

1 **ACCEPTED ARTICLE:**

2
3 **Comparing machine learning algorithms and linear model for detecting significant SNPs**
4 **for genomic evaluation of growth traits in F2 chickens**

5
6 **Hossein Bani Saadat^a, Rasoul Vaez Torshizi^{a*}, Ghader Manafiazar^b, Ali Akbar Masoudi^a, Alireza**
7 **Ehsani^a, Saleh Shahinfar^c**

8 ^a Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran

9 ^b Department of Animal Science and Aquaculture, Dalhousie University, Truro, NS, Canada.

10 ^c Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, Bundoora, Victoria 3083, Australia

11 ***Corresponding author:**

12 Rasoul Vaez Torshizi,

13 Department of Animal Science, Tarbiat Modares

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

15 rasoult@modares.ac.ir

16
17 **ABSTRACT**

18 High-density single nucleotide polymorphisms (SNP) panels are expensive, especially in
19 developing countries, but methods have been developed to detect critical SNPs from these panels
20 and design low-density chips for genomic evaluation at lower cost. This study aimed to determine
21 the efficiency of random forest (RF) and gradient boosting machine (GBM) algorithms, and Linear
22 Model (LM) in identification of SNPs subsets to predict genomic estimated breeding values
23 (GEBVs) for body weights at 6 (BW6) and 9 (BW9) weeks in broiler chickens and compare the
24 predicted GEBVs with those obtained by the 60k SNP panel. The data were collected on 312 F₂
25 chickens that genotyped with 60K Illumina SNP BeadChip. After applying quality control, the
26 remaining 45,512 SNPs were ranked based on p-values, mean square error percentage, and relative
27 influence, obtained by LM, RF and GBM methods, respectively. Then subsets of top 400, 1000,
28 3000 and 5000 SNPs, selected by each method, employed to construct genomic relationship
29 matrices for the prediction of GEBVs with genomic best linear unbiased prediction model. Results

30 indicated that predicted accuracies by RF and GBM were generally higher than LM. A Subset of
31 1000 SNPs selected by RF and GBM algorithms compared to the total SNPs increased accuracy
32 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
33 findings of the present study provide that machine learning methods, especially GBM, can perform
34 better than LM in selecting important SNPs and increase the accuracy of genomic prediction in
35 broiler chickens.

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

37

38 INTRODUCTION

39 Single nucleotide polymorphisms (SNPs) have been widely utilized in biological research, cancer
40 research, parentage testing, mapping of quantitative trait loci, and evaluation of genomic selection
41 due to their effectiveness as genetic markers. High-density (HD) SNP panels are now accessible
42 for many species due to advancements in high-throughput sequencing technology (Unterseer et al.,
43 2014). One of the important factors in using high-density SNPs is the cost, which is a big limiting
44 factor in utilizing it, especially in developing countries (Mrode et al., 2018). High-density SNP
45 panels used for genomic evaluations have a large number of SNPs that have little to no effect on
46 the traits and could decrease prediction accuracy (Ye et al., 2019). Therefore, various strategies
47 have been performed to select SNPs with large effect from high-density SNP chips, such as
48 selecting SNP evenly spaced across the genome (Habier et al., 2009) and based on allelic frequency
49 (Abdollahi et al., 2014).

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

61 Machine learning (ML) techniques have been used in GWASs (Mokry et al., 2013). In the context
62 of genome-enabled prediction of phenotypes, ML classification procedure was used by Long et al.
63 (2007) in selection of SNPs for prediction of mortality traits in poultry. Random Forest (RF)
64 (Breiman, 2001) has been applied to GWASs to identify SNP associated with phenotypes and to
65 map QTL on the genomic regions (Minozzi et al., 2014). Gradient Boosting Machine (GBM)
66 (Friedman, 2001) is another popular method of ML algorithm that has gained attention recently.
67 Piles et al. (2021) showed that compared to parametric methods, the best prediction quality in terms
68 of accuracy and stability was obtained with the GBM method for selecting SNPs in order to create
69 low-density SNP chips. The RF and GBM algorithms are suitable alternative to other methods used
70 for genomic evaluations at the expense of lower interpretability of results (González-Recio et al.,
71 2010) and are the most appealing alternatives to analyze complex traits using dense genomic
72 markers information (González-Recio and Forni, 2011).

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

84 85 MATERIALS AND METHODS

86 Experimental population, phenotypic and genotypic data

87 A population of F₂ crosses between the fast-growing Arian line (AA) and the slow-growing Urmia
88 Iranian indigenous chickens (NN) was used in this study. The F₁ birds were generated from the
89 mating of AA ♂ × NN ♀ and NN ♂ × AA ♀ birds and reared for 12 weeks in poultry research
90 farm of Tarbiat Modares University, Tehran, Iran. Then F₁ males from each reciprocal cross were
91 mated each to 4–8 females from other families and F₂ chickens were produced. Chickens of F₂
92 generation were raised individually in cages equipped with water nipples and feeders for 12 weeks

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

99 Genomic DNA was extracted from 312 blood samples using salting out method and stored at -
100 20°C. After extraction, spectrophotometry and agarose gel electrophoresis methods were used to
101 determine the quantity and quality of DNA. These DNA samples were genotyped with the Illumina
102 Chicken 60K SNP BeadChip, in cooperation with Cobb-Vantress Inc., and the Aarhus University,
103 Denmark. Quality control steps were applied to the original data with PLINK 1.9 software (Purcell
104 et al., 2007). SNPs with call rate of <95%, minor allele frequency of <5%, a Hardy–Weinberg
105 equilibrium test p-value <1× 10⁻⁶ were deleted (Emrani et al., 2017). After quality control, 45512
106 of SNPs for twenty-eight autosome chromosomes and 300 birds remained for final analysis.

