuals (Jang et al., 2023; Calus et al., 2008).

Significant differences between genomic prediction accuracy of the best subsets of SNPs (which had the highest increase in genomic prediction accuracy) with each other and all SNPs are presented in Table 1. The results showed that 1000 SNPs selected by ML algorithms was the best pre-selected SNPs for estimating genomic breeding value in broiler chickens for body weight traits in the present study. In BW6, there was no significant difference between RF and GBM algorithms in the best subsets (1000 SNPs) of selected SNPs, and they were superior to linear model with the best subset (3000 SNPs). However, in BW9, GBM was superior to other methods. These findings are consistent with the results of Kriaridou et al. (2020), who used different subsets of SNPs in four aquaculture datasets, ranging from 100 to 9000 SNPs, and observed that SNP densities between 1000 and 2000 SNPs had a very similar accuracy of genomic evaluation to high-density genotyping. Ye et al. (2019) used selected markers from whole-genome sequencing data based on the p-value obtained from GWAS and showed that the use of pre-selected markers for most traits did not increase the genomic prediction accuracy in broilers and even increased the bias. One of the possible reasons is the difficulty of discovering causative variants using GWAS due to the large number of variants (600k) and high LD between variants. On the contrary, Li et al. (2018) indicated an increase in the accuracy of genomic prediction by selecting a subset of significant SNPs from high-density SNP panel (651,253) using RF method. One of the advantages of ML method is its ability to analyze data with a high dimension, however, factors such as linkage disequilibrium and minor allele frequency can affect the performance of ML algorithms for selecting important

markers (Zhou and Troyanskaya, 2015). Decision trees are known to have low bias and high variance in prediction, but RF overcomes this issue by forming many trees on each bootstrap sample, to minimize prediction errors by lowering the variance of prediction. In the GBM, both bias and variance are expected to be reduced due to the boosting process which is the assembling multiple weak learners sequentially and using the weighted average of each tree for prediction (Li et al., 2018). Hence ML can be superior to linear models for selecting SNPs from high-density SNP panels.

Literature results on improving accurate prediction of breeding values using high-density SNP genotype, even with implementation of a specific model, are inconsistent. Several studies indicated that selecting markers from high-density genomic data can be resulted in a small improvement in genomic accuracy (Lopez et al., 2020). Our strategy for screening SNPs in two growth traits improved estimation of genomic breeding value accuracies. A Subset of 1000 SNPs selected by the RF and GBM methods compared to the total SNPs increased the accuracy of genomic prediction from 0.38 to 0.64 and 0.66 for BW6 and from 0.42 to 0.60 and 0.66 for BW9, respectively. Liu et al. (2020) improved genomic prediction accuracy for body weight traits in broiler chickens by selecting a subsets of SNPs based on p-values obtained from GWAS, revealing that high prediction accuracy for growth traits may be achieved even with a small number of markers. SNPs that are not close to causal mutations may have a negative impact on genomic prediction. Also, many SNPs may not tag any causative mutations when the number of markers is too large. Therefore, if only effective SNPs that tag any causative mutations are included in the model, the ability of the model to predict genomic breeding value may be increased and the model error is decreased by removing the unrelated markers. Druet et al. (2014) showed that the accuracy of genomic prediction depends largely on the coverage of key genes affecting target traits by genotyping platforms.

## CONCLUSIONS

The genomic selection has become one of the main techniques for animal breeding programs. High costs of genotyping has limited the use of genomic selection in poultry due to the large number of selection candidate, especially in developing countries. Therefore, selecting effective SNPs is useful in designing low-density panels which could provide broad potential and applicability in genomic evaluation. In the present study, the accuracy of GEBV for BW6 and BW9 obtained from a subset of pre-selected 1000 SNPs by RF and GBM performed better than the subset selected by

LM, indicating that ML algorithms can be used as a selection tools to find significant markers for designing and developing low-density SNP marker panels. However, due to the small population size of the current study, further studies with more data, different methods and a wide range of different SNP subsets are needed to find optimum and reliable set of subsets.

