Genomic Prediction With Different Heritability, QTL, and SNP Panel Scenarios Using Artificial Neural Network

SUNDAY O. PETERS\textsuperscript{1,2}, MAHMUT SINECEN\textsuperscript{3}, (Member, IEEE), KADIR KIZILKAYA\textsuperscript{4}, AND MILTON G. THOMAS\textsuperscript{5}

\textsuperscript{1}Department of Animal Science, School of Mathematical and Natural Sciences, Berry College, Mount Berry, GA 30149, USA
\textsuperscript{2}Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
\textsuperscript{3}Department of Computer Engineering, Faculty of Engineering, Aydin Adnan Menderes University, 09100 Aydin, Turkey
\textsuperscript{4}Department of Animal Science, Faculty of Agriculture, Aydin Adnan Menderes University, 09100 Aydin, Turkey
\textsuperscript{5}Department of Animal Science, Colorado State University, Fort Collins, CO 80526, USA

Corresponding author: Sunday O. Peters (speters@berry.edu)

This work was supported in part by the United States Department of Agriculture (USDA)- Agriculture and Food Research Initiative (AFRI) through the Brangus Genotypes under Grant 2008-35205-18751 and Grant 2009-35205-05100, and in part by the New Mexico Agriculture Experiment Station Project under Grant 216391.

\textbf{ABSTRACT} A simulation study was carried out to determine the genomic prediction performance of Artificial Neural Network model with 1 to 10 neurons (ANN-1–10) using 3361 SNP markers from BovineSNP50 (Infinium BeadChip, Illumina, San Diego, CA) on the first chromosome of Brangus beef cattle as a pilot study for two traits with heritabilities of 25\% (\(h^2_T_1 = 0.25\)) and 50\% (\(h^2_T_2 = 0.5\)) determined either by 50, 100, 250 or 500 QTL selected randomly from SNP markers. QTL effects were sampled from a multivariate normal distribution. Genomic predictions were carried out by Feed Forward Multi-Layer Perceptron ANN-1–10 with the back-propagation of errors algorithm employing the Levenberg–Marquardt algorithm to locate the optimal solution. Three sets of SNP panels were used for genomic prediction: only QTL (Panel-1), all SNP markers, including the QTL (Panel-2), and all SNP, excluding the QTL (Panel-3). Correlations between true genetic merits (breeding values) and predicted phenotypes from 10-fold disjoint cross-validation were used to assess predictive ability of ANN-1-10. Results indicated that an increase in heritability resulted in an increased predictive performance of ANN-1-10 for all scenarios. SNP Panels had a greater chance of including markers in LD with QTL, allowed the possibility of predicting the effect of each QTL from the collective action of several markers and performed better than the Panel including only QTL. In the other Panels, predictive performance of ANN-1-10 increased inconsistently with the number of neurons, which indicated that a few numbers of neurons were not be enough to learn specification of data and could cause the under fitting problem. Therefore, high number of neurons could be needed to learn relevant details of the data in the applications of ANN.

\textbf{INDEX TERMS} Genomic prediction, QTL, SNP, artificial neural network, heritability, number of neurons.

I. INTRODUCTION

Economic important (complex quantitative) traits such as milk/protein or grain yields, in animal and plant productions, respectively, are measured on continuous scales and assumed to be determined by large numbers of genes with small additive effects [1]. The genetic analysis of these complex traits was carried out using additive genetic relationships among individuals based on the pedigree information. However, high-throughput genotyping technologies in last two decades have resulted in enormous amounts of high-density single nucleotide polymorphisms (SNP) in an accurate and cost-efficient manner. The study of [2] pioneered the genomic selection which has been utilized to understand complexity of traits and to account for the effects that are contributed by these genes using SNP markers. For the application of genomic selection in plant and animal breeding, numerous parametric statistical methods were developed. The methods of BayesA and BayesB were developed by assuming Scaled-t and Scaled-t mixture distributions for SNP marker effects [2] and were extended by assuming Gaussian (called Bayesian ridge regression in [3]), Gaussian mixture (called BayesC and
BayesC π in [4]) and Double exponential (called Bayesian Lasso in [5]) for SNP marker effects. Also, reproducing kernel Hilbert spaces method as nonparametric genomic regression was developed for the application of genomic selection ([6], [7], [8]).

Complex quantitative traits are controlled by a network of numerous genes [9] and artificial neural networks (ANN) provide an interesting alternative to predict genomic (or phenotypic) values of individuals using SNP markers for complex traits in animal and plant breeding. The operations of nerve cells (known neurons) in the human brain inspired the development and use of ANN for artificial intelligence. The ANN being a statistical modeling of the human brain functions represents a new generation of information processing systems [10]. The learning ability of ANN without using any prior assumption and sophisticated statistical models for linear and nonlinear relationships in the information processing systems is a very attractive feature. In recent years, the use of ANN statistical models in the studies of genomic selection (prediction) has shown an increasing trend [11]–[16]. In the studies of plant and animal breeding, the number of parameters (SNP marker effects) to be estimated is much more than the available sample size and the computational costs of training the ANN applications are high. Therefore, at least in animal breeding, only subsets of markers have been used to make ANN computational feasible [17] [18]. Actual high-density SNP genotypes exhibit linkage disequilibrium (LD) that is not dependent upon assumed population sizes or mutation rates [19]. This study used actual high-density SNP genotypes on the first chromosome of Brangus beef cattle to simulate two traits with heritabilities of 25% ($h^2_T1 = 0.25$) and 50% ($h^2_T2 = 0.5$) determined by QTL-50, QTL-100, QTL-250 or QTL-500 and then aimed to determine how molecular architecture of the traits affect the likely accuracy of genomic prediction from ANN model with 1 to 10 neurons (ANN-1–10). We also aim to determine the effects of Linkage Disequilibrium and different number of inputs on predictive performance of ANN.

