PLOS ONE

RESEARCH ARTICLE

Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in *Coffea canephora*

Ithalo Coelho de Sousa[1,2], Moysés Nascimento[2], Isabela de Castro Sant'anna[3], Eveline Teixeira Caixeta[4], Camila Ferreira Azevedo[2], Cosme Damião Cruz[5], Felipe Lopes da Silva[6], Emilly Ruas Alkimim[7], Ana Carolina Campana Nascimento[2], Nick Vergara Lopes Serão[1]*

1 Department of Animal Science, Iowa State University, Ames, Iowa, United States of America, **2** Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **3** Rubber Tree and Agroforestry Systems Research Center, Campinas Agronomy Institute (IAC), Votuporanga, São Paulo, Brazil, **4** Brazilian Agricultural Research Corporation, Embrapa Coffee, Brasília, DF, Brazil, **5** Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **6** Department of Plant Science, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **7** Federal University of Triangulo Mineiro, Iturama, Minas Gerais, Brazil

* serao@iastate.edu

Citation: Coelho de Sousa I, Nascimento M, de Castro Sant'anna I, Teixeira Caixeta E, Ferreira Azevedo C, Damião Cruz C, et al. (2022) Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in *Coffea canephora*. PLoS ONE 17(1): e0262055. https://doi.org/10.1371/journal.pone.0262055

Editor: Muhammad Abdul Rehman Rashid, Government College University Faisalabad, PAKISTAN

Received: August 6, 2021

Accepted: December 15, 2021

Published: January 26, 2022

Data Availability Statement: The data are owned by Brazilian Coffee Breeding Program. Interested researchers may negotiate data sharing agreements with the participating companies (Embrapa, Epamig and Universidade Federal de Viçosa), which can be facilitated by Eveline Caixeta (eveline.caixeta@embrapa.br) or Antonio Carlos Baião de Oliveira (antonio.baiao@embrapa.br).

Abstract

Many methodologies are used to predict the genetic merit in animals and plants, but some of them require priori assumptions that may increase the complexity of the model. Artificial neural network (ANN) has advantage to not require priori assumptions about the relationships between inputs and the output allowing great flexibility to handle different types of complex non-additive effects, such as dominance and epistasis. Despite this advantage, the biological interpretability of ANNs is still limited. The aim of this research was to estimate the heritability and markers effects for two traits in *Coffea canephora* using an additive-dominance architecture ANN and to compare it with genomic best linear unbiased prediction (GBLUP). The data used consists of 51 clones of *C. canephora* varietal Conilon, 32 of varietal group Robusta and 82 intervarietal hybrids. From this, 165 phenotyped individuals were genotyped for 14,387 SNPs. Due to the high computational cost of ANNs, we used Bagging decision tree to reduce the dimensionality of the data, selecting the markers that accumulated 70% of the total importance. An ANN with three hidden layers was run, each varying from 1 to 40 neurons summing 64,000 neural networks. The network architectures with the best predictive ability were selected. The best architectures were composed by 4, 15, and 33 neurons in the first, second and third hidden layers, respectively, for yield, and by 13, 20, and 24 neurons, respectively for rust resistance. The predictive ability was greater when using ANN with three hidden layers than using one hidden layer and GBLUP, with 0.72 and 0.88 for yield and coffee leaf rust resistance, respectively. The concordance rate (CR) of the 10% larger markers effects among the methods varied between 10% and 13.8%, for additive effects and between 5.4% and 11.9% for dominance effects. The narrow-sense (h_a^2) and dominance-only (h_d^2) heritability estimates were 0.25 and 0.06, respectively, for yield, and 0.67 and 0.03, respectively for rust resistance. The ANN was able to estimate the

Funding: This work was financially supported by the Brazilian Coffee Research and Development Consortium (CBP&D/Café), the National Institute of Science and Technology of Coffee (INCT-Café), the Foundation for Research Support of the State of Minas Gerais (FAPEMIG), the National Council of Scientific and Technological Development (CNPq), and the Coordination for the Improvement of Higher Education people (CAPES) - Finance Code 001.

Competing interests: The authors have declared that no competing interests exist.

heritabilities from an additive-dominance genomic architectures and the ANN with three hidden layers obtained best predictive ability when compared with those obtained from GBLUP and ANN with one hidden layer.

Introduction

The interest in semi- and non-parametric statistical methods for genome-enabled prediction is increasing [1]. Methodologies based on machine learning, as Artificial Neural Networks (ANN), has been successfully used to predict the genetic merit in animals [2, 3] and plants [4, 5]. ANN is a methodology inspired by the biological behavior of human brain. ANN comprises layers divided into units called neurons. Each neuron's output is expressed as the sum of inputs to a neuron, regulating specific weights for the predictor variables through linear and nonlinear activation functions [1, 6]. ANN have been applied for genomic prediction of complex traits in some crops as maize, eucalypt [7], soybean [8] and wheat [9]. This approach does not require making a priori assumptions about the relationships between inputs (SNP markers) and the output (phenotypic observations). The non-priori assumptions allow for great flexibility to handle different types of complex non-additive effects, such as dominance and epistasis [1, 10, 11].

Despite this advantage, reports about the biological interpretation from the marker effects and genetic parameter (i.e., heritability) estimates are limited to the best of our knowledge. Glória et al, [1] using simulated data, aimed to evaluate Bayesian regularized ANNs' predictive performance and exploit SNP effects and heritability estimates. Considering only additive effects, the authors observed that based on the predictive ability and estimates of the heritabilities, the best ANN presented similar results to those obtained by Ridge Regression BLUP (RR_BLUP) and Bayesian Lasso (BLASSO).

For some species, for example, maize, eucalyptus, cotton, rice, pinus, and coffee [12–17], where there is commercial interest in hybrids and heterosis, the contribution of dominance presents high importance [16]. Coffee is globally one of the most important export crops and is a part of the economy in more than 50 countries in Latin America, Africa, and Asia. Besides the yield, traits associated with resistance to coffee rust are important in the selection in coffee, since the coffee production can be reduced in the presence of this disease [18]. Therefore, the identification of cultivars having resistance for diseases can improve the productivity of the culture. Despite its relevance, the effective selection of new cultivars depends on the ability to consider genomic models, which correctly represent complex traits with additive and dominance effects. Therefore, methods considering dominance effects, different numbers of layers, and neurons to exploit SNP effects and heritability can bring new insights for genomic selection in coffee.

Against this background, we aimed to exploit SNP effects and heritability from additive-dominance genomic model by ANN of traits associated with the yield and coffee leaf rust resistance, in *Coffea cenephora*. In addition, we predicted the individual genetic merits of the traits (yield and coffee leaf rust resistance) using ANN, and compared the predictive ability obtained for ANN and GBLUP for predicting genetic merit.

Material and methods

Phenotypic data

The used population consisted of 51 clones of *C. canephora* varietal group Conilon, 32 varietal group Robusta and 82 intervarietal hybrids. These hybrids were originated from crosses

between five Conilon genotypes (males) and five Robusta (females), obtained in a partial diallel model [19]. The Conilon genetic material was obtained from the Capixaba Institute for Research, Technical Assistance, and Rural Extension (INCAPER, Vitória, ES, Brazil). The Robusta material was obtained from the Tropical Agronomic Research and Teaching Center (CATIE, Cartago, Turrialba, Costa Rica). This population composes the breeding program of the Agricultural Research Company of Minas Gerais (Epamig, Belo Horizonte, MG, Brazil) in partnership with the Federal University of Viçosa (UFV, Viçosa, Minas Gerais, Brazil) and the Brazilian Agricultural Research Company—Café (Embrapa Café, Oratório, Minas Gerais, Brazil).

Individuals were phenotyped for two traits, coffee leaf rust resistance and yield, for three years (2014 to 2016). Coffee leaf rust resistance (Hemileia vastatrix) was evaluated using a 5-point scale (1 = fully resistant, 5 = highly susceptible). The yield per coffee plant was evaluated by harvesting all fruits present in a genotype and measuring the total volume of freshly harvested coffee liters.

