Comparison of Bivariate Machine Learning and Linear Model For Genomic Prediction With Different Heritability, QTL and SNP Panel Scenarios

Sunday O Peters[1], Mahmut Sinecen[2], Kadir Kizilkaya[2] and Milt Thomas[3]

[1]Berry College, Mount Berry, GA, USA [2]Aydin Adnan Menderes University, Aydin, Turkey [3]Colorado State University, Fort Collins, CO, USA

INTRODUCTION

Several statistical and computational methods have been devised to predict the genetic value of individuals from the analysis of SNPs datasets Useche et al. (2001). There is an important statistical problem to use SNPs marker for estimating the genetic merits because the number of SNP effects is usually much larger than the number of observed phenotypes. However, genomic best linear unbiased prediction (GBLUP) method was developed by replacing pedigree-based relationship with genomic relationships estimated from SNPs marker information. So, the dimensions of the genetic effects in the model is reduced the number of individuals in the population, which is computationally more efficient (Misztal et al., 2009).

Artificial neural networks (ANN) provide an interesting alternative for marker-based genomic predictions of complex traits in animal and plant breeding. The knowledge of the nervous system inspired the use of ANN which were developed in the field of artificial intelligence. The idea of ANN has been used to define statistical models in the form of neuron diagrams, as shown in Figure 1. However, ANN are computationally costly, especially when applied to high-dimensional genomic data, for which the number of parameters to be estimated typically exceeds the number of available samples. Therefore, at least in animal breeding, the use of genomic relationship have made ANN computational feasible (Gianola et al., 2011).

This simulation study used actual SNP genotypes on the first chromosome of Brangus beef cattle to simulate 0.50 genetically correlated two traits with heritabilities of 0.25 and 0.50 determined either by 50, 100, 250 or 500 QTL and then aimed to compare the accuracies of genomic prediction from bivariate linear and artificial neural network with 1 to 10 neurons models based on $\mathbf{G}$ genomic relationship matrix.

MATERIAL and METHODS

Single Nucleotide Polymorphism Genotypes: Single nucleotide polymorphism (SNP) genotypes were successfully obtained from 719 Brangus heifers (3/8 Brahman-*Bos indicus* × 5/8 Angus-*Bos taurus*) registered with International Brangus Breeders Association using BovineSNP50 (Infinium BeadChip, Illumina, San Diego, CA; 53,692 SNP). Genotype call rates averaged 98.1 ± 0.001% for 53,692 SNP. Genotypes were in the Illumina A/B allele format and were used to compute a value at each locus coded as 0, 1, or 2, representing the number of B alleles. All 3361 SNP genotypes on the first chromosome were used for this simulation study (Peters et al., 2012)

Simulation of Additive Genetic Merits and Phenotypic Performances: Correlated two traits with heritability of 25% (T_1h^2=0.25) and 50% (T_2h^2=0.5), determined by 50, 100, 250, or 500 additive bi-allelic QTL was simulated using Illumina SNP genotypes from Brangus heifers as described below.

A random sample of N = 50, 100, 250, or 500 SNP genotypes were chosen from the observed 3361 SNP genotypes on the first chromosome to represent QTL. Each locus had an equal probability of being included, regardless of minor allele frequency. Each QTL was assigned a parametric substitution effect by sampling from a multivariate normal distribution with mean $\begin{bmatrix}0\\0\end{bmatrix}$ and variance-covariance matrix $\frac{\boldsymbol{\Psi}}{(N2\bar{p}\bar{q})}$ with $\bar{p}$ and $\bar{q}$ estimated from all SNP loci (Fernando et al., 2007) and $\boldsymbol{\Psi} = \begin{bmatrix}1.0 & 0.7\\0.7 & 2.0\end{bmatrix}$. Residual effects for each animal were obtained by sampling from a multivariate normal distribution with mean $\begin{bmatrix}0\\0\end{bmatrix}$ and variance-covariance matrix $\boldsymbol{R} = \begin{bmatrix}3.0 & 0.7\\0.7 & 2.0\end{bmatrix}$.

Additive genetic merit of each animal for two traits was obtained as the sum of the substitution effects for each QTL allele. The simulated phenotypic performances of each animal for two traits were obtained by adding its residual effects to its additive genetic merits as follows:

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \sum_{j=1}^{N} g_{ij}\boldsymbol{\beta}_j + \boldsymbol{e}_i$$

where $\boldsymbol{y}_i = \begin{bmatrix}y_1\\y_2\end{bmatrix}$ is a vector of the simulated phenotypic performances of animal i,$\boldsymbol{\mu} = \begin{bmatrix}\mu_1 = 5\\\mu_2 = 10\end{bmatrix}$ is a vector of population means for two traits, $\sum_{j=1}^{N} g_{ij}\boldsymbol{\beta}_j$ is a vector of the additive genetic merits obtained by summing the genotypic values at each locus over all N loci for two traits, where the genotypic values at locus j is the product of the 0, 1, 2 covariate (g_{ij}) for animal i, and the substitution $\boldsymbol{\beta}_j$ for that locus, and $\boldsymbol{e}_i$ is a vector of the random residual effects (Kizilkaya et al., 2010). Data sets were generated for each of 4 QTL scenarios (QTL50, QTL100, QTL250, or QTL500) representing N = 50, 100, 250, or 500 QTL.