II. MATERIAL AND METHODS

A. SINGLE NUCLEOTIDE POLYMORPHISM (SNP) GENOTYPES

SNP marker genotypes were obtained from the studies of [20] describing the acquisition of 53,692 SNP marker genotypes using BovineSNP50 (Infinium BeadChip, Illumina, San Diego, CA) from 719 Brangus heifers registered in International Brangus Breeders Association. Call rates of 53,692 SNP marker genotypes averaged 98.1 ± 0.001%. SNP marker genotypes obtained in the A/B allele format were coded as 0 (AA), 1 (AB), or 2 (BB).

The application of ANN to high-density SNP marker genotypes is a significant challenge and needs improved algorithms because of the requirement of high computing time and memory in the study of genomic prediction [17]. Therefore, in this simulation study only 3,361 SNP marker genotypes on the first chromosome from 53,692 SNP marker genotypes were used [20].

B. SIMULATION OF CORRELATED ADDITIVE GENETIC MERITS AND CORRELATED PHENOTYPIC PERFORMANCES

Two traits ($T_1$ and $T_2$) based on the parameters of genetic correlation $r^2_{T1T2} = 0.50$ and heritabilities of 25% ($h^2_{T1} = 0.25$) and 50% ($h^2_{T2} = 0.5$) were simulated by using 50, 100, 250 or 500 additive bi-allelic QTL. In order to represent QTL, a random sample of $N = 50, 100, 250$ or 500 SNP markers was chosen from the observed 3,361 Illumina SNP markers on the first chromosome of Brangus heifers. The choice of numbers of SNP markers was based on the results of the analyses done by Kizilkaya et al. 2010.

An equal probability for each locus was given in the random sampling process without consideration of minor allele frequency. Parametric substitution effects for each selected QTL determining the two traits ($T_1$ and $T_2$) were assigned by sampling from a multivariate normal distribution with mean $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and variance-covariance matrix $\Psi = \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 2.0 \end{bmatrix}$.

Additive genetic merits of each animal for the two traits were calculated by summing the substitution effects for each QTL allele [19]. Residual effects for each animal were generated from a multivariate normal distribution with mean $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and variance-covariance matrix $R = \begin{bmatrix} 3.0 & 0.7 \\ 0.7 & 2.0 \end{bmatrix}$. Phenotypic values of each animal for two traits were generated by adding its residual effects to its additive genetic merits (breeding values) as in Eq. (1):

$$y_i = \mu + \sum_{j=1}^{N} g_{ij} \beta_j + e_i$$

where $y_i = \begin{bmatrix} y_{T1} \\ y_{T2} \end{bmatrix}$ is a vector of the simulated phenotypic values of animal $i$, $\mu = \begin{bmatrix} \mu_{T1} = 5 \\ \mu_{T2} = 10 \end{bmatrix}$ is a vector of population means for two traits, $u = \sum_{j=1}^{N} g_{ij} \beta_j$ is a vector of the additive genetic merits of animal $i$ generated by multiplying the covariate $g_{ij}$ (0, 1 or 2) by the substitution effects $\begin{bmatrix} \beta_j \end{bmatrix}$ for the locus $j$ and summing them over all $N$ loci and $e_i$ is a vector of the random residual effects [19]. Datasets of two traits were simulated for each of $N = 50, 100, 250$, or 500 QTL scenario (QTL-50, QTL-100, QTL-250, or QTL-500). All simulations were repeated 5 times.

C. multilayer feed-forward neural networks

Multilayer feed-forward neural network is one of the most common ANN used in genome-enable prediction of complex traits in animal and plant production [22]. As seen in
Weights (\(w_{ij}^{[1]}\)) from input to hidden layer

Weights (\(w_{kt}^{[2]}\)) from hidden to output layer

\[ Z_{kt}^{[1]} = f_1 \left( \alpha_t + \sum_{j=1}^{N} g_{ij} w_{ij}^{[1]} \right) \]

\[ Z_{kt}^{[2]} = f_2 \left( \alpha_t + \sum_{j=1}^{N} g_{ij} w_{ij}^{[2]} \right) \]

\[ y_{ki}^{[2]} = \delta_{\phi k}^{[2]} = \delta_{\phi t}^{[2]} + e_{ki} \]

\[ \delta_{\phi t}^{[2]} = \beta + \sum_{t=1}^{s} w_{kt}^{[2]} Z_{kt}^{[2]} \]

\[ \delta_{\phi t}^{[1]} = \beta + \sum_{t=1}^{s} w_{kt}^{[1]} Z_{kt}^{[1]} \]

\[ y_{ki}^{[1]} = \phi \left( \beta + \sum_{t=1}^{s} w_{kt}^{[1]} Z_{kt}^{[1]} \right) + e_{ki} \] (Eq. 2)

FIGURE 1. Architecture of a two-layer feed-forward (single hidden layer perceptron) neural network with ten neurons in the hidden layer. \(g_{ij}\): QTL/SNP marker genotype \(j\) of animal \(i\) in input layer; \(w_{ij}^{[1]}\): network weight from the input layer to neuron \(t\) of the hidden layer; \(w_{kt}^{[2]}\): network weight from the hidden layer to the output layer for neuron \(t\) \((t = 1, \ldots, s = 10)\); \(y_{ki}^{[2]}\): predicted phenotype of individual \(i\) for trait \(k \in \{T_1, T_2\}\) from neuron \(t\) in the output layer; \(f_1\): non-linear (tangent sigmoid) activation function in the hidden layer; \(\phi()\): linear activation function in the output layer; \(\alpha_t\) and \(\beta\) represent the intercepts in the hidden and output layers, respectively.