SNP genotyping

DNA samples of 165 young and fully expanded leaves coffee were genotyped using the methodology described by Diniz et al. [20]. The concentration of DNA was verified in NanoDrop 2000, and its quality was evaluated in 1% agarose gel. The sample's DNA concentration was standardized and sent to Rapid Genomics (Florida, Orlando, USA) for identification of SNP molecular markers. The data was genotyped using the Capture Seq methodologie [21], totalizing 14,387 markers.

Marker genotypes were coded according to the effects assumed. For additive effects, homozygous markers containing only alleles with minor frequency, the value is 0. For heterozygous markers, the value is 1, and for homozygous markers containing only alleles with major frequency, the value is 2. For dominant codification, we used 0 for homozygous marker and 1 for heterozygous marker.

Phenotypic data analysis

Prior to genomic analyses, the phenotypic data of both traits were independently adjusted for systematic effects using Selegen REML/BLUP software [22] according to the following statistical model:

$$y = Xu + Tc + Wf + Zm + Qs + Sb + e \quad (1)$$

where y is the observed phenotype; **μ** is the effect of the overall mean in each evaluation year (assumed as fixed effect) added to the general mean; **c** is the dominance effect of combination between the parents Conilon and Robusta (assumed as random effect and distributed as $N \sim I\sigma_c^2$); **f** is the additive effect of combination of the parent Robusta (assumed as random effect and distributed as $N \sim A\sigma_f^2$); ***m*** is the additive effect of combination of the parent Conilon (assumed as random effect and distributed as $N \sim A\sigma_m^2$); ***s*** is the effect of permanent environment of individuals (assumed as random effect and distributed as $N \sim I\sigma_s^2$); ***b*** is the effect of permanent environment of blocks (assumed as random effect and distributed as $N \sim I\sigma_b^2$); ***e*** is the residuals (assumed as random effect and distributed as $N \sim I\sigma_e^2$); and X, T, W, Z, Q, and S are the design matrices for the effects of μ, c, f, m, s, and, b, respectivaly. From this, adjusted phenotypes ($\mathbf{Y}^*$) were calculated as the sum of the estimates of random effects ***c***, ***f***, and ***m***, and the residual, and used for subsequent genomic analyses that were carried out in R [23].

Genomic analyses

Genomic BLUP (GBLUP). The additive dominance model for the REML/GBLUP (restricted maximum likelihood/genomic linear unbiased predictor) method is given by:

$$\boldsymbol{Y}^{*} = \boldsymbol{Xb} + \boldsymbol{Z\mu_a} + \boldsymbol{Z\mu_d} + \boldsymbol{e}, \quad (2)$$

where $\mathbf{Y}^{*}$ is the vector of adjusted phenotypic observations obtained in Eq (1), **b** is the vector of fixed effects, $\boldsymbol{\mu_a}$ is the vector of random of additive marker effects, $\boldsymbol{\mu_d}$ is the vector of random of dominance marker effects, **e** refers to the vector of random errors; and X, Z, are the design matrix. The variance structure is given by:

$$\begin{bmatrix} \mu_a \\ \mu_d \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{G_a}\sigma^2_{\mu_a} & 0 & 0 \\ 0 & \boldsymbol{G_d}\sigma^2_{\mu_d} & 0 \\ 0 & 0 & \boldsymbol{I}\sigma^2_e \end{bmatrix} \right)$$

where $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ are the genomic relationship matrices for the additive and dominance effects, respectively, and $\boldsymbol{I}$ is the identity matrix.

An equivalent model [24] at the marker level is given by

$$\boldsymbol{Y}^{*} = \boldsymbol{Xb} + \boldsymbol{ZUm_a} + \boldsymbol{ZSm_d} + \boldsymbol{e}, \quad (3)$$

where: $\boldsymbol{\mu_a} = \mathbf{Um_a}$; $\mathrm{Var}(\mathbf{Um_a}) = \mathbf{UI}\sigma^2_{\mathrm{m_a}}\mathbf{U'} = \mathbf{UU'}\sigma^2_{\mathrm{m_a}}$;$\boldsymbol{\mu_d} = \boldsymbol{Sm_d}$; $\mathrm{Var}(\mathbf{Sm_d}) = \mathbf{SI}\sigma^2_{m_d}\mathbf{S'} = \mathbf{SS'}\sigma^2_{\mathrm{m_a}}$; **X** is the design matrix for the vector **b** and **Z** is the design matrix for the vectors additive ($\mathbf{m_a}$) and dominance ($\mathbf{m_d}$) marker genetic effects. The variance components associated to these effects are $\sigma^2_{\mathrm{m_a}}$ and $\sigma^2_{m_d}$, respectively. The quantity $\mathbf{m_a}$ in one locus is the allele substitution effect and is given by $m_a = \alpha_i = a_i + (q_i - p_i)d_i$, where p_i and q_i are allelic frequencies and a_i and d_i are the genotypic values for one homozygote and heterozygote, respectively, at locus i. In turn, the quantity m_d can be directly defined as $m_{di} = d_i$. The matrices **U** and **S** are defined based on the values 0, 1 and 2 for the number of one of the alleles at the i^{th} marker locus in a diploid individual. The correct parameterization of **U** and **S** is as follows, according to the marker genotypes at a locus m.

$$\boldsymbol{U} = \begin{cases} MM : 2 - 2p \ \rightarrow 2q \\ Mm : 1 - 2p \ \rightarrow q - p \\ mm : 0 - 2p \ \rightarrow -2p \end{cases}$$

$$\boldsymbol{S} = \begin{cases} MM : 0 \ \rightarrow -2q^2 \\ Mm : 1 \ \rightarrow 2pq \\ mm : 0 \rightarrow -2p^2 \end{cases}$$

The covariance matrix for the additive effects is given by $\boldsymbol{G_a}\sigma^2_a = Var(Um_a) = UU'\sigma^2_{m_a}$, which leads to: $\boldsymbol{G_a} = \boldsymbol{UU'}/(\sigma^2_{m_a}/\ \sigma^2_a) = \boldsymbol{UU'}/\sum_{i=1}^{n}[2p_i(1-p_i)]$, as $\sigma^2_a = \sum_{i=1}^{n}[2p_i(1-p_i)]\sigma^2_{m_a}$. The covariance matrix for the dominance effects is given by $\boldsymbol{G_d} = Var(\boldsymbol{Sm_d})\ \boldsymbol{SS'}\sigma^2_{m_d}$. Thus, $\boldsymbol{G_d}\sigma^2_d = \boldsymbol{SS'}/(\sigma^2_{m_d}/\ \sigma^2_d) = \boldsymbol{SS'}/\sum_{i=1}^{n}[2p_i(1-p_i)]$ as $\sigma^2_d\ = \sum_{i=1}^{n}[2p_i(1-p_i)]\sigma^2_{m_d}$. The additive (i.e., narrow-sense) heritability was calculated as $\hat{h}^2_{\alpha} = \hat{\sigma}^2_{\alpha}/(\hat{\sigma}^2_{\alpha} + \hat{\sigma}^2_d + \hat{\sigma}^2_e)$ and the dominant heritability as $\hat{h}^2_d = \hat{\sigma}^2_d/(\hat{\sigma}^2_{\alpha} + \hat{\sigma}^2_d + \hat{\sigma}^2_e)$. The additive-dominance GBLUP method was fitted using GenomicLand software [25] via REML through mixed model equations.

Artificial neural network. The ANN is composed by a combination of neurons in a single or multiple layers. A vector of real values enters as input in each neuron, with the values 0, 1 and 2, which are computed the weighted average of these values followed by a transformation, then the output of neurons can be directly fed as input into other neurons in the next layer [26].