107

108 **Methods for selecting markers**

109 The linear model for conventional GWAS was as follows:

110

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

111 where \mathbf{y} is the vector of corrected phenotypic values for BW6 and BW9, $\mathbf{1}$ is an n-vector of ones,
112 $\boldsymbol{\mu}$ is the population mean, \mathbf{q} is the effect of the marker in the model, which is treated as a fixed
113 regression of observation on genotype, \mathbf{Z} is a vector containing genotypes of the marker with 0, 1
114 and 2 for A₁A₁, A₁A₂ and A₂A₂, respectively, and \mathbf{e} is a vector of random residual effects, assuming
115 $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance and \mathbf{I} is the identity matrix. The genetic association
116 tests were conducted using the '--Linear' command in PLINK v1.9 (Purcell et al., 2007). The SNPs
117 were selected based on the p-values from GWAS results.

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

124 permuted (shuffled) and prediction error of OOB samples were calculated again. The mean square
125 error percentage (MSEP) difference between permuted and non-permuted samples (averaged over
126 all the trees in the forest) indicated the importance or predictive ability of that particular predictor.
127 The ‘randomForest’ package was used to perform this analysis in R software (Breiman, 2013).
128 In the GBM algorithm, the basic functions are weak learners such as a decision trees. The purpose
129 of the Boosting algorithm is to enhance ensemble of weak learners into a strong learner. In this
130 method, a basic learner such as a decision trees are added sequentially to the residuals of the
131 previous tree, and it is expected that by focusing on the incorrectly predicted data in the previous
132 tree, error rate in the next tree will be lessened and as long as the error rate is decreasing, the
133 boosting algorithm will continue (Friedman, 2001). In the present study, important markers in the
134 GBM method are identified by relative influence (RI), which is the average of reduction in MSEP
135 over all the trees when that particular SNP to split the data (Friedman, 2001). The ‘gbm’ package
136 was used to perform this algorithm in R software (Greenwell et al., 2019). For GBM and RF
137 methods, hyper-parameters tuning performed via nested grid search within a 3-fold cross-
138 validation on the 75 percent randomly selected subset of the data.

139

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

141 All SNPs were ranked from the most important to the least important SNP by criteria values of RF
142 (increase in MSEP), GBM (RI), and LM (p-value) using ‘dplyr’ package implemented in R
143 (Wickham et al., 2023). For the 5000 number of important SNPs, obtained from LM, RF and GBM,
144 venn diagrams were drawn by the ‘VennDiagram’ package (Chen and Boutros, 2011). Top 400,
145 1,000, 3000, and 5000 SNPs with the above-mentioned criteria were used to create genomic
146 relationship matrices.

147

148 **Genomic estimated breeding value**

149 Genomic estimated breeding values (GEBV) were derived using genomic best linear unbiased
150 prediction (GBLUP) model. The statistical model of GBLUP is written as follows (Gianola et al.,
151 2006):

152

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

153 where \mathbf{y} is an n-vector of corrected phenotypes, $\mathbf{1}$ is an n-vector of ones, $\boldsymbol{\mu}$ is the population mean,
154 \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

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

164 165 **Cross-validation for the accuracy of genomic breeding values**

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

$$168 \text{ Accuracy} = \frac{r_{\text{GEBV,phen}}}{\sqrt{h^2}}$$

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

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

$$172 b_{\text{GEBV,phen}} = r_{\text{GEBV,phen}}(S_{\text{phen}}/S_{\text{GEBV}})$$

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

180 181 **RESULTS AND DISCUSSION**

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

185 the importance of SNPs changes from positive to negative values. The highest positive value in RF
186 indicates an increase in the MSEP when the SNP is randomly permuted compared to the prediction
187 error before SNP permutation. In this model, 47%, 7%, and 46% of SNPs for BW6 and 47%, 9%
188 and 44% of SNPs for BW9 had positive, zero, and negative effects, respectively. About 5% of
189 SNPs for BW6 and 2.8% for BW9 (5000 pre-selected SNPs) had a MSEP increase more than 0.2,
190 respectively. In the GBM method, 26% and 16% of SNPs had larger than zero effect for BW6 and
191 BW9, respectively. In 5000 pre-selected SNPs with GBM method, none of the SNPs had a zero
192 RI, however, 65.3% of SNPs for BW6 had a RI less than one and close to zero. For BW9, the
193 amount of RI for last SNP of the 5000 pre-selected SNPs was 58.10%, and 52.52% of SNPs had a
194 RI less than 1000. Based on this method, about 3.62% of SNPs for BW6 and 5.06% for BW9
195 (5000 pre-selected SNPs) had a RI more than 10000, respectively.