As seen in Figure 1, phenotypes of animals for the two traits were predicted separately (independently each other) in the output layer of the two-layer feed-forward neural network. The two-layer feed-forward neural network is trained by estimating the weights (\(w_{1j}\) and \(w_{2j}\)) and biases (\(\alpha_t\) and \(\beta\)) based on the minimization of the differences between the observed and predicted phenotypes. The most widely used training or learning algorithm for the two-layer feed-forward neural network is the back-propagation of errors algorithm [22] employing the Levenberg–Marquardt algorithm to locate the optimal solution [25].

MATLAB® Neural Network Toolbox was used to fit the two-(hidden and output)-layer-feed-forward neural network (Figure 1a and 1b). As seen in Figure 1a and 1b, there were different number (50 to 3361) of QTL and SNP marker combinations from Panel-1 to Panel-3 as input, one to ten neurons (hidden nodes) in a single hidden layer and one node in the output layer. Levenberg–Marquardt algorithm with

\[ y_{ki}^{[1]} = \phi \left( \beta + \sum_{t=1}^{s} w_{kt}^{[1]} Z_{kt}^{[1]} \right) + e_{ki} \] (Eq. 2)

\[ \delta_{\phi t}^{[2]} = \beta + \sum_{t=1}^{s} w_{kt}^{[2]} Z_{kt}^{[2]} \]

\[ \delta_{\phi t}^{[1]} = \beta + \sum_{t=1}^{s} w_{kt}^{[1]} Z_{kt}^{[1]} \]

\[ y_{ki}^{[2]} = \phi \left( \beta + \sum_{t=1}^{s} w_{kt}^{[1]} Z_{kt}^{[1]} \right) + e_{ki} \] (Eq. 2)
backpropagation in MATLAB® Neural Network Toolbox was used to minimize the mean squared error (MSE) between true and predicted phenotypes, \( \text{MSE} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \) where \( y_i \) and \( \hat{y}_i \) are true and predicted phenotypes in the network. In training sets with between 50 and 3361 QTL and SNP markers and 10 neurons, there were about from 500 to 33610 weights and biases to estimate from only 719 animals. The number of epochs used was 1000 and a forward activation occurred to produce a solution for weights and a backward propagation of the computed error occurred to modify the weights in each epoch (iteration). When the maximum number of 1000 epochs were reached or when the averaged mean squared error (MSE) reached a certain threshold (MSE \( \leq 10^{-6} \)) the algorithm was stopped.

### D. USE OF MARKER PANELS

Genome-enabled prediction of phenotypic records of animals for two traits with \( h^2_1 = 0.25 \) and \( h^2_2 = 0.5 \) were obtained from the two-layer feed-forward neural network. In the architecture of the two-layer feed-forward neural network, the number of neurons between 1 and 10 were used to test the predictive performance of neural network. Also, three SNP marker panels were applied through the input layer of the two-layer feed-forward neural network: Panel-1 for only QTL genotypes (\( N = 50, 100, 250 \) or 500 QTL); Panel-2 for all QTL and SNP markers (\( N = 3361 \text{SNP} \)) and Panel-3 for only SNP markers (\( N = 3311, 3261, 3111 \) or 2861), excluding the QTL (50, 100, 250 or 500 QTL). The correlations between true genetic and predicted phenotypes from 10-fold training and testing (cross validation) datasets for two traits with \( h^2_1 = 0.25 \) and \( h^2_2 = 0.5 \) were used to assess genome-enabled predictive ability of the two-layer feed-forward neural network with 1 to 10 neurons based on 4 QTL scenarios with 3 panels of SNP genotypes.

### E. LINKAGE DISEQUILIBRIUM ANALYSIS

Linkage disequilibrium (LD) is defined as a nonrandom association between alleles of different loci in a given population and is estimated using the squared correlation (\( r^2 \)) between pairwise combination of all SNP markers expressed as in Eq. (3):

\[
\frac{(p_{AB} - p_A p_B)^2}{(p_A p_A) (p_B p_B)}
\]

where \( p_{AB} \) is defined as the observed frequency of haplotype \( AB, p_A = 1 - p_a \) and \( p_B = 1 - p_b \) the observed frequencies of allele \( A \) at locus \( i \) and allele \( B \) at locus \( j \) [26]. As the squared correlation (\( r^2 \)) ranges from 0 to 1 as \( p_{AB}, p_A \) and \( p_B \) vary. The estimates of LD among all 3361 SNP markers were obtained from using the function of pairwise LD in the R package Synbreed [27]. The distribution and decay of LD were examined by plotting the squared correlation (\( r^2 \)) by the distance between SNP markers and calculating the descriptive statistics.

### F. ASSESSMENT OF THE MODELS’ PREDICTIVE ABILITY

The predictive performance of the two-layer feed-forward neural network for 4-QTL and 2-heritability scenarios in the simulation of genomic data and 3-panel scenario in the analysis of genomic data was evaluated in a 10-fold training and testing (cross-validation, CV) datasets using five replicates. All QTL, heritability and panel scenarios for the data simulation and analysis are summarized in Table 1.