## REFERENCES


Abdollahi-Arpanahi, R., Nejati-Javaremi, A., Pakdel, A., Moradi-Shahrbabak, M., Morota, G., Valente, B.D., Kranis, A., Rosa, G.J.M. and Gianola, D. 2014. Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens. *J. Anim. Breed. Genet*., **131 (2)**: 123-133.

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

Breiman, L. 2013. Breiman and Cutler's random forests for classification and regression. Package 'randomForest'. Institute for Statistics and Mathematics, Vienna University of Economics and Business.

Brown, D.J. and Reverter, A. A. 2002. Comparison of methods to pre-adjust data for systematic effects in genetic evaluation of sheep. *Livest. Prod. Sci.*, **75**:281–91.

Calus, M.P.L., Meuwissen, T.H.E., De Roos, A.P.W. and Veerkamp, R.F. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genet.*, **178**: 553–561.

Chen, H. and Boutros, P.C. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**:35.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B. and Woolliams, J.A. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genet.*, **185**:1021–31.

Demeure, O., Duclos, M.J., Bacciu, N., Le Mignon, G., Filangi, O., Pitel, F., Boland, A., Lagarrigue, S., Cogburn, L.A., Simon, J. and Le Roy, P. 2013. Genome-wide interval mapping using SNPs identifies new QTL for growth, body composition and several physiological variables in an F 2 intercross between fat and lean chicken lines. *Genet. Sel. Evol*., **45 (1)**: 1-12.

Druet, T., Macleod, I.M. and Hayes, B.J. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, **112 (1)**: 39-47.

Emrani, H., Torshizi, R.V., Masoudi, A.A. and Ehsani, A. 2017. Identification of new loci for body weight traits in F2 chicken population using genome-wide association study. *Livest. Sci*., **206**: 125-131.

Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Ann. statist*., **29 (5)**.1189-1232.

Gianola, D., Fernando, R.L. and Stella, A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genet*., **173 (3)**: 1761-1776.

González-Recio, O.; Forni, S. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet. Sel. Evol*., **43 (1)**: 7.

González-Recio, O.; Weigel, K. A.; Gianola, D.; Naya, H., Rosa, G. J. M. 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet. Res*., **92 (3)**: 227–237.

Greenwell, B., Boehmke, B., Cunningham, J., Developers, G.B.M. and Greenwell, M.B. 2019. Package 'gbm'. *R package version*, **2 (5)**.

Habier, D., Fernando, R.L. and Dekkers, J.C. 2009. Genomic selection using low-density marker panels. *Genet*., **182 (1)**: 343-353.

Jang, S., Tsuruta, S., Leite, N.G., Misztal, I. and Lourenco, D. 2023. Dimensionality of genomic information and its impact on genome-wide associations and variant selection for genomic prediction: a simulation study. *Genet. Sel. Evol.* **55**, 49.

Kriaridou, C., Tsairidou, S., Houston, R.D. and Robledo, D. 2020. Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. *Front. Genet.*, 11:124.

Li, B., Zhang, N., Wang, Y.G., George, A.W., Reverter, A. and Li, Y. 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet*., **9**: 237.

Li, Y., Raidan, F.S.S., Vitezica, Z. and Reverter, A. 2018. Using Random Forests as a prescreening tool for genomic prediction: impact of subsets of SNPs on prediction accuracy of total genetic values. World Congress on Genetics Applied to Livestock Production., 1130. Massey University.

Liu, T., Luo, C., Ma, J., Wang, Y., Shu, D., Su, G. and Qu, H. 2020. High-throughput sequencing with the preselection of markers is a good alternative to SNP chips for genomic prediction in broilers. *Front. Genet*., **11**: 108.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A., Avendaño, S. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J. Anim. Breed. Genet*., **124 (6)**: 377–389.