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

209 The highest estimates of heritability for BW6 with pre-selected SNPs by LM, RF, and GBM
210 methods were, 0.42, 0.39 and 0.48, respectively, in subsets of 5000 SNPs (in LM) and 1000 SNPs
211 (in RF and GBM). For BW9, the highest heritability was 0.43, 0.46 and 0.58 in subsets of 5000
212 SNPs (in LM), 1000 SNPs (in RF) and 3000 SNPs (in GBM), respectively. In comparison to all
213 45,512 SNPs, the heritability estimates were increased from 0.28 to 0.35 and from 0.30 to 0.46
214 through preselection of 5000 SNPs using LM model for BW6 and BW9, respectively. By using
215 1000 pre-selected SNPs, the increases in the heritability estimates were ranged from 0.28 to 0.37

216 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
217 for BW9 with GBM model. Ren et al. (2022) indicated that high-density SNP data provide more
218 information for genomic evaluation compared to medium-density SNP data but they do not confer
219 any advantage for heritability estimation. Literature studies reported that the estimates of genomic
220 heritability were very sensitive to differences in LD between SNPs, suggesting that genomic
221 heritability is overestimated in region with high LD and underestimated in region with low LD
222 (Speed et al., 2012). The stronger LD of the remaining SNPs and the removal of the imperfect LD
223 between the causal mutations may improve the genomic relationships between individuals and
224 increase the heritability of the trait (Abdollahi et al., 2014; Ye et al., 2019). Abdollahi et al. (2014)
225 estimated genomic heritability for body weight at 6 weeks in broilers chickens using the genomic
226 relationship matrix consisting of all SNPs and a subset of selected SNPs and reported that the
227 genomic heritability with selected SNP (0.59) is expected to be overestimated in comparison to all
228 SNPs (0.30), however, the subsets of SNPs could increase the GEBV accuracy. The increase in the
229 accuracy of GEBV has been reported by Luo et al, (2021) who proposed a strategy for genomic
230 selection in aquaculture using a subset of markers selected by the p-value of GWAS and indicated
231 that the prediction accuracy of a subset of top SNPs was higher than using total SNPs. Li et al.
232 (2018) reported that ML methods can consider complex and nonlinear relationships. Therefore,
233 they can produce a smaller error variance and increase genetic variance and heritability. These
234 authors estimated the heritability of a subset of 3000 SNPs with GBM method to be higher than all
235 of 38082 SNPs for body weight in Brahman cattle.

236 Figure 4 and Figure 5 show the mean accuracy and regression coefficient (as a measurement of
237 unbiasedness) of genomic breeding value for BW6 and BW9 traits using SNP subsets in a 5-fold
238 cross-validation scheme, respectively. Accuracy of genomic prediction for BW6 and BW9 using
239 all SNPs was estimated to be 0.38 and 0.42, respectively, which was lower than the genomic
240 prediction accuracy obtained from the subsets of selected SNPs (400, 1000, 3000 and 5000 selected
241 with three methods). Average accuracy of genomic breeding value (\pm standard error) with top 400,
242 1000, 3000, and 5000 SNP subsets selected by LM, RF and GBM methods were estimated to be
243 0.56 (\pm 0.02), 0.61 (\pm 0.04), 0.52 (\pm 0.02) and 0.47 (\pm 0.02) for BW6, and 0.58 (\pm 0.02), 0.61
244 (\pm 0.02), 0.55 (\pm 0.02) and 0.51 (\pm 0.01) for BW9, respectively. Mean regression coefficient of
245 genomic prediction on phenotype using total SNPs for BW6 and BW9 were estimated to be 0.76
246 and 0.90, respectively. With top 400, 1000, 3000, and 5000 SNP subsets selected by three methods,