Each of the five simulated datasets included 719 animals having phenotype and SNP genotypes and was randomly divided into 10 disjoint approximately equal datasets (\( CV_1, CV_2, \ldots, CV_{10} \)), with nine-tenths (647 observations) used for training and one-tenth (72 observations) used for testing, and the datasets used for training and testing were

---

**TABLE 1.** The scenarios with respect to the number of QTL, the heritability of the traits, QTL and SNP marker panels and the number of neurons in ANN.

| Panels for QTL scenarios | Trait-1 | Trait-2 |
|--------------------------|---------|---------|
|                          | Simulation of genomic data | Analysis of genomic data |
| Heritability (h^2)       | 0.25    | 0.50    |
| Number of QTL            | 50      | 100     | 250     | 500     | 50      | 100     | 250     | 500     |
| Cross-Validation         | 10      | 10      | 10      | 10      | 10      | 10      | 10      | 10      |
| Number of replications   | 5       | 5       | 5       | 5       | 5       | 5       | 5       | 5       |

For two traits with genome-enabled prediction of phenotypic records of animals. The number of epochs used was 1000 and a forward activation occurred to produce a solution for weights and a backward propagation of the computed error occurred to modify the weights in each epoch (iteration). When the maximum number of 1000 epochs were reached or when the averaged mean squared error (MSE) reached a certain threshold (MSE \( \leq 10^{-6} \)) the algorithm was stopped.

#### E. LINKAGE DISEQUILIBRIUM ANALYSIS

Linkage disequilibrium (LD) is defined as a nonrandom association between alleles of different loci in a given population and is estimated using the squared correlation (\( r^2 \)) between pairwise combination of all SNP markers expressed as in Eq. (3):

\[
\frac{(p_{AB} - p_A p_B)^2}{(p_A p_A) (p_B p_B)}
\]

where \( p_{AB} \) is defined as the observed frequency of haplotype \( AB, p_A = 1 - p_a \) and \( p_B = 1 - p_b \) the observed frequencies of allele \( A \) at locus \( i \) and allele \( B \) at locus \( j \) [26]. As the squared correlation (\( r^2 \)) ranges from 0 to 1 as \( p_{AB}, p_A \) and \( p_B \) vary. The estimates of LD among all 3361 SNP markers were obtained from using the function of pairwise LD in the R package Synbreed [27]. The distribution and decay of LD were examined by plotting the squared correlation (\( r^2 \)) by the distance between SNP markers and calculating the descriptive statistics.

#### F. ASSESSMENT OF THE MODELS’ PREDICTIVE ABILITY

The predictive performance of the two-layer feed-forward neural network for 4-QTL and 2-heritability scenarios in the simulation of genomic data and 3-panel scenario in the analysis of genomic data was evaluated in a 10-fold training and testing (cross-validation, CV) datasets using five replicates. All QTL, heritability and panel scenarios for the data simulation and analysis are summarized in Table 1.

Each of the five simulated datasets included 719 animals having phenotype and SNP genotypes and was randomly divided into 10 disjoint approximately equal datasets (\( CV_1, CV_2, \ldots, CV_{10} \)), with nine-tenths (647 observations) used for training and one-tenth (72 observations) used for testing, and the datasets used for training and testing were
rotated ten times. Then the two-layer feed-forward neural network was trained and tested 10 times using 10-fold training and testing datasets. In the training analysis of the two-layer feed-forward neural network, the simulated dataset excluded one testing (CV_{-i}) of CV datasets \( i \in \{1, 2, \ldots, 10\} \) to estimate the parameters of the two-layer feed-forward neural network. In the testing analysis of the two-layer feed-forward neural network, SNP genotypes of animals in the CV_{-i} were presented to the two-layer feed-forward neural network to obtain the predicted phenotypes.

The correlations between true genetic and predicted phenotypes from 10-fold cross validation for two traits with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) were used to assess genome-enabled predictive ability of the two-layer feed-forward neural network with 1 to 10 neurons based on 4 QTL scenarios with 3 panels of SNP marker genotypes. Pearson’s correlation coefficient \( (r_{uk,\hat{y}_k}) \) between the true genetic merits (breeding values) \((u_k)\) and the predicted \((\hat{y}_k)\) phenotypes in training and testing datasets from 10-fold cross validation for trait \( k \) (\( T_1 \) or \( T_2 \)) was calculated using Eq. (4):

\[
r_{uk,\hat{y}_k} = \frac{S_{u_k,\hat{y}_k}}{\sqrt{S^2_{u_k}S^2_{\hat{y}_k}}} \tag{4}
\]

where \( u_k \) and \( \hat{y}_k \) are the true genetic merits and the predicted phenotypes for training or testing dataset, \( S_{u_k,\hat{y}_k} \) is the covariance between the true genetic merits and the predicted phenotypes, \( S^2_{u_k} \) and \( S^2_{\hat{y}_k} \) are the variances for the true genetic merits and the predicted phenotypes for trait \( k \) (\( T_1 \) or \( T_2 \)), respectively.

### III. RESULTS AND DISCUSSION

Genomic prediction for two traits with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) determined either by QTL-50, QTL-100, QTL-250 or QTL-500 were studied by using ANN-1-10 architecture with QTL and SNP markers through Panel-1 (only QTL), Panel-2 (QTL and SNP markers) and Panel-3 (only SNP markers) as input to the ANN. The average Pearson’s correlation coefficients were obtained for 4-QTL scenarios from 10-fold training and testing datasets across 5 replicates analyzed by ANN-1-10 architecture and given for the traits with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) in columns, with Panels in rows of Figure 2 for training datasets and of Figure 3 for testing datasets, respectively. As seen in Figures 2 and 3, the average Pearson’s correlation coefficients showed that the predictive performance of ANN depended tightly on the number of neurons used (architecture of network) in the hidden layer of the network, the number of features (QTL and SNP markers in Panels) of network and genetic architecture of trait (heritability of trait and the number of QTL determining trait).