Lopez, B.I., Lee, S.H., Shin, D.H., Oh, J.D., Chai, H.H., Park, W., Park, J.E. and Lim, D. 2020. Accuracy of genomic evaluation using imputed high-density genotypes for carcass traits in commercial Hanwoo population. *Livest. Sci*., **241**:104256.

Lu, S., Liu, Y., Yu, X., Li, Y., Yang, Y., Wei, M., Zhou, Q., Wang, J., Zhang, Y., Zheng, W. and Chen, S. 2020. Prediction of genomic breeding values based on pre-selected SNPs using ssGBLUP, WssGBLUP and BayesB for Edwardsiellosis resistance in Japanese flounder. *Genet. Sel. Evol*., **52**: 49.

Luo, Z., Yu, Y., Xiang, J. and Li, F. 2021 Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. *Aquaculture*., 539:736620.

Minozzi, G., Pedretti, A., Biffani, S., Nicolazzi, E.L. and Stella, A. 2014. Genome wide association analysis of the 16th QTL- MAS Workshop dataset using the Random Forest machine learning approach. *BMC proc*., **5**: 1-6.

Mokry, F.B., Higa, R.H., de Alvarenga Mudadu, M., Oliveira de Lima, A., Meirelles, S.L.C., Barbosa da Silva, M.V.G., Cardoso, F.F., Morgado de Oliveira, M., Urbinati, I., Meo Niciura, S.C. and Tullio, R.R. 2013. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. *BMC genet*., **14 (1)**: 1-11.

Mrode, R., Tarekegn, G.M., Mwacharo, J.M. and Djikeng, A. 2018. Invited review: Genomic selection for small ruminants in developed countries: how applicable for the rest of the world? *Anim.*, **12 (7)**: 1333-1340.

Pérez-Rodríguez, P. and de los Campos, G. 2022. Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. *Genetics*., **222(1)**: 112.

Piles, M., Bergsma, R., Gianola, D., Gilbert, H. and Tusell, L. 2021. Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. *Front. Genet*., **12**: 137.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81 (3)**: 559-575.

Ren, D., Cai, X., Lin, Q., Ye H., Teng, J., Li, J., Ding, X. and Zhang, Z. 2022. Impact of linkage disequilibrium heterogeneity along the genome on genomic prediction and heritability estimation. *Genet. Sel. Evol*.,**54 (1)**: 47.

Schiavo, G., Bertolini, F., Galimberti, G., Bovo, S., Dall'Olio, S., Nanni Costa, L., Gallo, M., Fontanesi, L. 2020. A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds. *Anim*., **14 (2)**: 223-232.

Seo, D., Cho, S., Manjula, P., Choi, N., Kim, Y.K., Koh, Y.J., Lee, S.H., Kim, H.Y., Lee, J.H. 2021. Identification of Target Chicken Populations by Machine Learning Models Using the Minimum Number of SNPs. *Anim*., **11 (1)**: 241.

Speed, D., Hemani, G., Johnson, Michael, R., Balding, David, J. 2012. Improved Heritability Estimation from Genome- wide SNPs. *Am. J. Hum. Genet*., **91**: 1011–1021.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R., Pausch, H. and Bertani, C. 2014. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC genomics.*, **15(1)**, pp.1-15.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. 2023. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E. and Visscher, P.M. 2013. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet*., **14 (7)**: 507–515.

Ye, S., Gao, N., Zheng, R., Chen, Z., Teng, J., Yuan, X., Zhang, H., Chen, Z., Zhang, X., Li, J. and Zhang, Z. 2019. Strategies for obtaining and pruning imputed whole-genome sequence data for genomic prediction. *Front. Genet*., **10**: 673.

Zhou, J. and Troyanskaya, O.G. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*., **12 (10):** 931-934.