Use of Marker Panels for Genomic Relationship (G) Matrix: Three sets of SNP panels were used for genomic relationship (**G**) matrix: only QTL genotypes (Panel1), all SNP markers, including the QTL (Panel2), and all SNP markers, excluding the QTL (Panel3). Following the approach of Habier et al. (2007) and vanRaden (2008), $\mathbf{G}$ was defined as $\mathbf{G} = \frac{\mathbf{MM}^{\mathrm{T}}}{2\sum_{i=1}^{s} p_i(1-p_i)}$ where $\mathbf{M} = \mathbf{W} - \mathbf{P}$, $\mathbf{W}$ is the (*n*×*s*) matrix of SNP genotype vectors for the *n* animals with the *s* SNPs code as 0, 1, 2 and $\mathbf{P}$ contains the allele frequencies multiplied by 2, p_i is the allele frequency of SNP marker *i*, and the sum is over all loci.

Genomic Prediction: Genomic predictions for T_1h^2=0.25 and T_2h^2=0.5 were carried out by bivariate Genomic Best Linear Unbiased Prediction (GBLUP) and bivariate Feed Forward MultiLayer Perceptron ANN-1-10 neurons with **G** matrix. The correlations between true genetic and predicted phenotypes from 10-fold cross validation for T_1h^2=0.25 and T_2h^2=0.5 were used to assess predictive ability of bivariate GBLUP and ANN-1-10 neurons based on 4 QTL scenarios with 3 Panels of SNP genotypes.

Genomic Best Linear Unbiased Prediction (GBLUP): GBLUP is the method combining genomic information into the method of best linear unbiased prediction by using genomic relationship matrix (**G**) and was introduced by Habier et al. (2007) and VanRaden (2008).

The underlying statistical model is

$$\mathbf{y} = \mathbf{X\mu} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypes of animals for traits T_1h^2=0.25 and T_2h^2=0.5, $\mathbf{X}$ is a matrix of 1, $\boldsymbol{\mu}$ is the overall mean for traits T_1h^2=0.25 and T_2h^2=0.5, $\mathbf{Z}$ is a design matrix allocating phenotypes to genetic values of animals, $\mathbf{g}$ is a vector of additive genetic effects of animals for T_1h^2=0.25 and T_2h^2=0.5 and $\mathbf{u} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{G} \otimes \boldsymbol{\Omega})$ $\mathbf{u}$ is assumed to be multivariate normal, with $\mathbf{G}$ the genomic relationship matrix and $\boldsymbol{\Omega}$ the additive genetic variance covariance matrix. $\mathbf{e} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I} \otimes \boldsymbol{R})$ is a vector of the normally distributed residuals for T_1h^2=0.25 and T_2h^2=0.5, where $\boldsymbol{R}$ is the residual variance covariance matrix.

Artificial Neural Network (ANN) with G Matrix: Feed Forward Multilayer Perceptron (FFMP) Artificial Neural Network (ANN) was applied with three layers: input layer (Genomic relationships values (g) from **G** matrix), one hidden layer (neuron numbers are increased 1 to 10 for finding best performance) and output layer (the simulated traits, T_1h^2=0.25 and T_2h^2=0.5) (Figure 1). In the analysis, FFMP-ANN parameters were tangent sigmoid transfer function and linear transfer function for output. The train algorithm was the scaled conjugate gradient algorithm.

FFMP-ANN-1-10 was used to predict the simulated traits, T_1h^2=0.25 or T_2h^2=0.5 of Brangus. In the training phase of FFMP-ANN-1-10, genomic relationship values (g) were linearly combined with a vector of weights. The resulting linear score is then transformed using an activation function to produce the output of the single hidden neuron.