#### A. PREDICTIVE PERFORMANCE OF ANN WITH THE DIFFERENT NUMBER OF NEURONS

The predictive performance of ANN models in 10-fold training and testing datasets differed substantially in terms of the number of neurons in the hidden layer. The average Pearson’s correlation coefficients for the predictive performance of ANN models across the number of neurons are given in Figure 2 for training datasets and Figure 3 for testing datasets. As seen in Figure 2 and 3, the pattern about predictive performance of ANN models through the number of neurons was similar for training and testing datasets when different QTL and SNP Panels, and heritabilities were used. When the number of neurons increased in the hidden layer of ANN an increasing trend was observed for the predictive performance of ANN models in Panel-2 (all QTL and SNP markers, \( N = 3361 \) SNP) and Panel-3 (only SNP markers, \( N = 3311, 3261, 3111 \) or 2861). However, there were no clear increasing trend across the number of neurons in Panel-1 (only QTL genotypes, \( N = 50, 100, 250 \) or 500 QTL). The number of neurons in the hidden layer influenced the predictive performance of ANN models. Reference [28] studied the impact of the number of neurons and type of genomic information used as input to the network on the predictive performance. They indicated that a few numbers of neurons (2-6 neurons in the hidden layer) was not enough to learn specification of data and could cause the under fitting problem, whereas high number of neurons (over 15 neurons in the hidden layer) could learn irrelevant details of the data and could cause the over fitting problem in the applications of ANN. In this study, the minimum and maximum values of the average Pearson’s correlation coefficients given for QTL-50, QTL-100, QTL-250 and QTL-500 scenarios within trait 1 and 2 with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) indicated the improvements in the predictive ability of ANN through neurons within all 3 Panels. Reference [29] compared the predictive ability of different Bayesian regularized backpropagation ANN using SNP markers and varying number of neurons (1-7) in the hidden layer and indicated that in terms of predictive ability, a network with five neurons in the hidden layer attained the smallest error and highest correlation in the test data although differences among networks were negligible. References [28], [30] and [31] also showed that the ANN models differed little in predictive performance in terms of number of neurons in the hidden layer when the genomic relationship G matrix was used. These results showed that predictive ability of ANN models did not depend on the number of neurons (network architecture) when sample size was larger than the number of markers used in the ANN analyses. Our ANN predictive performance results from Panel-1 for QTL-50 scenarios in Figure 2, 3 and 4 with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) as well as in the hidden layer of the network when sample size was equal to or higher than the number of features of the network. However, as seen in Figures 2 and 3, the pattern about predictive performance of ANN models in Panel-2 and Panel-3 for QTL-50, QTL-100, QTL-250 and QTL-500 scenarios within trait 1 and 2 with \( h^2_{11} = 0.25 \) and \( h^2_{22} = 0.5 \) depended tightly on the number of neurons used in the hidden layer of the network when SNPs were used as input to the network. Neurons in the neural network are interconnected compute units calculating outputs from weighted information and a well-chosen number of neurons is necessary to learn the features of SNP markers as input...
FIGURE 2. Comparison of predictive abilities of ANN for all scenarios in training dataset. Traits with heritability are in the columns. Panels in rows show the average Pearson’s correlation coefficients over 10-traing datasets across 5 replicates for 4-QTL scenarios on the vertical axis, and the number of hidden neurons on the horizontal axis.

and to capture structure in the data [32]. Reference [28] reported that the correlations about predictive performance of ANN differed across the number of neurons. In the study of [29], and [31], correlation coefficients increased slightly with the number of neurons in the training dataset and decreased substantially in the testing datasets. However,
in this simulation study, the comparison of average Pearson’s correlation coefficients from training and testing datasets showed that there were small differences ranging from 1.19 to 4.78% which indicated that the ANN application with the number of neurons had a consistent pattern in the prediction performance and learned the specification from datasets.
B. EFFECTS OF LINKAGE DISEQUILIBRIUM AND DIFFERENT NUMBER OF INPUTS ON PREDICTIVE PERFORMANCE OF ANN

The number of inputs (features) in the network analysis has impact on the predictive performance of ANN models. In this simulation study, the number of features was determined by QTL and SNP Panels. QTL genotypes \( (N = 50, 100, 250 \text{ or } 500 \text{ QTL}) \) in Panel-1, QTL and SNP marker genotypes \( (N = 3361 \text{ SNP}) \) in Panel-2 and only SNP marker genotypes \( (N = 3311, 3261, 3111 \text{ or } 2861) \) in Panel-3 were fed into the ANN system for genomic selection. Figure 2 and 3 showed the effect of the number of features (QTL and SNP markers) on the predictive performance of ANN-1-10 architecture. The number of QTL genotypes \( (N = 50, 100, 250 \text{ or } 500 \text{ QTL}) \) in Panel-1 was lower than the sample size; however, the number of QTL and/or SNP markers in Panel-2 and in Panel-3 were higher than the sample size. Therefore, results from testing datasets in Figure 3 showed that the predictive ability of the ANN models did depend on the number of features of network regardless of sample size. Panel-1 resulted in the lowest correlation coefficients \( (0.446, 0.343, 0.473 \text{ and } 0.446) \) in trait 1 with \( h^2_{T1} = 0.25 \) and \( (0.653, 0.540, 0.620 \text{ and } 0.601) \) in trait 2 with \( h^2_{T2} = 0.5 \) for QTL genotypes. However, Panel-2 including QTL and SNP markers provided the highest correlation coefficients \( (0.555, 0.501, 0.554 \text{ and } 0.529) \) in trait 1 with \( h^2_{T1} = 0.25 \) and \( (0.664, 0.630, 0.690 \text{ and } 0.642) \) in trait 2 with \( h^2_{T2} = 0.5 \) and Panel-3 including only SNP markers, which is more realistic, resulted in slightly lower correlation coefficients than Panel-2. Overall, Panel-2 and Panel-3 produced between 14.7\% and 44.3\% higher correlation coefficients for trait 1 with \( h^2_{T1} = 0.25 \) and between 17.1\% and 57.3\% higher correlation coefficients for trait 2 with \( h^2_{T2} = 0.5 \) compared to the Panel-1. Although the study results of [33] showed the less predictive capacity in the use of a higher density of markers for different scenarios of SNP markers, heritabilities and linkage disequilibrium, [34] and [35] found the increasing prediction accuracy with larger marker density in the studies of complex human traits using different number of SNP markers. Reference [36] investigated the performance of random forest and genomic BLUP for the different scenarios according to heritability, number of QTL, marker density, and the LD structure of the genotyped population and observed the improvements in the prediction performance of random forest with increasing marker density. Reference [37] also studied the accuracies in genomic selection for wood growth and quality using different subsets of markers and identified lower accuracies by reducing density to less than 500 SNPs in a black spruce (Picea Mariana) population while assessing wood growth and quality. Reference [17] analyzing marbling score EPD using ANN models indicated that correlation coefficients for prediction accuracy in training dataset increased almost monotonically with panel size until reaching a plateau. However, in testing dataset, correlation coefficients reached its peak, then started decreasing which indicates possible over-fitting problem in the training dataset resulted from the SNP panel size exceeding the training sample size. Although in theory, the more markers there are, the better the prediction, accuracy is difficult to improve meaningfully when the density reaches a certain degree ([38]). The studies of [39], [40] and [41] showed that the increase of marker density resulted in a rapid increase in the prediction accuracy until a plateau, then very little or marginal improvement in prediction accuracy occurred when the marker density kept to increase. Reference [36] explained that an increase of SNP markers shortened the distances between markers and functional mutations; then, SNP markers close to a QTL were sampled with sufficient frequency, implying that the signal of the QTL is captured by distinct SNP in close map distance. Reference [19] also indicated that Panels with more SNP had a greater chance of including markers in LD with QTL and allowed the possibility of predicting each QTL from the collective action of several markers.