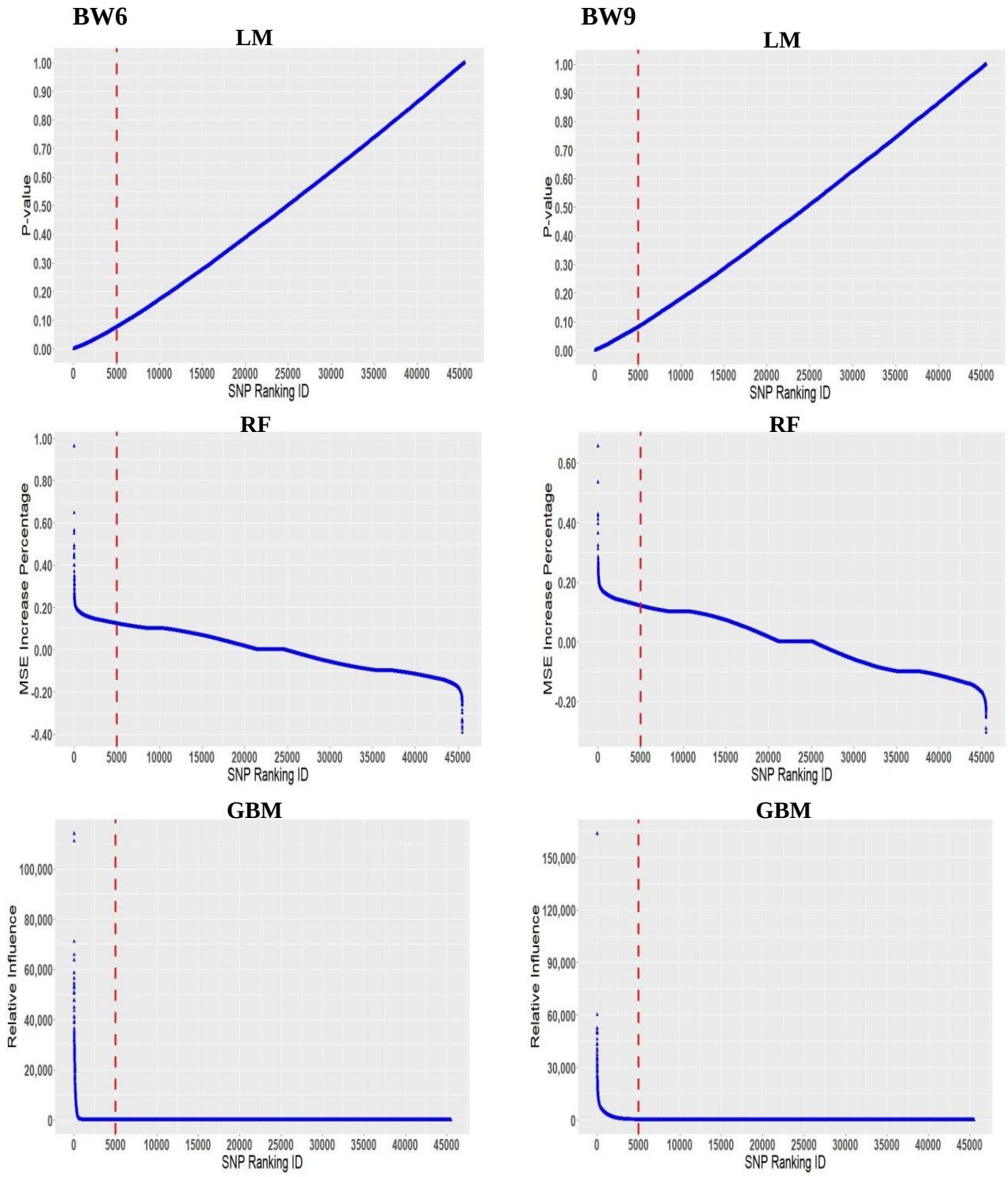


**Figure 1.** The distribution of ranked SNP for BW6 and BW9 from LM, RF and GBM methods.


**Figure 2.** Venn diagram showing the 5000 number of important SNPs from RF, GBM, and LM methods. Circle represents the number of identified SNPs and the intersection areas represent the number of overlapping SNPs.