247 mean regression coefficient (\pm standard error) were 0.94 (\pm 0.05), 0.97 (\pm 0.04), 0.94 (\pm 0.02) and
248 0.95 (\pm 0.01) for BW6, and 1.06 (\pm 0.01), 1.03 (\pm 0.02), 1.05 (\pm 0.04) and 1.06 (\pm 0.05) for BW9,
249 respectively. The best average accuracy of genomic breeding value and regression coefficient
250 provided by 1000 **SNP subset was** 0.61 (\pm 0.04) and 0.97 (\pm 0.04) for BW6 and 0.61 (\pm 0.02) and
251 1.03 (\pm 0.02) for BW9, respectively. **In the study of Liu et al. (2020), the highest accuracy of**
252 **genomic breeding value by a subset of 817 SNPs selected from high-density SNP panels was** 0.60
253 **for body weight at the age of 12 weeks and by a subset of 354 SNPs was** 0.45 for feed conversion
254 **ratio in broiler. Furthermore, several studies indicated a direct relationship between effective**
255 **population size and the accuracy of GEBVs. The significant impact of smaller effective population**
256 **size on the prediction accuracy of GBLUP has been revealed by Daetwyler et al. (2010), which is**
257 **a reflection of strong linkage disequilibrium between variants due to close genetic relatedness**
258 **between individuals (Jang et al., 2023; Calus et al., 2008).**
259 **Significant differences between genomic prediction accuracy of the best subsets of SNPs (which**
260 **had the highest increase in genomic prediction accuracy) with each other and all SNPs are**
261 **presented in Table 1.** The results showed that 1000 SNPs selected by ML algorithms was the best
262 pre-selected SNPs for estimating genomic breeding value in broiler chickens for body weight traits
263 in the present study. In BW6, there was no significant difference between RF and GBM algorithms
264 in the best subsets (1000 SNPs) of selected SNPs, and they were superior to linear model with the
265 best subset (3000 SNPs). However, in BW9, GBM was superior to other methods. **These findings**
266 **are consistent with the results of Kriaridou et al. (2020), who used different subsets of SNPs in four**
267 **aquaculture datasets, ranging from 100 to 9000 SNPs, and observed that SNP densities between**
268 **1000 and 2000 SNPs had a very similar accuracy of genomic evaluation to high-density**
269 **genotyping.** Ye et al. (2019) used selected markers from whole-genome sequencing data based on
270 the p-value obtained from GWAS and showed that the use of pre-selected markers for most traits
271 did not increase the genomic prediction accuracy in broilers and even increased the bias. One of
272 the possible reasons is the difficulty of discovering causative variants using GWAS due to the large
273 number of variants (600k) and high LD between variants. On the contrary, Li et al. (2018) indicated
274 an increase in the accuracy of genomic prediction by selecting a subset of significant SNPs from
275 high-density SNP panel (651,253) using RF method. **One of the advantages of ML method is its**
276 **ability to analyze data with a high dimension, however, factors such as linkage disequilibrium and**
277 **minor allele frequency can affect the performance of ML algorithms for selecting important**

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

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

302 303 **CONCLUSIONS**

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

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

314

315 REFERENCES

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

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

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

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

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

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

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

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

334 Druet, T., Macleod, I.M. and Hayes, B.J. 2014. Toward genomic prediction from whole-genome sequence
335 data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, **112** (1):
336 39-47.

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

339 Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Ann. statist.*, **29**
340 (5).1189-1232.

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

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

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

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

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

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

354 **Kriaridou, C., Tsairidou, S., Houston, R.D. and Robledo, D. 2020. Genomic prediction using low density**
355 **marker panels in aquaculture: performance across species, traits, and genotyping platforms. *Front. Genet.*,**
356 **11:124.**

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

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

362 Liu, T., Luo, C., Ma, J., Wang, Y., Shu, D., Su, G. and Qu, H. 2020. High-throughput sequencing with the
363 preselection of markers is a good alternative to SNP chips for genomic prediction in broilers. *Front. Genet.*,
364 **11**: 108.

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

368 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
369 of genomic evaluation using imputed high-density genotypes for carcass traits in commercial Hanwoo
370 population. *Livest. Sci.*, **241**:104256.

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

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

376 Minozzi, G., Pedretti, A., Biffani, S., Nicolazzi, E.L. and Stella, A. 2014. Genome wide association analysis
377 of the 16th QTL- MAS Workshop dataset using the Random Forest machine learning approach. *BMC proc.*,
378 **5**: 1-6.

379 Mokry, F.B., Higa, R.H., de Alvarenga Mudadu, M., Oliveira de Lima, A., Meirelles, S.L.C., Barbosa da
380 Silva, M.V.G., Cardoso, F.F., Morgado de Oliveira, M., Urbinati, I., Meo Niciura, S.C. and Tullio, R.R.
381 2013. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest
382 approach. *BMC genet.*, **14** (1): 1-11.

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

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

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

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

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

396 Schiavo, G., Bertolini, F., Galimberti, G., Bovo, S., Dall'Olio, S., Nanni Costa, L., Gallo, M., Fontanesi, L.
397 2020. A machine learning approach for the identification of population-informative markers from high-
398 throughput genotyping data: application to several pig breeds. *Anim.*, **14** (2): 223-232.