A strong and positive relationship between LD and predictive performance of models for genomic selection was observed in previous studies (e.g., [42], [43]). The distribution of LD with respect to the physical distance (Mb) separating loci in this study is presented in Figure 4. LD was calculated as a square of Pearson’s correlation coefficient \( r^2 \) between actual 3,361 SNP marker genotypes. Actual LD had overall mean of 0.02 and the average of LD at distance of less than 0.05 Mb between SNP markers was 0.16. As seen in Figure 4, LD was higher especially for SNP markers at close physical distances and declined gradually with increasing physical distance between SNP markers. However, as seen in Figure 4, some higher LD also was observed for SNP markers at longer distances. Results from ANN-1-10 architecture in this simulation study were consistent with those of [19], who showed the predictive ability of Bayes-C method applied to the similar simulated genomic data for genomic selection. As seen in the results from [19], Panels including more SNP markers such as Panel-2 and -3 have a higher chance of including SNP markers in LD with QTL and predicting the genetic contribution of each QTL from the collective action of many SNP markers. Reference [31] investigated the effect of the genomic architecture of traits and reported a strong influence of LD on the accuracies of genomic prediction. Results from [43] also revealed a stronger effect of LD on prediction accuracies when applying Random Forest than when applying GBLUP.

C. PREDICTIVE ABILITY ACROSS TRAITS WITH \( h^2_{T1} = 0.25 \) AND \( h^2_{T2} = 0.5 \)

The effect of heritabilities of 25\% \( (h^2_{T1} = 0.25) \) and 50\% \( (h^2_{T2} = 0.5) \) on the predictive performance of ANN model with 1 to 10 neurons for 4-QTL scenarios and marker panels is depicted in Figure 2 and 3 for training and testing datasets, respectively. As expected, increase in heritability was associated with increased predictive performance of ANN-1-10.
FIGURE 4. Distribution of linkage disequilibrium (LD) between actual 3,361 SNP marker genotypes from Brangus cattle with respect to the distance (Mb) separating loci.

for 4-QTL scenarios and marker panels and this trend was similar for training and testing datasets. The percentage changes (increases) for 4-QTL scenarios were 42, 55, 28 and 31% in Panel-1, 19, 25, 23% in Panel-2 and 20, 28, 24 and 17% in Panel-3 for training datasets. They were 46, 57, 31 and 35% in Panel-1, 20, 26, 25 and 21% in Panel-2 and 21, 29, 25 and 17% in Panel-3 for testing datasets. However, the increases on predictive performances of ANN-1-10 in Panel-1 were higher than those in Panel-2 and -3 within each QTL scenario and predictive performances of ANN-1-10 of Panle-2 and -3 were similar within each QTL scenarios. The effect of genetic architecture (the number of QTL and heritability) of trait on genomic selection in the comparison of different genomic models were studied by [44] and [45] and they found that models for genomic selection were sensitive to the number of QTL and heritability and decreasing heritability resulted in the decrease in the predictive performance of genomic models. Reference [4] also determined significant trends for the number of SNP depended on heritability, number of QTL and the distribution of QTL effects.