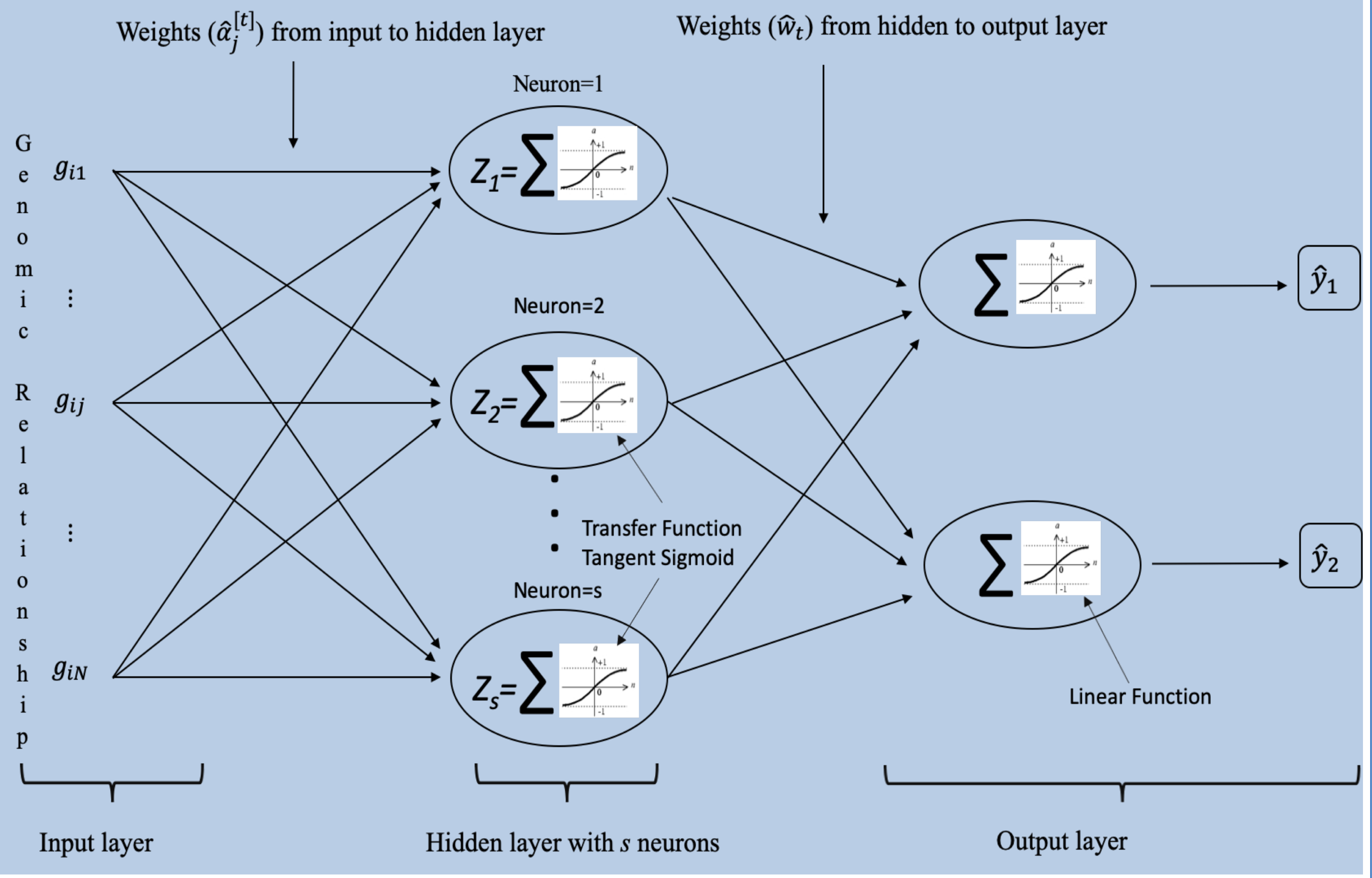

Figure 1. Feed Forward Multilayer Perceptron Artificial Neural Network (ANN)

RESULTS AND DISCUSSION

The predictive performance of GBLUP and ANN-1-10 models varied for 4-QTL scenarios and marker panels. However, GBLUP resulted in 42% higher predictive performance than ANN-1-10 model across QTL and Panel scenarios and the predictive performance of GBLUP was also higher for (46%) low heritability (T_1h^2=0.25) trait than (38%) high heritability (T2h2=0.5) trait. In ANN-1-10 model, the number of neurons resulted in the varied correlations among 10-fold cross-validation datasets. Sinecen (2019) determined that GBLUP model resulted in higher correlation than Bayesian Regularization of Neural Network with different number of neurons among 10-fold cross-validation datasets.

The effect of heritabilities of 25% and 50% on the predictive performance of GBLUP and ANN-1-10 model for 4-QTL scenarios and marker panels (through $\mathbf{G}$ matrix) is shown in Table 1 for testing dataset. As expected, increase in heritability was associated with increased predictive performance of GBLUP and ANN-1-10 models for 4-QTL scenarios and marker panels and this trend was similar for training and testing datasets. The average percentage changes (increases) for 4-QTL scenarios in Panels were 48% for GBLUP and 58% for ANN-1-10 model. The effect of the number of QTL and heritability (genetic architecture) of trait on genomic prediction in the comparison of different genomic models were studied by Daetwyler et al. (2010) and Zhang et al. (2010) and they found that models for genomic prediction were sensitive to the number of QTL and heritability and decreasing heritability resulted in the decrease in the predictive performance of genomic models. Habier et al. (2011) also determined significant trends for the number of SNP depended on heritability, number of QTL and the distribution of QTL effects. Zhang et al. (2019) studied the factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations and reported that traits with higher heritability have higher prediction accuracy.