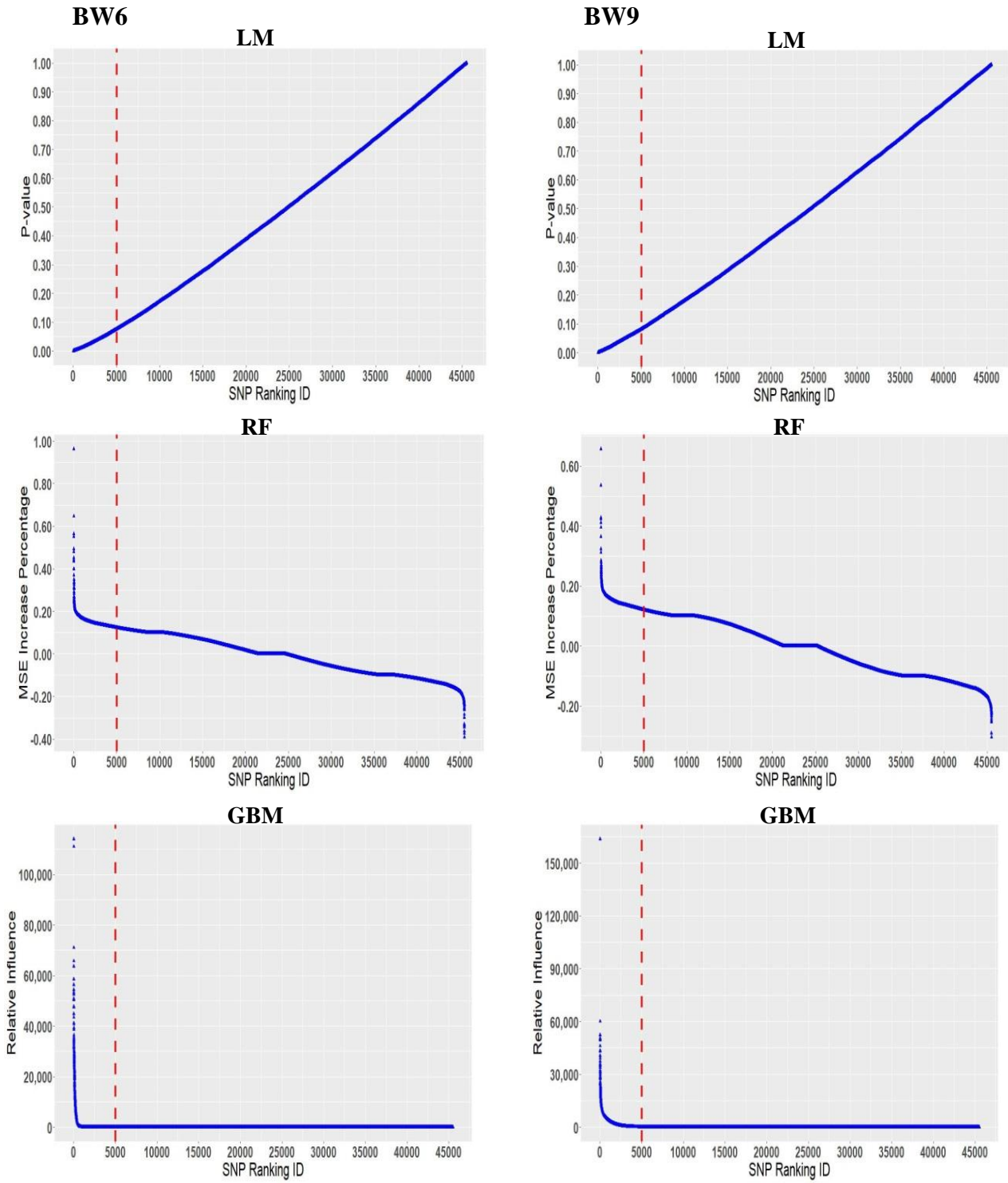
399 Seo, D., Cho, S., Manjula, P., Choi, N., Kim, Y.K., Koh, Y.J., Lee, S.H., Kim, H.Y., Lee, J.H. 2021.
400 Identification of Target Chicken Populations by Machine Learning Models Using the Minimum Number of
401 SNPs. *Anim.*, **11** (1): 241.

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

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

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

- 409 Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E. and Visscher, P.M. 2013. Pitfalls of predicting
410 complex traits from SNPs. *Nat. Rev. Genet.*, **14** (7): 507–515.
- 411 Ye, S., Gao, N., Zheng, R., Chen, Z., Teng, J., Yuan, X., Zhang, H., Chen, Z., Zhang, X., Li, J. and Zhang,
412 Z. 2019. Strategies for obtaining and pruning imputed whole-genome sequence data for genomic prediction.
413 *Front. Genet.*, **10**: 673.
- 414 Zhou, J. and Troyanskaya, O.G. 2015. Predicting effects of noncoding variants with deep learning–based
415 sequence model. *Nat. Methods.*, **12** (10): 931-934.



416

417

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

418



419

420

421

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.