One of the most common families of architectures for connecting neurons into a network is the feed-forward, which can have multiple layers [27]. This architecture is composed by an input layer (IL), $j = 1,2, \ldots, J$ hidden layers (HL), and an output layer (OL). The IL is composed by n_{il} neurons corresponding to the number of markers, the HL are composed by $n_1, n_2, \ldots, n_j$ neurons respectively, and the OL is composed by n_{ol} neurons representing the output values of the application. In this architecture every neuron of the layer j is connected only to the neurons of the layer j +1 producing matrixes of weights W^i, where the output is generated by a linear combination of the last HL.

As we can see in Fig 1, the output of the neurons in the first HL (HL1) is given by $a_i^{[1]} = f\left(\sum_{t=1}^{P} w_{1t}^{[1]} x_{ti} + b_1\right)$, in the second HL (HL2), the outputs of the neurons is given by a linear combination of the outputs from HL1: $a_i^{[2]} = g\left(\sum_{t=1}^{n_1} w_{1t}^{[2]} a_t^{[1]} + b_2\right)$. The third HL (HL3) output is obtained using the same thoughts we use to obtain those from HL2. Finally, the outputs from the OL is obtained by $y_i = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} a_t^{[3]} + b_4) = y_i = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} a_t^{[2]} + b_3) + b_4) = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} g(\sum_{t=1}^{n_1} w_{1t}^{[2]} a_t^{[1]} + b_2) + b_3) + b_4)$ $= z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} g(\sum_{t=1}^{n_1} w_{1t}^{[2]} f(\sum_{t=1}^{P} w_{1t}^{[1]} x_{ti} + b_1) + b_2) + b_3) + b_4)$.

Once an ANN demands high computational processing, it is necessary the use of methodologies to reduce the dimensionality of the data [28]. The reduction of the markers was made by Bagging decision tree. This procedure is an ensemble methodology consisting of training many decision trees built using a random part of the same original data. The variables that, on average, reduces more the residual sum of squares (RSS) are classified as the most important variables. We selected the variables that accumulated 70% of the total importance and used them in the ANN. The network structure considers 1,302 markers as input for resistance to coffee leaf and 1,086 markers as inputs for yield, three hidden layers, and the output that

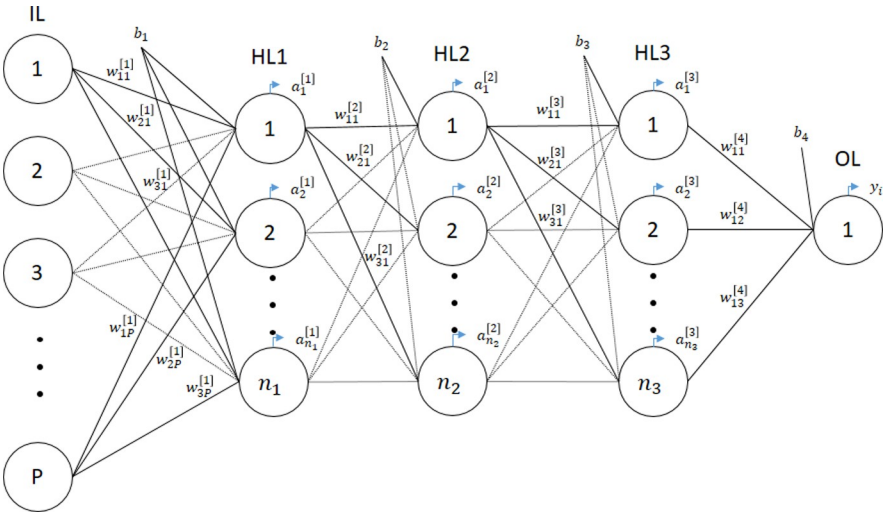

Fig 1. Multilayer perceptron architecture. Feed forward neural network architecture with three hidden layers.

https://doi.org/10.1371/journal.pone.0262055.g001

predicts traits. The ANNs architecture uses the backpropagation as a learning algorithm [29] and the logistic function as activation function. The three hidden layers varied from 1 to 40 neurons, and the architecture was chosen according to the best predictive ability.

To estimate the heritability and SNP effects, the relative importance (RI) of markers were obtained. Olden et al. [30] proposed a methodology that uses all the connection weights even when the ANN has multiple hidden layers to obtain the RI. To calculate the vector of RI of all markers, the connection weights matrices were multiplied. Considering $\boldsymbol{W}^{[i=1]}$ as the matrix of estimated weights connecting the $(j—1)^{th}$ layer to the j^{th} layer where j is the number of layers of the ANN, the RI is obtained multiplying $\boldsymbol{W}^{[j]} * \boldsymbol{W}^{[2]} * \ldots \boldsymbol{W}^{[j-1]}$. To estimate the additive and dominant SNP effect vectors ($\boldsymbol{\beta_a}$ *and* $\boldsymbol{\beta_d}$) using RI, a linear approximation adapted from [31] was used. The estimators are given by $\hat{\boldsymbol{\beta}} = \boldsymbol{ZM}'(\boldsymbol{MZM}')^{-1}\hat{\boldsymbol{y}}$ changing only the codification of the matrix $\boldsymbol{M}$ to obtain the additive or dominant effect, $\boldsymbol{Z}$ is a diagonal matrix composed by the RI values, the matrix $\boldsymbol{M}$ is the matrix of markers and $\hat{y}$ is the genomic estimated breeding values (GEBV) from ANN.

To estimate heritabilities, the additive and dominant variance (σ^2_α and σ^2_d) were estimated using $\hat{\beta}_\alpha$ and $\hat{\beta}_d$ in the following equations: $\hat{\sigma}^2_\alpha = \sum_{j=1}^{P} 2p_j(1-p_j)\hat{\beta}^2_{\alpha_j}$ and $\hat{\sigma}^2_d = \sum_{j=1}^{P}(2p_j(1-p_j))^2\hat{\beta}^2_{d_j}$. The residual variance (σ^2_e) was estimated through the difference of the real phenotype and GEBV, thus $\hat{\sigma}^2_e = Var(\hat{\boldsymbol{e}})$, being $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

Results

The input layer (IL) was composed of a genotype matrix **X** with 165 rows (plants) and 1302 columns (markers) for coffee leaf rust resistance. For yield, the matrix was made up of 165 rows and 1086 markers. The markers were selected using bagging. After reducing dimensionality, 64,000 neural networks were performed, with each hidden layer ranging from 1 to 40 neurons, and the ANN was chosen based on the best predictive ability. For yield, the best ANN has 4, 15, and 33 neurons for the first, second, and third hidden layers, respectively. For coffee leaf rust resistance, the best ANN has 13, 20, and 24 neurons for the first, second, and third hidden layers, respectively. In Fig 2, we can observe the map of each trait with the effects (in absolute terms) of each marker estimated by the ANNs cited above.

The predictive ability mean was calculated (Fig 3) by fixing the number of neurons in one HL and varying the number of neurons in the other. The data showed that in Fig 3, the predictive ability is more affected when we change the number of neurons in the first hidden layer. In the second and the third hidden layers, the average predictive ability does not change significantly as we change the number of neurons.

The chosen ANNs were compared with GBLUP and with the simplest ANN containing one hidden layer with one neuron and the logistic function as activation function according to predictive ability. The most complex ANNs showed a better predictive ability, 0.72 and 0.88 for yield and coffee leaf rust resistance, respectively, indicating that the traits are complex. The ANNs with a single HL with one neuron showed the worse predictive ability, 0.18 and 0.57 for yield and coffee leaf rust resistance, respectively (Fig 4). The ANNs has the ability to capture non-additive effects as dominance and espistasis [1, 10, 11]. It occurs because the interactions between the markers are implicit in the neuron´s outputs.

For both traits, the additive and dominance heritabilities captured by ANN with 3HL (ANN/3HL) were similar to those obtained by GBLUP (Table 1). The ANN with 1HL (ANN/1HL) showed only additive heritability from coffee leaf rust resistance was similar to the other methodologies.