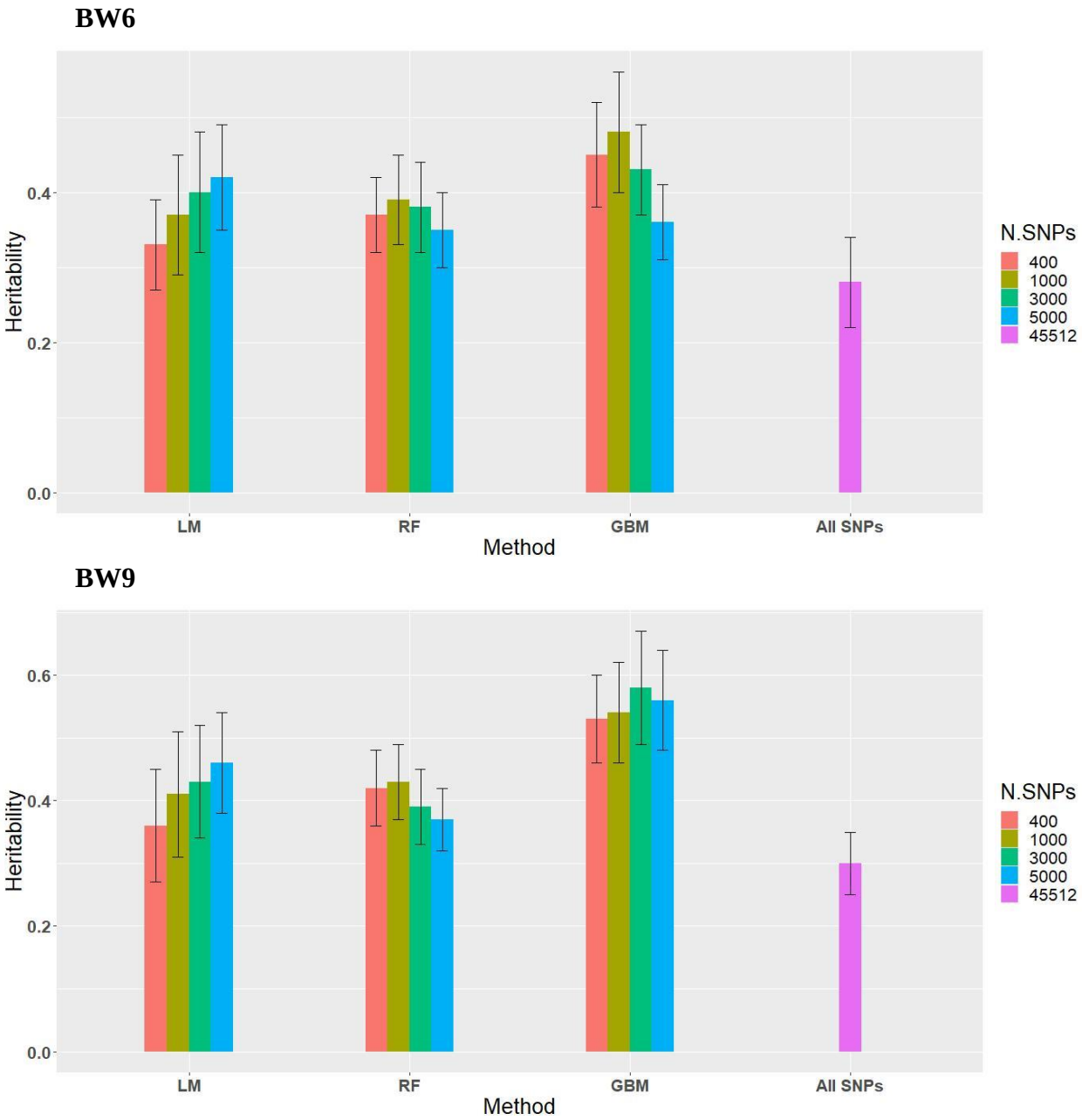


**Figure 3.** Genomic heritability of different subsets of SNPs for BW6 and BW9 from LM, RF and GBM methods.

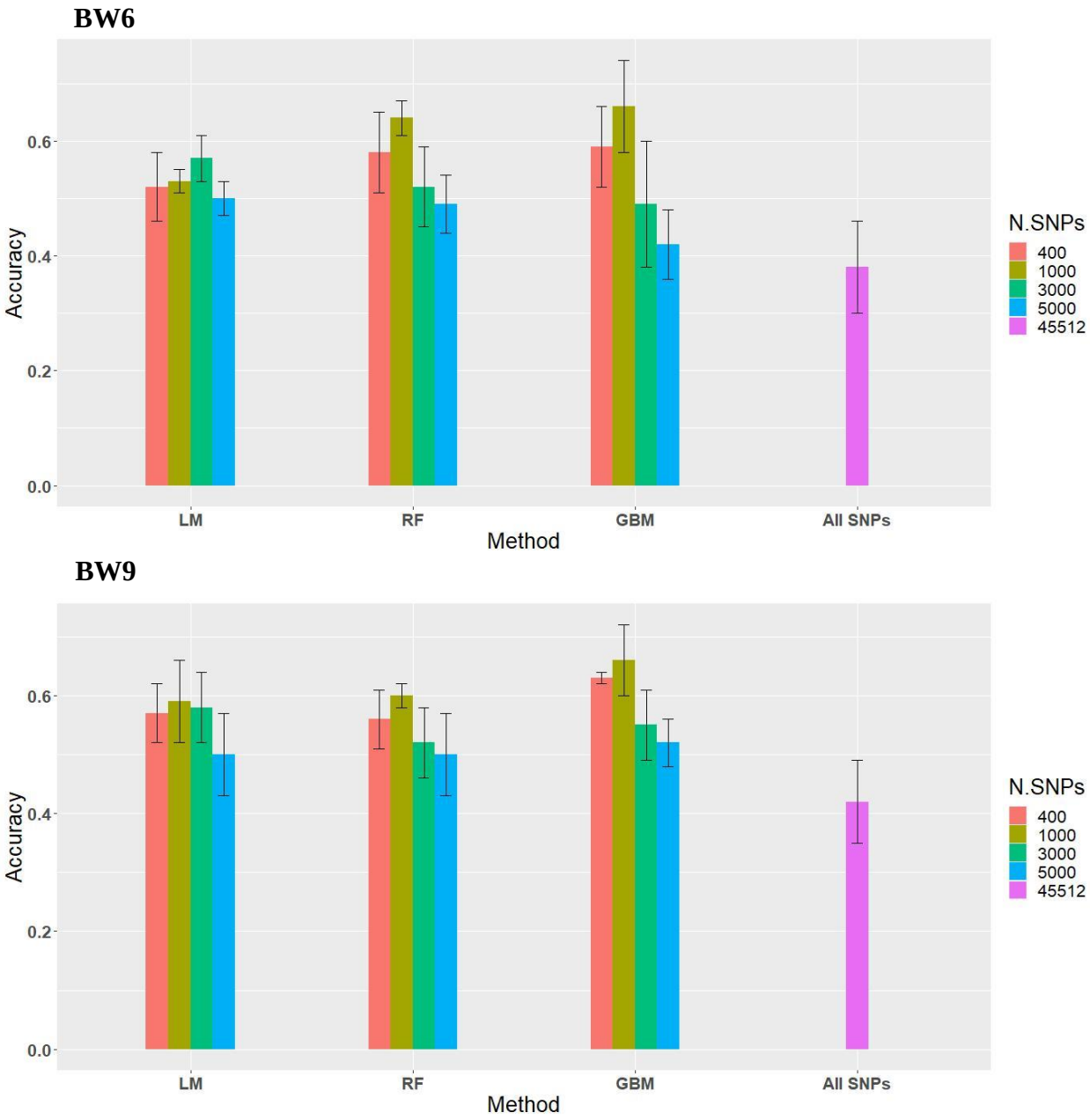


**Figure 4.** Accuracy of genomic prediction of different subsets of SNPs using a 5-fold cross-validation approach for BW6 and BW9 from LM, RF and GBM methods.