422

423

424

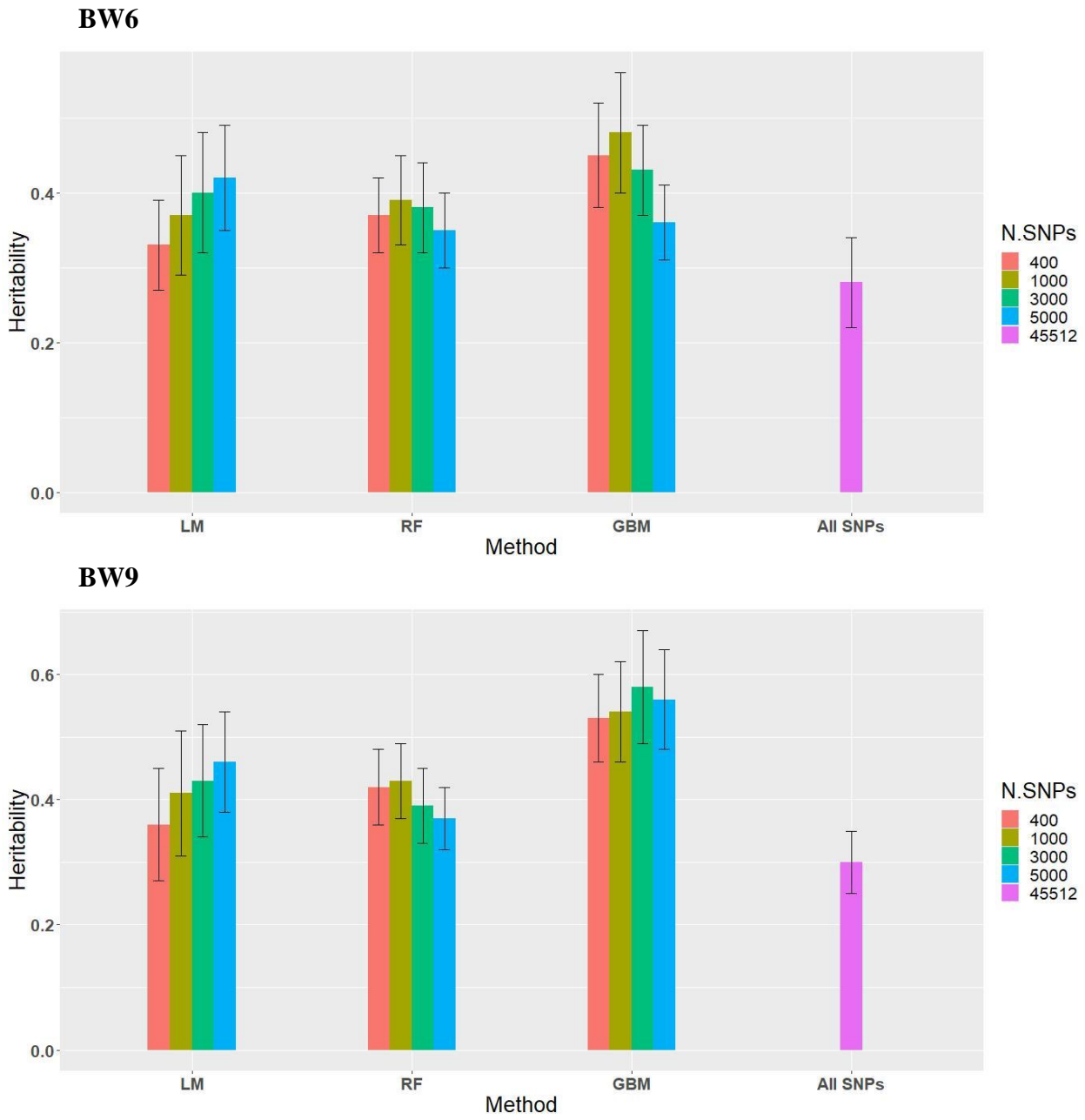
425

426

427

428

429



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

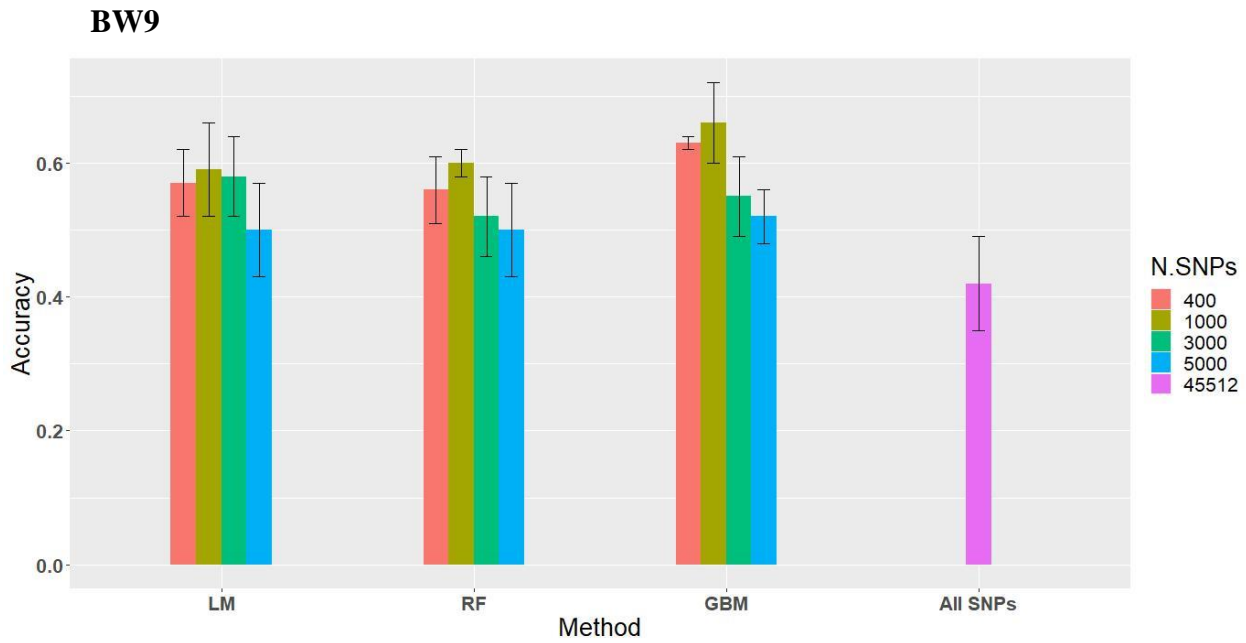