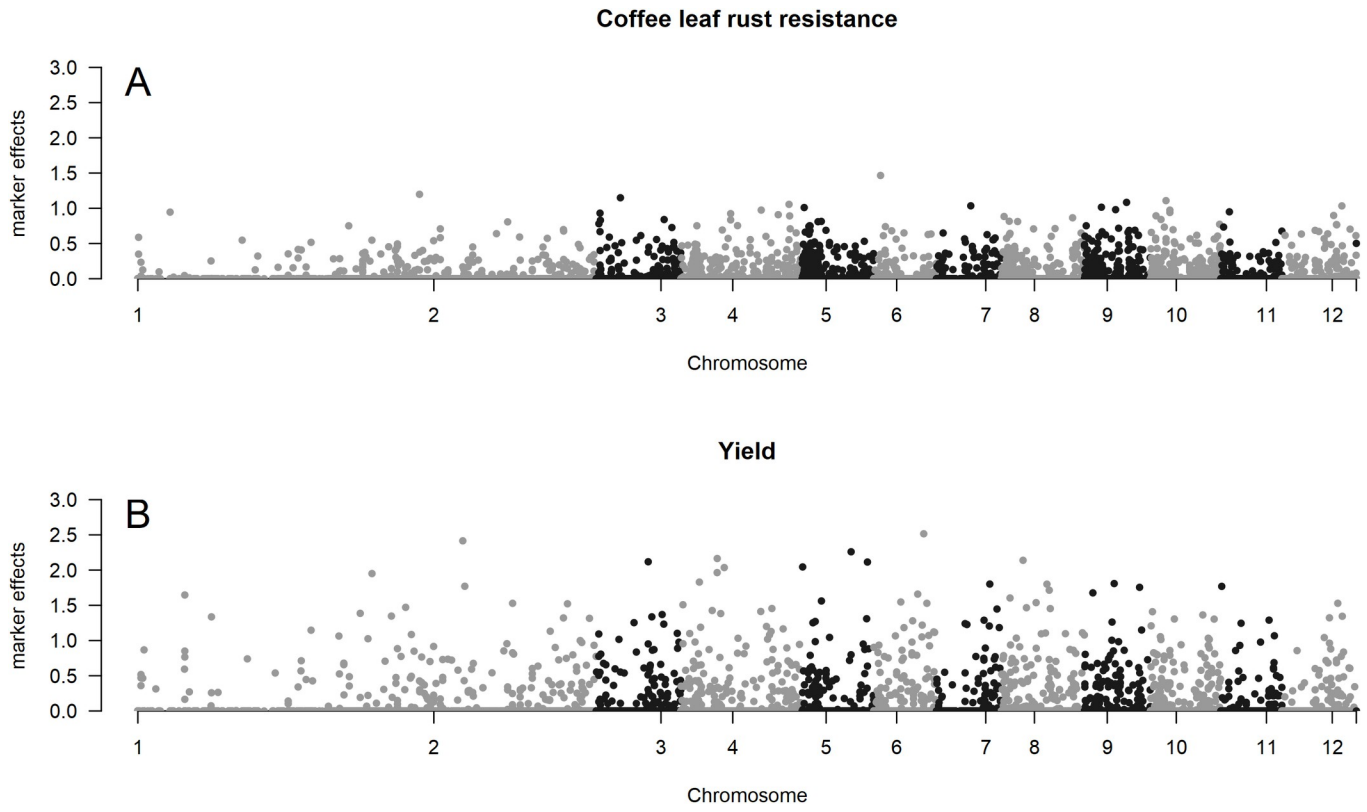

Fig 2. Manhattan plot. A, Manhattan plot showing the effects (in absolute terms) of each marker for coffee leaf rust resistance according to the chromosome position. B, Manhattan plot showing the effects (in absolute terms) of each marker for yield according to the chromosome position.

https://doi.org/10.1371/journal.pone.0262055.g002

The marker effects were estimated using linear approximation [31] based on the method of Olden et al. [30] for ANN. For GBLUP, the marker effects were estimated through a fitted regression model. The absolute values of marker effects from the yield trait are plotted in Fig 5. For this trait, ANN/3HL obtained bigger values than other methodologies evaluated.

The absolute values of marker effects from the coffee leaf rust resistance trait are in Fig 6. For this trait, ANN/1HL obtained bigger values than other methodologies evaluated. In both traits, there is not a strong pattern when comparing the important markers among the methodologies.

Looking at the top 10% larger marker effects in each methodology (Table 2), the concordance rate (CR) among additive marker effects was bigger than dominance marker effects. For the yield trait, the CR between ANN/1HL and GBLUP for additive marker effects was bigger (0.14), and between GBLUP and ANN/1HL for dominance, marker effects were the lowest (0.06). For rust resistance, the biggest CR was between ANN/1HL and GBLUP for the additive marker (0.12), the lowest CR was between GBLUP and ANN/1HL for dominance marker effects (0.05).

Discussion

The use of ANN for predicting the individual genetic merit of plants considering yield and coffee leaf rust resistance in *Coffea canephora* was efficient. The ANN/3HL presented higher values of predictive ability compared with those obtained by GBLUP, a result also obtained by Glória et al. [1], Waldmann [32] and Maldonado [7]. Indeed, the better result was expected