433

434

435

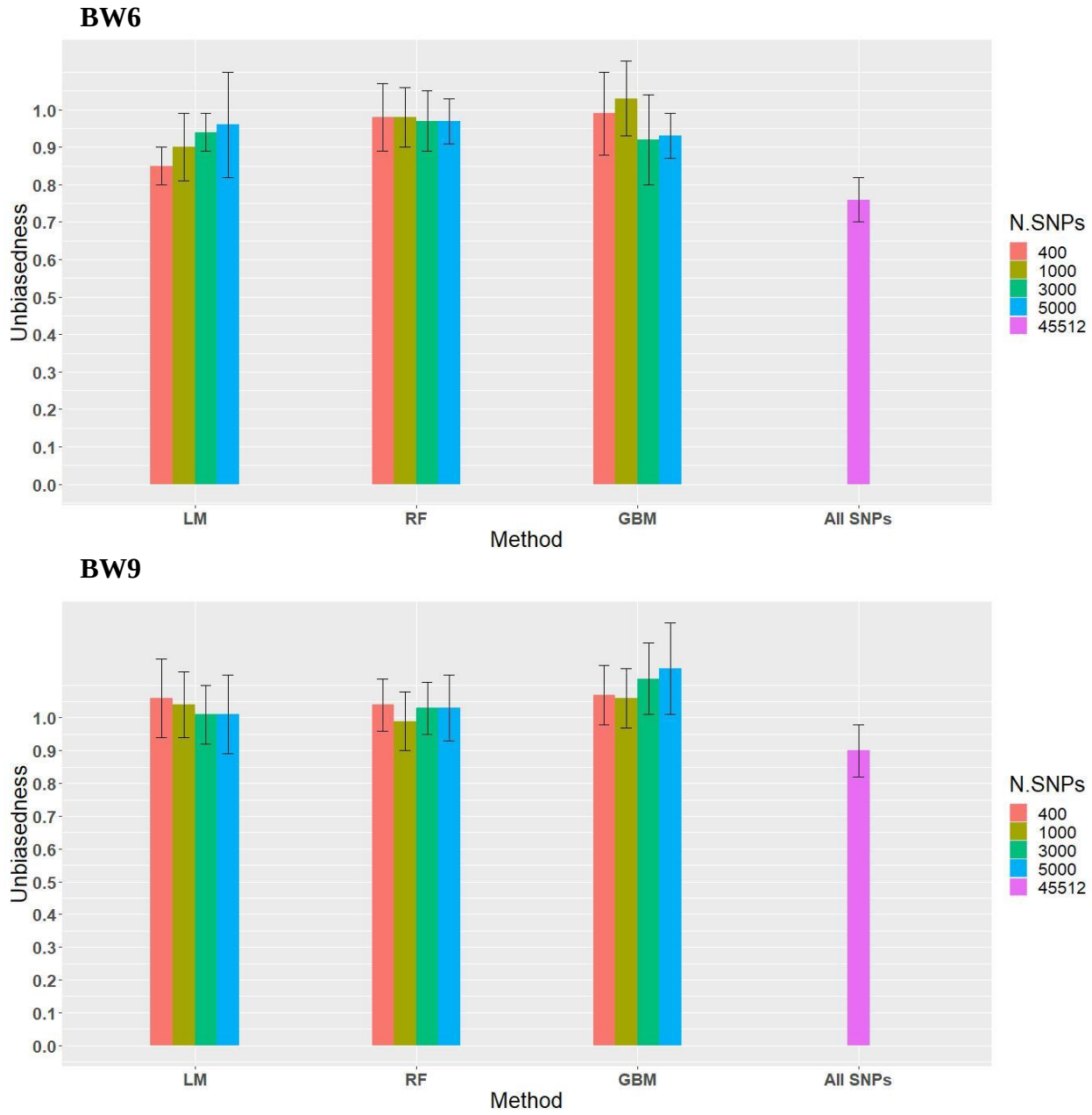


**Figure 5.** Unbiasedness of genomic prediction by different subsets of SNPs using a 5-fold cross-validation approach for BW6 and BW9 from LM, RF and GBM methods.