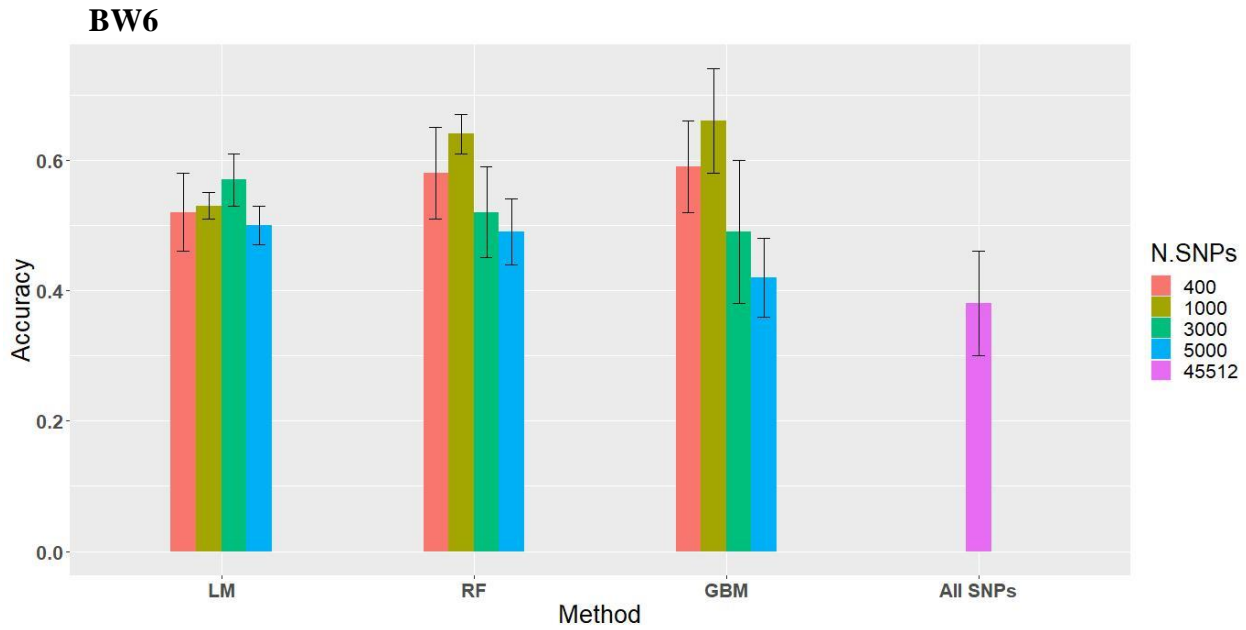


Figure 4. Accuracy of genomic prediction of different subsets of SNPs using a 5-fold cross-validation approach for BW6 and BW9 from LM, RF and GBM methods.