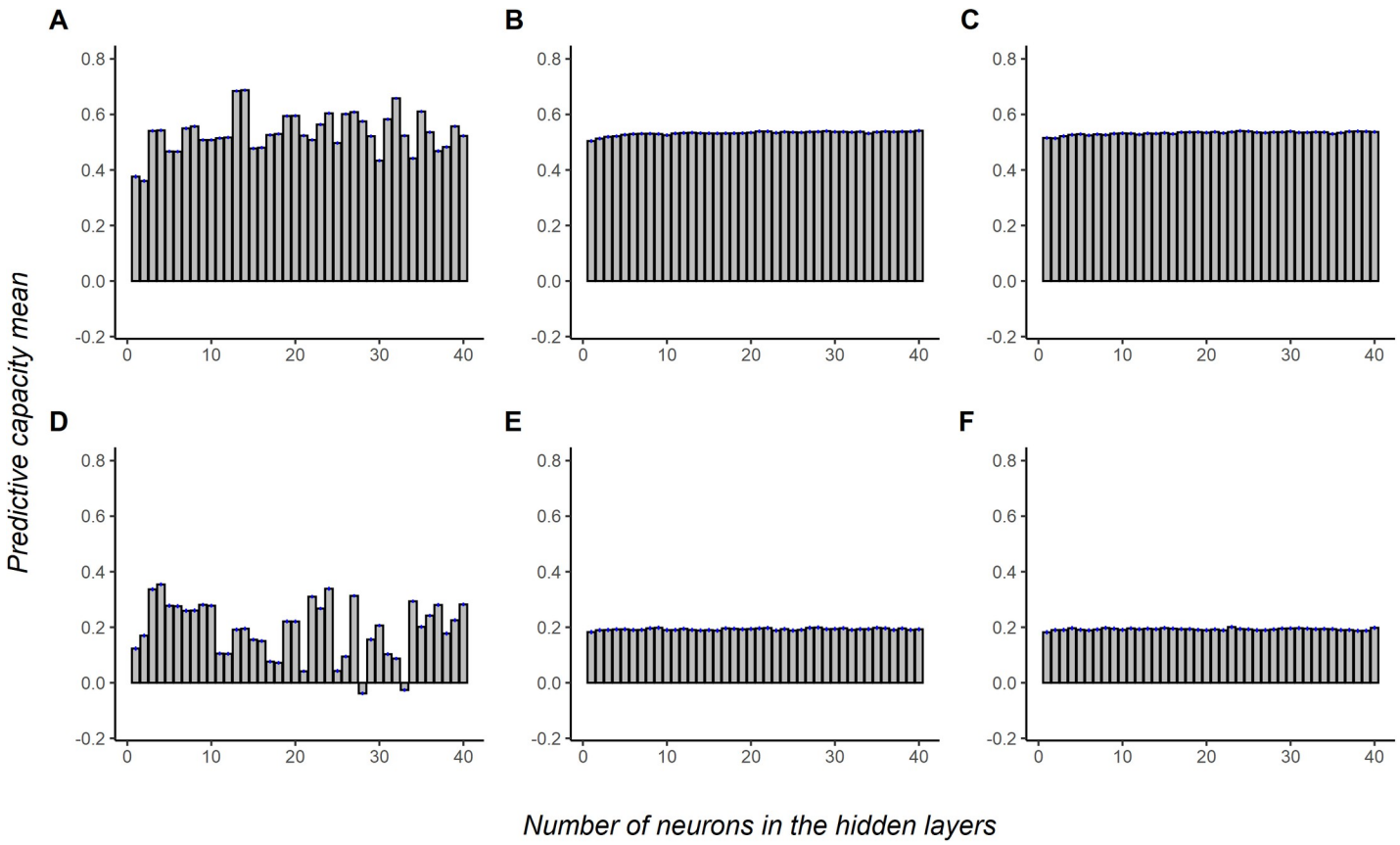

Fig 3. Average predictive capacity of the neural networks according to the numbers of neurons in each hidden layers. A, B, and C are the average predictive capacity when varying the number of neurons in the first, second, and third hidden layers, respectively, for coffee leaf rust resistance in coffee Canephora. D, E, and F are the predictive capacity overage when varying the number of neurons in the first, second, and third hidden layers, respectively, for yield in coffee Canephora.

https://doi.org/10.1371/journal.pone.0262055.g003

since the ANN allows to estimate the functional relationships between the variables using non-linear functions [33]. Thus, the ANN allows great flexibility to handle different types of complex non-additive effects such as dominance and epistasis [34]. The interactions between inputs (SNPs genotypes) and between inputs and the output (phenotypic observations) are naturally modelling from the data. In other words, differently than the traditional methods proposed for genomic selection [11, 35], ANN does not require a priori assumptions about the model relationships allowing to infer the trait architecture directly from the data set [1, 11, 36].

The heritability estimated by ANN/3HL for yield ($h_a^2 = 0.25$; $h_d^2 = 0.06$) and coffee leaf rust resistance ($h_a^2 = 0.67$; $h_d^2 = 0.31$) were similar to those obtained by GBLUP (yield - $h_a^2 = 0.26$; $h_d^2 = 0.05$; coffee leaf rust resistance - $h_a^2 = 0.55$; $h_d^2 = 0.22$). In addition, these estimates were consistent with those reported in the literature. The heritability estimate for yield was within the range of estimates for coffee (0.15–0.79 [37]). For coffee leaf rust resistance, the estimate was close to that reported by Alkimin et al. [37] (0.37).

Glória et al [1] considering only additive effects showed that it is possible to obtain estimates from heritabilities through fitting an ANN composes by one layer, one neuron, and identity activation function. However, for some species, for example maize [38, 39], eucalyptus [40, 41], cotton [42, 43], rice [44, 45], pinus [16, 46] and coffee [47, 48], where there is commercial interest in hybrids, the contribution of dominance presents importance. In fact, an ANN composed by one layer, one neuron, and identity activation function can seem like

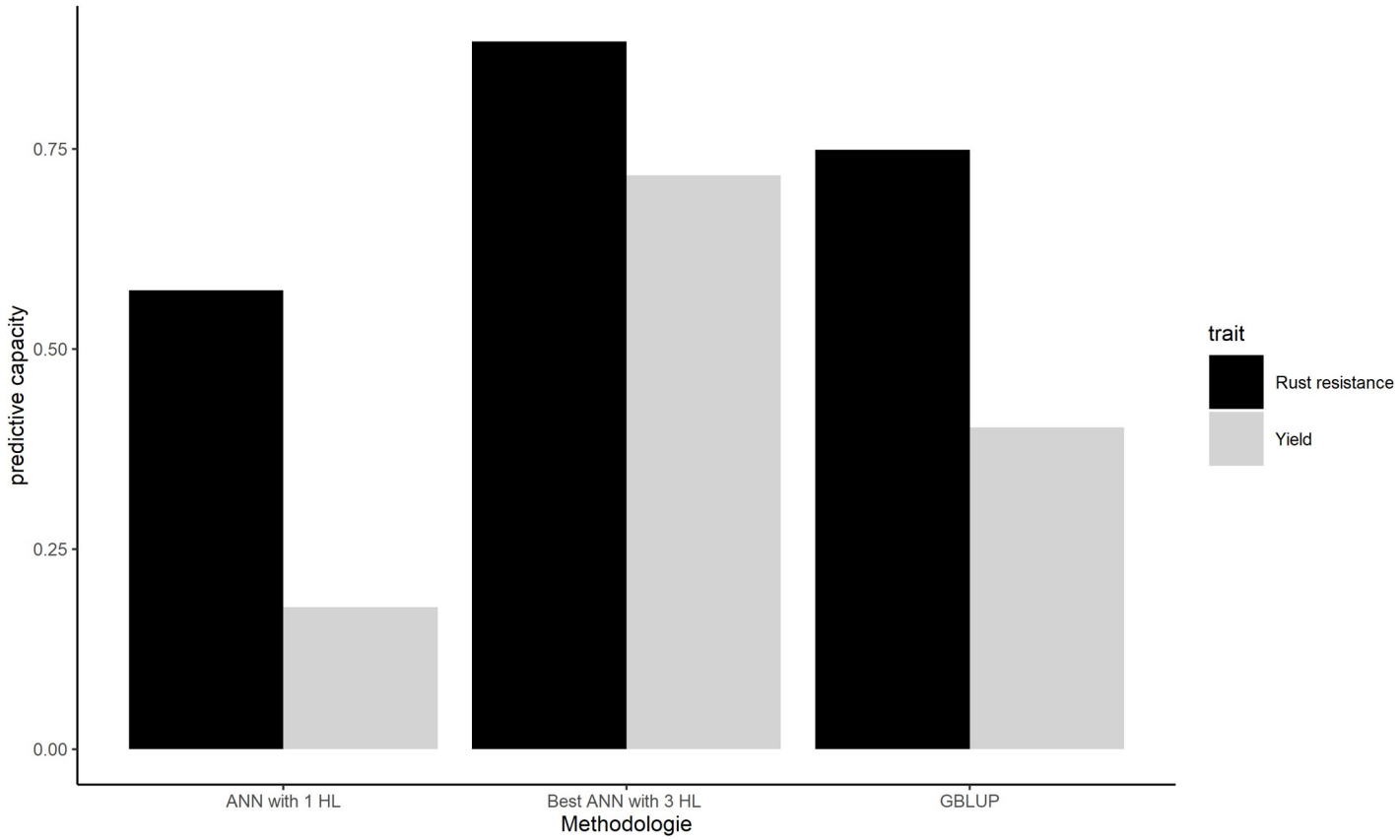

Fig 4. Estimated predictive ability. Yield's estimated predictive ability and coffee leaf rust resistance's estimated predictive ability according to artificial neural network with 1 and 3 hidden layers and Genomic BLUP (GBLUP).

https://doi.org/10.1371/journal.pone.0262055.g004

multiple regression. Differently from [1], the ANN/3HL fitted in this work presents more than one hidden layer, and the activation function is not the identity. Nevertheless, the ANN/3HL was able to obtain heritability estimates similar to those obtained by GBLUP. Therefore, besides increasing the predictive ability, the ANN/3HL allows to access the marker effects and consequently the heritability estimate.