The decrease in predictive performances of ANN-1-10 from training to testing datasets was also obvious within heritabilities of 25% ($h^2_{T1} = 0.25$) and 50% ($h^2_{T2} = 0.5$). The percentage decreases in predictive performance ranged from 2.04 to 4.78% in heritabilities of 25% ($h^2_{T1} = 0.25$) and from 1.19 to 3.44% in heritabilities of 50% ($h^2_{T2} = 0.5$) across 4-QTL and marker panels. Increase in heritability resulted in less decrease in predictive performance of ANN-1-10 of testing datasets. The association between predictive performance and heritabilities of training and testing datasets was also determined in the study of [46], indicating that the increasing genomic heritability was due to higher genetic variations in training and testing sets, contributing to accurate predictions of marker effects. Reference [47] investigated the performance of several ANN architectures for prediction of breeding values and determined on average 1% and 0.5% increase in prediction accuracy estimates by ANN considering the heritabilities of 40 and 70%, respectively. In the study of evaluating the efficiency of ANN in predicting genetic value under the scenarios of heritabilities, variations and genotype blocks, [48] determined that the difference between the correlation between network value and genetic value and the correlation between phenotypic value and genetic value was higher in experiments with lower heritability (10 and 20%), which means, the environmental effect of this characteristic is increased, i.e., noise is increased. Reference [36] also found that decreases in heritability of traits and reduction in QTL were associated with decreasing prediction accuracy in genomic predictions using Random Forest and GBLUP for a simulated binary trait. Increasing effects of heritability on the prediction accuracy has been previously confirmed in other studies on domestic animals using simulated ([49]) and real genomic data ([40], [50], [51] investigated the prediction accuracies of GBLUP methods in American mink using simulated data with different levels of heritability and found that accuracy of genomic methods was increased with enhancement in heritability levels and marker densities. Reference [38] indicated that heritability was positively related to prediction accuracy and using the same genic
selection method, prediction accuracy of a high-heritability trait (such as thousand-grain weight) was often higher than that of a low-heritability trait (such as grain yield) [52] noted a positive linear association between predictive ability of models and heritabilities, i.e., traits with lower heritabilities (below 0.25) exhibited the lowest predictive ability while higher predictive ability were detected for traits with moderate heritabilities (above 0.30). Reference [53] 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.

IV. CONCLUSION

Results from ANN-1-10-neuron analysis with different QTL and SNP marker panels applied to two traits with heritabilities of 25% ($h^2_1 = 0.25$) or 50% ($h^2_2 = 0.5$) determined either by QTL-50, QTL-100, QTL-250 or QTL-500 scenarios suggest that:

- Higher genomic predictive performance was achieved for the trait with high heritability than the trait with low heritability. The increase in heritability resulted in the increased predictive performance of ANN-1-10 neurons for 4-QTL scenarios and marker panels.
- The Panels including SNP markers had a greater chance of including markers in LD with QTL, allow the possibility of predicting each QTL from the collective action of several markers and perform better prediction than the Panel including only QTL.
- In ANN, as the number of neurons increases, the number of parameters to be estimated also increases; and as the number of parameters rises, the risk of over-fitting also increases. However, predictive performance results from training and testing datasets were consistent with increasing number of neurons within each heritability, QTL and panel scenarios and over-fitting problem in the predictive performance of ANN-1-10 neurons were not observed.
- ANN model with one neuron had less parameters and resulted in predictive performance similar with those from ANN model with more than one neuron within the Panel-1 with scenarios QTL-50 for heritabilities of 25% ($h^2_1 = 0.25$) and 50% ($h^2_2 = 0.5$).
- In the other Panels, predictive performance of ANN-1-10-neuron models increased with the number of neurons. However, there was no consistent increase in the predictive performance across the number of neurons, which indicated that a few numbers of neurons could not be enough to learn specification of data and could cause the under fitting problem. Therefore, high number of neurons could be needed to learn relevant details of the data in the applications of ANN.

REFERENCES

[1] D. S. Falconer and T. F. C. Mackay, Introduction to Quantitative Genetics, 4th ed. London, U.K.: Pearson, 1996.

[2] B. J. Hayes and M. E. Goddard, “Prediction of total genetic value using genome-wide dense marker maps,” Genetics, vol. 157, no. 4, pp. 1819–1829, Apr. 2001.

[3] J. B. Endelman, “Ridge regression and other kernels for genomic selection with r package rrBLUP,” Plant Genome, vol. 4, no. 3, pp. 250–255, Nov. 2011, doi: 10.3835/plantgenome2011.08.0024.

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

[5] T. Park and G. Casella, “The Bayesian lasso,” J. Amer. Stat. Assoc., vol. 103, no. 482, pp. 681–686, Jun. 2008, doi: 10.1198/016214508000000337.

[6] G. de los Campos, D. Gianola, and G. J. M. Rosa, “Reproducing kernel Hilbert spaces regression: A general framework for genomic evaluation,” J. Animal Sci., vol. 87, no. 6, pp. 1883–1887, Jun. 2009, doi: 10.2527/jas.2008-1259.

[7] G. De Los Campos, D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa, “Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods,” Genet. Res., vol. 92, no. 4, pp. 295–308, Aug. 2010, doi: 10.1017/S0016672310000285.

[8] D. Gianola, R. L. Fernando, and A. Stella, “Genomic-assisted prediction of genetic value with semiparametric procedures,” Genetics, vol. 173, no. 3, pp. 1761–1776, Jul. 2006, doi: 10.1534/genetics.105.049510.

[9] R. Howard, A. L. Carriquiry, and W. D. Beavis, “Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures,” G3, Genes, Genomes, Genet., vol. 4, no. 6, pp. 1027–1046, Jun. 2014, doi: 10.1534/g3.114.010298.

[10] B. de Bragança Pereira, C. R. Rao, and M. Rao, Data Mining using Neural Networks: A Guide for Statisticians, 1st ed. London, U.K.: Chapman & Hall, 2000.

[11] J. M. González-Camacho, G. de los Campos, P. Pérez, D. Gianola, J. E. Cairns, G. Mahuku, R. Babu, and J. Crossa, “Genome-enabled prediction of genetic values using radial basis function neural networks,” Theor. Appl. Genet., vol. 125, no. 4, pp. 759–771, Aug. 2012, doi: 10.1007/s00122-011-1986-9.

[12] S. Chesang, T. K. Musaaya, and K. Ngeno, “Predictions of indigenous chicken phenotypes from genotypes: Comparison between machine learning and conventional linear models,” Res. Square, vol. 1, no. 1, pp. 8–14, 2019, doi: 10.21203/rs.2.13280v1.