433

434

435

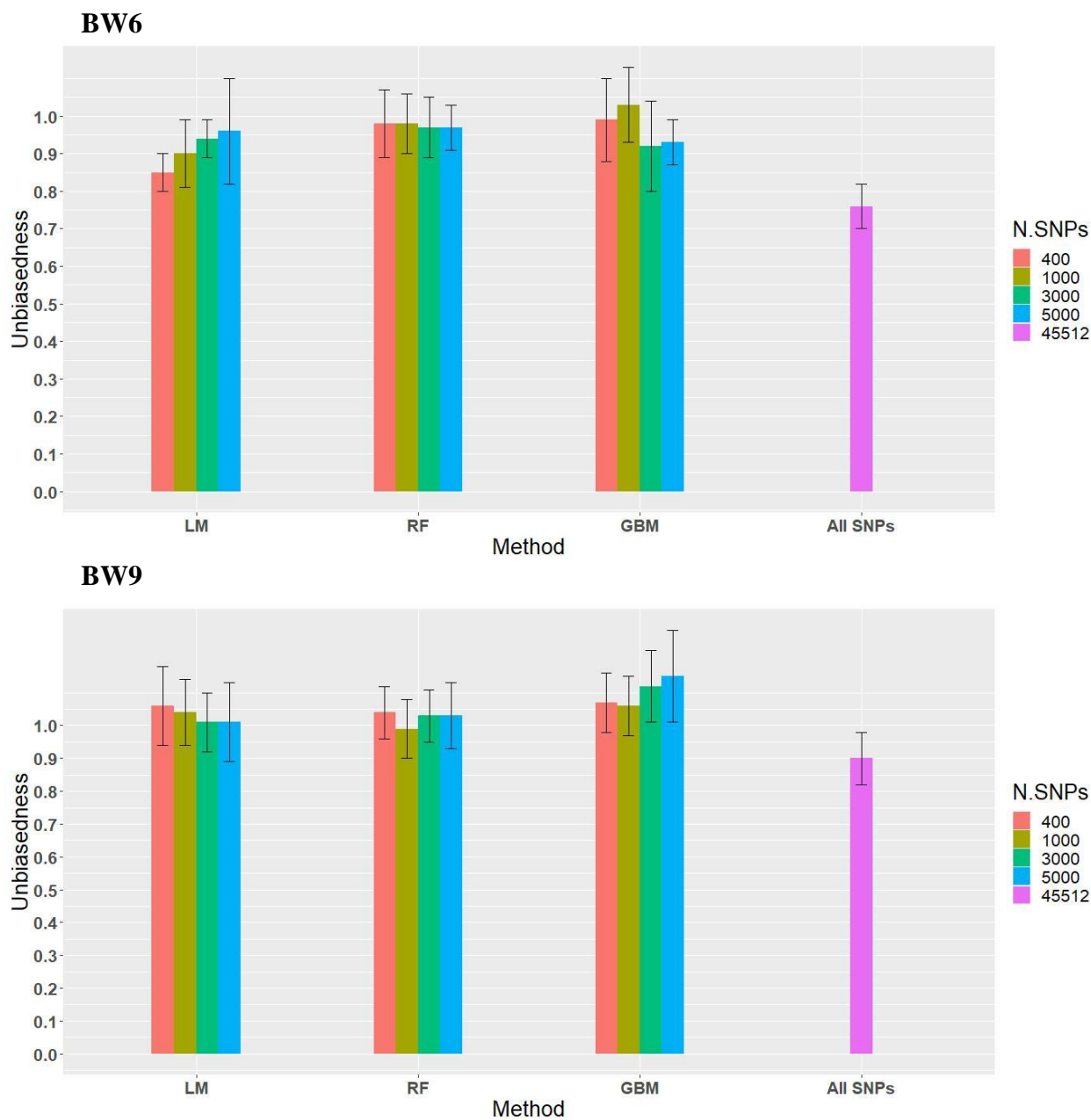


Figure 5. Unbiasedness of genomic prediction by different subsets of SNPs using a 5-fold cross-validation approach for BW6 and BW9 from LM, RF and GBM methods.

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.

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457 کاربرد الگوریتم‌های یادگیری ماشین در شناسایی SNP های تاثیرگذار برای ارزیابی ژنومی صفات رشد در جوجه های

458 F_2

459 حسین بانی سعادت، رسول واعظ ترشیزی، قادر منافی آذر، علی اکبر مسعودی، علیرضا احسانی و صالح شاهین فر

460

461 چکیده

462 استفاد از تراشه‌های چندشکلی‌های تک‌نوکلئوتیدی (SNP) با چگالی بالا به‌خصوص در کشورهای درحال توسعه بسیار هزینه‌بر هستند، اما

463 روش‌هایی برای شناسایی SNP های تاثیرگذار از این تراشه‌ها و طراحی تراشه‌های با چگالی کم برای ارزیابی ژنومی با هزینه کمتر توسعه یافته

464 است. هدف از مطالعه حاضر، تعیین کارایی الگوریتم‌های جنگل تصادفی (RF)، گرادیان بوستینگ (GBM) و مدل خطی (LM) در

465 شناسایی زیرمجموعه‌های SNP ها از یک تراشه 60K برای پیش‌بینی ارزش‌های اصلاحی (GEBVs) وزن بدن در سن 6 (BW6) و 9

466 (BW9) هفتگی جوجه‌های گوشتی و مقایسه GEBVs پیش‌بینی شده از زیر مجموعه‌ها با کل SNP های تراشه 60K است. داده‌های

467 312 جوجه F_2 جمع‌آوری شده با تراشه 60K ایلومینا تعیین ژنوتیپ شدند. پس از اعمال کنترل کیفیت، 45512 SNP های باقیمانده

468 براساس مقادیر p (p-values)، افزایش درصد خطای میانگین (increase in mean square error percentage) و تأثیر نسبی

469 (relative influence) به‌دست‌آمده به ترتیب از روش‌های LM، RF و GBM رتبه‌بندی شدند. سپس زیرمجموعه‌هایی از 400،

470 1000، 3000 و 5000 SNP برتر به دست آمده از هر روش برای ایجاد ماتریس‌های روابط ژنومی برای پیش‌بینی GEBVs با روش

471 بهترین پیش‌بینی ناریب خطی ژنومی استفاده شدند. نتایج نشان داد که دقت GEBV های پیش‌بینی شده توسط RF و GBM به طور کلی

472 بیشتر از مدل خطی بود. زیر مجموعه‌ای از 1000 SNP انتخاب شده توسط الگوریتم‌های RF و GBM در مقایسه با کل SNP ها، دقت

473 GEBV ها 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

474 حاضر نشان داد که روش‌های یادگیری ماشین، به‌ویژه GBM، می‌توانند بهتر از روش خطی معمولی در انتخاب SNP های مهم عمل کنند و

475 دقت پیش‌بینی ژنومی را در جوجه‌های گوشتی افزایش دهند.

476

477