A different pattern in marker effects was obtained in the two traits (Figs 5 and 6). A bigger dominance markers effects were observed for yield when compared with the additive marker effect. In comparison, the additive marker effects were bigger than dominance for coffee leaf rust resistance. This can be explained due yield be a polygenic trait and coffee leaf rust resistance oligogenic. According to Cruz [49], when the trait is polygenic, and there is none or

Table 1. Estimates of additive and dominance heritabilities.

	Yield			Rust resistance		
	ANN/1HL	ANN/3HL	GBLUP	ANN/1HL	ANN/3HL	GBLUP
h_a^2	0.07	0.25	0.26	0.55	0.67	0.55
h_d^2	0.02	0.06	0.05	0.45	0.30	0.22

ANN/1HL, an artificial neural network with one hidden layers; ANN/3HL, an artificial neural network with three hidden layer; GBLUP, genomic best linear unbiased predictor; h_a^2, additive heritability; h_d^2, dominance heritability.

https://doi.org/10.1371/journal.pone.0262055.t001

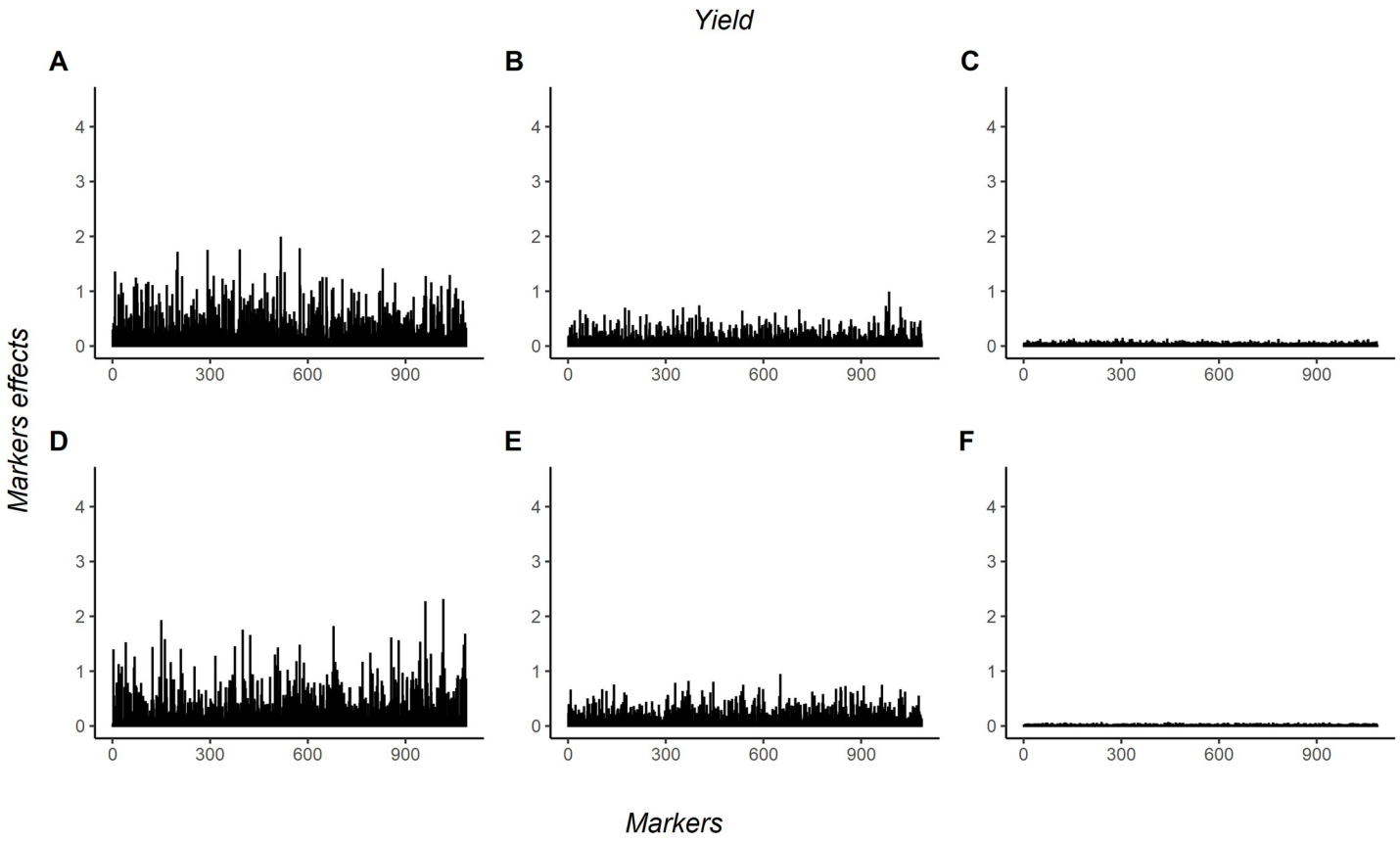

Fig 5. Additive and dominance markers effects for yield in coffee canephora. 1086 markers effects for yield in coffee Canephora. A, B and C are the additive markers effects estimated by a neural network with three hidden layers, a neural network with one hidden layer, and GBLUP, respectively. D, E, and F are the dominance markers effects estimated by a neural network with three hidden layers, a neural network with one hidden layer, and GBLUP, respectively.

https://doi.org/10.1371/journal.pone.0262055.g005

fewer dominance, the phenotype distribution becomes symmetric and starts to obtain asymmetry as the dominance starts to increase. Observing the histogram of both traits (S1 Fig), we see that yield has symmetry distribution and coffee leaf rust resistance an asymmetry distribution.

An issue related to using an ANN approach is the computational cost [50]. Once it is necessary to choose the best network topology, the ANN fitting requires a high computational cost. The ANN/3HL was 409.36 and 1331.49 times slower than GBLUP for yield and coffee leaf rust resistance, respectively. Some approaches can be used to minimize the computational cost. For example, it is possible to reduce the number of inputs of an ANN using some reduction dimensionality methods [51]. Other approaches to select markers used in this work are based on machine learning [52]. Sousa et al. [53] used bagging to select the most important markers. However, since, in general, the number of markers is huge in genomic selection problems, the use of a methodology to reduce the computational cost cannot be effective.

Conclusions

The Artificial Neural Network was able to access the marker effects and heritability estimates from additive-dominance genomic architectures by neural networks in *Coffea canephora*. In addition, considering the estimates of predictive ability, ANN/3HL presented better results compared with those obtained from GBLUP and ANN/1HL.