[13] A. A. C. Alves, “Applying machine learning methods for genomic analysis of reproductive traits in nellore cattle,” Ph.D dissertation, Genética e Melhoramento Animal FCAV, Universidade Estadual Paulista, São Paulo, Brazil, 2019. Accessed: Dec. 12, 2019. [Online]. Available: http://hdl.handle.net/11449/1191316.

[14] A. M. F. Ribeiro, “Application of artificial neural networks to genome-enabled prediction in Nellore,” Ph.D dissertation, Genética e Melhoramento Animal FCAV, Universidade Estadual Paulista, São Paulo, Brazil, 2019. Accessed: Jul. 29, 2019. [Online]. Available: http://hdl.handle.net/11449/190702.

[15] F. B. Lopes, C. U. Magnabosco, T. L. Passafaro, L. C. Brunes, M. F. O. Costa, E. C. Eifert, M. G. Narciso, G. J. M. Rosa, R. B. Lobo, and F. Baldi, “Improving genomic prediction accuracy for meat tenderness in Nellore cattle using artificial neural networks,” J. Animal Breed. Genet., vol. 137, no. 5, pp. 438–448, 2020, doi: 10.1111/jbg.12468.

[16] P. Waldmann, “Approximate Bayesian neural networks in genomic prediction,” Genet. Selection Evol., vol. 50, no. 1, pp. 1–9, Dec. 2018, doi: 10.1186/s12711-018-0439-1.

[17] H. Okut, X.-L. Wu, G. J. M. Rosa, S. Bauk, B. W. Woodward, R. D. Schnabel, J. F. Taylor, and D. Gianola, “Predicting expected progeny difference for marbling score in angus cattle using artificial neural networks and Bayesian regression models,” Genet. Selection Evol., vol. 45, no. 1, p. 34, Dec. 2013, doi: 10.1186/s12711-012-0034-1.

[18] J. M. González-Camacho, J. Crossa, P. Pérez-Rodríguez, L. Ormella, and D. Gianola, “Genome-enabled prediction using probabilistic neural network classifiers,” BMC Genomics, vol. 17, no. 1, pp. 1–16, Dec. 2016, doi: 10.1186/s12864-016-3553-1.

[19] K. Kizilkaya, 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, Feb. 2010, doi: 10.2527/jas.2009-2004.

[20] S. O. Peters, K. Kizilkaya, D. J. Garrick, R. L. Fernando, J. M. Reecy, R. L. Webster, G. A. Silver, and M. G. Thomas, “Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcase traits in Brangus heifers,” J. Animal Sci., vol. 90, no. 10, pp. 3398–3409, Oct. 2012, doi: 10.2527/jas.2011-4507.
P. J. Werbos, The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting, Hoboken, NJ, USA: Wiley, 1988.

W. G. Hill and A. Robertson, “Linkage disequilibrium in finite populations,” Theor. Appl. Genet., vol. 38, no. 6, pp. 226–231, Jun. 1968. doi: 10.1007/BF01245622.

S. Naderi, T. Yin, and S. König, “Random forest estimation of genomic predictions,” G3, Genes|Genomes|Genet., vol. 8, no. 8, pp. 749–752, Aug. 2018. doi: 10.1534/g3.118.200311.

Z. Zhang, J. Liu, X. Ding, P. Bijjuna, 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, Sep. 2010. Art. no. e12648, doi: 10.1371/journal.pone.0012648.

S. O. Peters et al.: Genomic Prediction With Different Heritability, QTL, and SNP Panel Scenarios Using ANN.
MAHmut SineCen (Member, IEEE) received the M.Sc. degree from Pamukkale University, Denizli, Turkey, and the Ph.D. degree in machine learning from Dokuz Eylul University, Izmir, Turkey. He was an Instructor with the Computer Center, Pamukkale University, from 2005 to 2011. He was an Assistant Professor with the Computer Engineering Department, Canakkale 18 Mart University, from 2011 to 2013. He is currently an Assistant Professor with the Department of Computer Engineering, Faculty of Engineering, Aydin Adnan Menderes University, Aydin, Turkey. His research interests include industrial image processing, software development, fuzzy logic, artificial neural networks, and pattern recognition.

Kadir KizilKayA received the bachelor’s degree in animal science from Ankara University, Ankara, Turkey, the M.Sc. degree from Texas A&M University, and the Ph.D. degree from Michigan State University. He is currently a Professor with the Department of Animal Science, Biometry and Genetics Unit, Aydin Adnan Menderes University, Aydin, Turkey. He completed his Postdoctoral study about developing statistical models for analysis of genomic data with Iowa State University. His research interests include genomic selection and genome-wide association study for livestock animals.

Milton G. Thomas received the B.S. and M.S. degrees in animal science and molecular biology from the University of Missouri, in 1988 and 1990, respectively, and the Ph.D. degree in animal science and molecular biology from Texas A&M University, in 1994. He was with the Center for Molecular Medicine, The University of Texas Institute of Biotecnology, from 1994 to 1996. He progressed through the faculty ranks in the Department of Animal and Range Sciences, New Mexico State University, from 1997 to 2011. He joined Colorado State University, in December 2011, as a Professor and the John E. Rouse Chair of Beef Cattle Breeding and Genetics with the Department of Animal Sciences. He is also an Adjunct Professor with the Animal Reproduction Biotechnology Laboratory, College of Veterinary Medicine and Biomedical Sciences. Within the Breeding and Genetics Team, he brings expertise and effort in the area physiology and omics, genomics (DNA), transcriptomics (RNA), and proteomics (protein). These disciplines are applied to research projects encompassing difficult to measures traits such as health, fertility, and grazing distribution.