436

437

438

439

440

**Table1**

The Tukey HSD test for Accuracy of genomic prediction for body weights using pre-selected markers with the best subsets of SNPs.

| Method | BW6 | BW9 |
|---|---|---|
| All SNP | 0.38[a] | 0.42[a] |
| LM1000 | 0.53[b] | 0.59[b] |
| LM3000 | 0.57[c] | 0.58[bc] |
| RF400 | 0.58[c] | 0.56[c] |
| RF1000 | 0.64[d] | 0.60[b] |
| GBM400 | 0.59[c] | 0.62[d] |
| GBM1000 | 0.66[d] | 0.66[e] |

BW6 = 6 weeks body weight; BW9 = 9 weeks body weight; LM = Linear Model; RF = Random Forests; GBM = Gradient Boosting Machine.

# کاربرد الگوریتم‌های یادگیری ماشین در شناسایی SNP های تاثیرگذار برای ارزیابی ژنومی صفات رشد در جوجه های $F_2$

حسین بانی‌سعادت، رسول واعظ ترشیزی، قادر منافی‌آذر، علی اکبر مسعودی، علیرضا احسانی و صالح شاهین‌فر

## چکیده

استفاد از تراشه‌های چندشکلی‌های تک‌نوکلئوتیدی (SNP) با چگالی بالا به‌خصوص در کشورهای درحال توسعه بسیار هزینه‌بر هستند، اما روش‌هایی برای شناسایی SNPهای تاثیرگذار از این تراشه‌ها و طراحی تراشه‌های با چگالی کم برای ارزیابی ژنومی با هزینه کمتر توسعه یافته است. هدف از مطالعه حاضر، تعیین کارآیی الگوریتم‌های جنگل تصادفی (RF)، گرادیان بوستینگ (GBM) و مدل خطی (LM) در شناسایی زیرمجموعه‌های SNPها از یک تراشه 60K برای پیش‌بینی ارزش‌های اصلاحی (GEBVs) وزن بدن در سن 6 (BW6) و 9 (BW9) هفتگی جوجه‌های گوشتی و مقایسه GEBVs پیش بینی شده از زیر مجموعه‌ها با کل SNPهای تراشه 60K است. داده‌های 312 جوجه $F_2$ جمع‌آوری شده با تراشه 60K ایلومینا تعیین ژنوتیپ شدند. پس از اعمال کنترل کیفیت، 45512 SNPهای باقیمانده براساس مقادیر p (p-values)، افزایش درصد خطای میانگین (increase in mean square error percentage) و تأثیر نسبی (relative influence) به‌دست‌آمده به ترتیب از روش‌های LM، RF و GBM رتبه‌بندی شدند. سپس زیرمجموعه‌هایی از 400، 1000، 3000 و 5000 SNP برتر به دست آمده از هر روش برای ایجاد ماتریس‌های روابط ژنومی برای پیش‌بینی GEBVs با روش بهترین پیش بینی ناااریب خطی ژنومی استفاده شدند. نتایج نشان داد که دقت GEBVهای پیش‌بینی شده توسط RF و GBM به طور کلی بیشتر از مدل خطی بود. زیر مجموعه‌ای از 1000 SNP انتخاب شده توسط الگوریتم‌های RF و GBM در مقایسه با کل SNPها، دقت GEBVها را به ترتیب از 0/38 به 0/64 و 0/66 برای BW6 و از 0/42 به 0/60 و 0/66 برای BW9 افزایش داد. یافته‌های مطالعه حاضر نشان داد که روش‌های یادگیری ماشین، به‌ویژه GBM، می‌توانند بهتر از روش خطی معمولی در انتخاب SNPهای مهم عمل کنند و دقت پیش‌بینی ژنومی را در جوجه‌های گوشتی افزایش دهند.