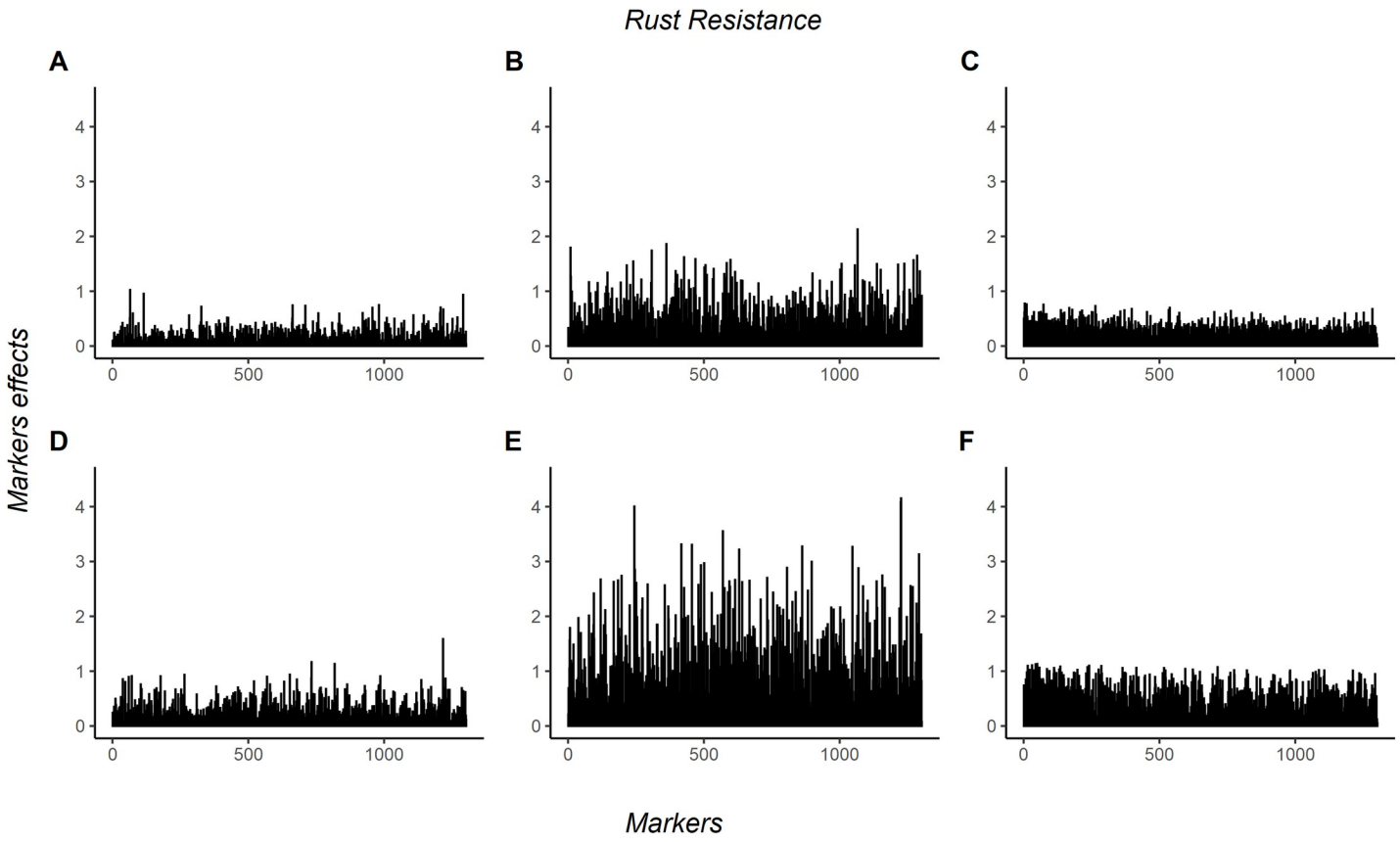

Fig 6. Additive and dominance markers effects for coffee leaf rust resistance in coffee canephora. 1302 markers effects for coffee leaf rust resistance in coffee Canephora. A, B and C are the additive markers effects estimated by neural network with three hidden layers, neural network with one hidden layer, and GBLUP, respectively. D, E, and F are the dominance markers effects estimated by neural network with three hidden layers, neural network with one hidden layer, and GBLUP, respectively.

https://doi.org/10.1371/journal.pone.0262055.g006

Table 2. Concordance of top 10% bigger marker effect among methodologies, in upper triangular matrix refers to additive marker effects, in lower triangular matrix refers to dominance marker effects.

Methodologies	Yield			Rust Resistance		
	ANN/3HL	ANN/1HL	GBLUP	ANN/3HL	ANN/1HL	GBLUP
ANN/3HL	109	12	13	130	13	15
ANN/1HL	13	109	15	14	130	16
GBLUP	8	6	109	11	7	130

ANN/3HL, Artificial neural network with three hidden layers; ANN/1HL, Artificial neural network with one hidden layer; GBLUP, Genomic Best Linear Unbiased Prediction.

https://doi.org/10.1371/journal.pone.0262055.t002

Supporting information

S1 Fig. Histogram. Histogram of yield and rust resistance.
(TIF)

Author Contributions

Conceptualization: Ithalo Coelho de Sousa, Moysés Nascimento, Isabela de Castro Sant'anna.

Data curation: Eveline Teixeira Caixeta, Felipe Lopes da Silva, Emilly Ruas Alkimim.

Formal analysis: Ithalo Coelho de Sousa, Moysés Nascimento, Camila Ferreira Azevedo.

Investigation: Ithalo Coelho de Sousa.

Methodology: Ithalo Coelho de Sousa, Moysés Nascimento, Isabela de Castro Sant'anna, Camila Ferreira Azevedo.

Resources: Ithalo Coelho de Sousa, Moysés Nascimento.

Software: Ithalo Coelho de Sousa, Moysés Nascimento, Camila Ferreira Azevedo.

Supervision: Moysés Nascimento, Isabela de Castro Sant'anna, Camila Ferreira Azevedo, Cosme Damião Cruz, Nick Vergara Lopes Serão.

Writing – original draft: Ithalo Coelho de Sousa, Moysés Nascimento, Isabela de Castro Sant'anna.

Writing – review & editing: Ithalo Coelho de Sousa, Moysés Nascimento, Isabela de Castro Sant'anna, Eveline Teixeira Caixeta, Camila Ferreira Azevedo, Cosme Damião Cruz, Ana Carolina Campana Nascimento, Nick Vergara Lopes Serão.

References

1. Glória LS, Cruz CD, Vieira RAM, de Resende MDV, Lopes PS, de Siqueira OHGBD, et al. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livest Sci. 2016; 191: 91–96. https://doi.org/10.1016/j.livsci.2016.07.015
2. Ehret A, Hochstuhl D, Gianola D, Thaller G. Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. Genet Sel Evol. 2015; 47: 22. https://doi.org/10.1186/s12711-015-0097-5 PMID: 25886037
3. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol. 2020; 52: 1–15. https://doi.org/10.1186/s12711-019-0522-2 PMID: 31941436
4. González-Camacho JM, Crossa J, Pérez-Rodríguez P, Ornella L, Gianola D. Genome-enabled prediction using probabilistic neural network classifiers. BMC Genomics. 2016; 17: 1–16. https://doi.org/10.1186/s12864-015-2294-6 PMID: 26818753
5. Khaki S, Wang L. Crop Yield Prediction Using Deep Neural Networks. 2019; 139–147. https://doi.org/10.1007/978-3-030-30967-1_13
6. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science. Elsevier Ltd; 2017. pp. 961–975. https://doi.org/10.1016/j.tplants.2017.08.011 PMID: 28965742
7. Maldonado C, Mora-Poblete F, Contreras-Soto RI, Ahmar S, Chen JT, do Amaral Júnior AT, et al. Genome-Wide Prediction of Complex Traits in Two Outcrossing Plant Species Through Deep Learning and Bayesian Regularized Neural Network. Front Plant Sci. 2020; 11: 1808. https://doi.org/10.3389/fpls.2020.593897 PMID: 33329658
8. Liu Y, Wang D, He F, Wang J, Joshi T, Xu D. Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. Front Genet. 2019; 10: 1091. https://doi.org/10.3389/fgene.2019.01091 PMID: 31824557
9. Gianola D, Okut H, Weigel KA, Rosa GJM. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. BMC Genet. 2011; 12: 1–14. https://doi.org/10.1186/1471-2156-12-1 PMID: 21205287
10. Felipe VPS, Okut H, Gianola D, Silva MA, Rosa GJM. Effect of genotype imputation on genome-enabled prediction of complex traits: An empirical study with mice data. BMC Genet. 2014; 15: 1–10. https://doi.org/10.1186/1471-2156-15-1 PMID: 24387126
11. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes, Genomes, Genet. 2014; 4: 1027–1046. https://doi.org/10.1534/g3.114.010298 PMID: 24727289

12. Liu R, Wang B, Guo W, Qin Y, Wang L, Zhang Y, et al. Quantitative trait loci mapping for yield and its components by using two immortalized populations of a heterotic hybrid in Gossypium hirsutum L. Mol Breed. 2012; 29: 297–311. https://doi.org/10.1007/s11032-011-9547-0

13. Technow F, Riedelsheimer C, Schrag TA, Melchinger AE. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet. 2012; 125: 1181–1194. https://doi.org/10.1007/s00122-012-1905-8 PMID: 22733443