Our results indicated that linear genome-enabled prediction (GBLUP) model outperform neural network models with one to ten neurons based on genomic relationship created from different marker panels. The higher differences in predicting future phenotypes for the traits of T_1h^2=0.25 and T_2h^2=0.5 different heritabilities in validation datasets is acknowledged.

Table 1. Correlations from 10-fold cross validation for traits were used to assess predictive ability of bivariate linear (GBLUP) and artificial neural network models based on 4 QTL scenarios with 3 Panels of SNP panels.

Scenarios		Neural Network		GBLUP	
Panel	**QTL**	**r_y1,yhat1**	**r_y2,yhat2**	**r_y1,yhat1**	**r_y2,yhat2**
QTL50	QTL50	0.425	0.562	0.432	0.642
QTL100	QTL100	0.340	0.550	0.433	0.628
QTL250	QTL250	0.279	0.474	0.419	0.645
QTL500	QTL500	0.374	0.429	0.444	0.585
QTLSNP50	QTL50	0.297	0.399	0.347	0.551
QTLSNP100	QTL100	0.271	0.420	0.378	0.565
QTLSNP250	QTL250	0.230	0.404	0.384	0.604
QTLSNP500	QTL500	0.224	0.421	0.416	0.550
SNP50	QTL50	0.189	0.375	0.341	0.543
SNP100	QTL100	0.205	0.413	0.371	0.557
SNP250	QTL250	0.232	0.388	0.377	0.591
SNP500	QTL500	0.320	0.308	0.406	0.536

REFERENCES

Daetwyler, HD, R. Pong-Wong, B. Villanueva, and J. A. Woolliams, “The Impact of Genetic Architecture on Genome-Wide Evaluation Methods,” *Genetics*, vol. 185, no. 3, pp. 1021–1031, Jul. 2010, doi: 10.1534/genetics.110.116855.

Fernando, RL., Habier, D., Sticker, C.., Dekkers, JCM. and Totir, LR. “Genomic selection”. Acta Agriculturae Scand. Section A, 57:192–195, 2007.

Gianola, D., H. Okut, K. A. Weigel, and G. J. Rosa, ‘‘Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat,’’ BMC Genetics, vol. 12, no. 1, p. 87, 2011.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, ‘‘The impact of genetic relationship information on genome-assisted breeding values,’’ Genetics, vol. 177, no. 4, pp. 2389–2397, 2007.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, “Extension of the bayesian alphabet for genomic selection,” *BMC Bioinformatics*, vol. 12, no. 1, p. 186, Dec. 2011, doi: 10.1186/1471-2105-12-186.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick, ‘‘Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes,’’ J. Animal Sci., vol. 88, no. 2, pp. 544–551, 2010.

Misztal, I., A. Legarra, and I. Aguilar, ‘‘Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information,’’ J. Dairy Sci., vol. 92, pp. 4648–4655, Sep. 2009.

Peters, SO., Kizilkaya, K., Garrick, DJ., Fernando, RL., Reecy, JM., Weaber, RL., Silver, GA. and Thomas, MG. “Bayesian quantitative loci inference from whole genome analyses of growth and yearling ultrasound measures of carcass traits in Brangus heifers”. J. Anim. Sci., 90: 3398–3409, 2012.

Sinecen, M. “Comparison of Genomic Best Linear Unbiased Prediction and Bayesian Regularization Neural Networks for Genomic Selection,” *IEEE Access*, vol. 7, pp. 79199–79210, 2019, doi: 10.1109/ACCESS.2019.2922006.

Useche, F.J., G. Gao, M. Hanafey, and A. Rafalski, ‘‘High-throughput identification, database storage and analysis of SNPs in EST sequences,’’ Genome Inform., vol. 12, pp. 194–203, Feb. 2001.

vanRaden, P. M. ‘‘Efficient methods to compute genomic predictions,’’ J. Dairy Sci., vol. 91, no. 11, pp. 4414–4423, 2008.

Zhang, H., L. Yin, M. Wang, X. Yuan, and X. Liu, “Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations,” *Front. Genet.*, 2019, doi: 10.3389/fgene.2019.00189.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning, and Q. Zhang, “Best Linear Unbiased Prediction of Genomic Breeding Values Using a Trait-Specific Marker-Derived Relationship Matrix,” *PLoS One*, vol. 5, no. 9, p. e12648, Sep. 2010, doi: 10.1371/journal.pone.0012648.