14. Denis M, Bouvet JM. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet Genomes. 2013; 9: 37–51. https://doi.org/10.1007/s11295-012-0528-1

15. Liang Q, Shang L, Wang Y, Hua J. Partial dominance, overdominance and epistasis as the genetic basis of heterosis in Upland cotton (Gossypium hirsutum L.). PLoS One. 2015;10. https://doi.org/10.1371/journal.pone.0143548 PMID: 26618635

16. De Almeida Filho JE, Guimarães JFR, E Silva FF, De Resende MDV, Muñoz P, Kirst M, et al. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity (Edinb). 2016; 117: 33–41. https://doi.org/10.1038/hdy.2016.23 PMID: 27118156

17. Sousa TV, Caixeta ET, Alkimim ER, Oliveira ACB, Pereira AA, Sakiyama NS, et al. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. Front Plant Sci. 2019; 9: 1934. https://doi.org/10.3389/fpls.2018.01934 PMID: 30671077

18. Alkimim ER, Caixeta ET, Sousa TV, Pereira AA, de Oliveira ACB, Zambolim L, et al. Marker-assisted selection provides arabica coffee with genes from other Coffea species targeting on multiple resistance to rust and coffee berry disease. Mol Breed. 2017; 37: 6. https://doi.org/10.1007/s11032-016-0609-1

19. Alkimim ER, Caixeta ET, Sousa TV, Da Silva FL, Sakiyama NS, Zambolim L. High-throughput targeted genotyping using next-generation sequencing applied in Coffea canephora breeding. Euphytica. 2018; 214: 1–18. https://doi.org/10.1007/s10681-018-2126-2

20. Diniz LEC, Sakiyama NS, Lashermes P, Caixeta ET, Oliveira AC., Zambolim EM, et al. Analysis of AFLP markers associated to the Mex-1 resistance locus in Icatu progenies. Crop Breed Appl Biotechnol. 2005; 5: 387–393.

21. Ruas Alkimim,Eveline Teixeira Caixeta, Tiago Vieira Sousa, Felipe Lopes da Silva, Ney Sussumu Sakiyama, Laércio Zambolim E. High-throughput targeted genotyping using next-generation sequencing applied in Coffea canephora breeding. [cited 21 Jun 2021]. https://doi.org/10.1007/s10681-018-2126-2

22. Resende MDV de. Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breed Appl Biotechnol. 2016; 16: 330–339. https://doi.org/10.1590/1984-70332016v16n4a49

23. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2019. Available: https://www.r-project.org/.

24. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177: 2389–2397. https://doi.org/10.1534/genetics.107.081190 PMID: 18073436

25. Azevedo CF, Nascimento M, Fontes VC, E Silva FF, De Resende MDV, Cruz CD. Genomicland: Software for genome-wide association studies and genomic prediction. Acta Sci—Agron. 2019;41. https://doi.org/10.4025/actasciagron.v41i1.45361

26. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019; 51: 12–18. https://doi.org/10.1038/s41588-018-0295-5 PMID: 30478442

27. Silva IN da, Spatti DH, Flauzino RA. Redes Neurais Artificiais para engenharia e ciências aplicadas. São Paulo: Artliber; 2010.

28. Verleysen M, Francois D, Simon G, Wertz V. On the effects of dimensionality on data analysis with neural networks. In: Mira J, Álvarez JR, editors. Artificial Neural Nets Problem Solving Methods. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. pp. 105–112.

29. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323: 533–536. https://doi.org/10.1038/323533a0

30. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Modell. 2004; 178: 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

31. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 2012; 94: 73–83. https://doi.org/10.1017/S0016672312000274 PMID: 22624567

32. Waldmann P. Approximate Bayesian neural networks in genomic prediction. Genet Sel Evol. 2018; 50: 70. https://doi.org/10.1186/s12711-018-0439-1 PMID: 30577737

33. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, et al. A review of deep learning applications for genomic selection. BMC Genomics. 2021; 22: 1–23. https://doi.org/10.1186/s12864-020-07350-y PMID: 33388042

34. Sant'Anna I de C, Silva GN, Nascimento M, Cruz CD, Sant'Anna I de C, Silva GN, et al. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Sci Agron. 2020; 43: e46307. https://doi.org/10.4025/actasciagron.v43i1.46307

35. Long N, Gianola D, Rosa GJM, Weigel KA. Marker-assisted prediction of non-additive genetic values. Genetica. 2011; 139: 843–854. https://doi.org/10.1007/s10709-011-9588-7 PMID: 21674154

36. Sant'Anna I de C, Nascimento M, Silva GN, Cruz CD, Azevedo CF, Gloria LS, et al. Genome-enabled prediction of genetic values for using radial basis function neural networks. Funct Plant Breed J. 2020;1. Available: http://www.fpbjournal.com/fpbj/index.php/fpbj/article/view/57

37. Alkimim ER, Caixeta ET, Sousa TV, Resende MDV, da Silva FL, Sakiyama NS, et al. Selective efficiency of genome-wide selection in Coffea canephora breeding. Tree Genet Genomes. 2020; 16: 1–11. https://doi.org/10.1007/s11295-020-01433-3

38. Ferrão LF V., Marinho CD, Munoz PR, Resende MFR Jr. Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. Crop Sci. 2020; 60: 666–677. https://doi.org/10.1002/csc2.20096

39. Ramstein GP, Larsson SJ, Cook JP, Edwards JW, Ersoz ES, Flint-Garcia S, et al. Dominance effects and functional enrichments improve prediction of agronomic traits in hybrid maize. Genetics. 2020; 215: 215–230. https://doi.org/10.1534/genetics.120.303025 PMID: 32152047

40. Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, et al. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Edinb). 2017; 119: 245–255. https://doi.org/10.1038/hdy.2017.37 PMID: 28900291

41. Tan B, Grattapaglia D, Wu HX, Ingvarsson PK. Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. Plant Sci. 2018; 267: 84–93. https://doi.org/10.1016/j.plantsci.2017.11.011 PMID: 29362102

42. Shang L, Liang Q, Wang Y, Zhao Y, Wang K, Hua J. Epistasis together with partial dominance, overdominance and QTL by environment interactions contribute to yield heterosis in upland cotton. Theor Appl Genet. 2016; 129: 1429–1446. https://doi.org/10.1007/s00122-016-2714-2 PMID: 27138784

43. Ma L, Wang Y, Ijaz B, Hua J. Cumulative and different genetic effects contributed to yield heterosis using maternal and paternal backcross populations in Upland cotton. Sci Rep. 2019; 9: 3984. https://doi.org/10.1038/s41598-019-40611-9 PMID: 30850683

44. Lin T, Zhou C, Chen G, Yu J, Wu W, Ge Y, et al. Heterosis-associated genes confer high yield in super hybrid rice. Theor Appl Genet. 2020; 133: 3287–3297. https://doi.org/10.1007/s00122-020-03669-y PMID: 32852584

45. Chen L, Bian J, Shi S, Yu J, Khanzada H, Wassan GM, et al. Genetic analysis for the grain number heterosis of a super-hybrid rice WFYT025 combination using RNA-Seq. Rice. 2018; 11: 1–13. https://doi.org/10.1186/s12284-017-0196-8 PMID: 29305728

46. Juranović-Cindrić I, Zeiner M, Starčević A, Liber Z, Rusak G, Idžojtić M, et al. Influence of F1 hybridization on the metal uptake behaviour of pine trees (Pinus nigra x Pinus thunbergiana; Pinus thunbergiana x Pinus nigra). J Trace Elem Med Biol. 2018; 48: 190–195. https://doi.org/10.1016/j.jtemb.2018.04.009 PMID: 29773180

47. Geneti D. Progress of Coffee (Coffea arabica L) Hybridization Development Study in Ethiopia: A Review. 2019;92. https://doi.org/10.7176/FSQM/92-03

48. Geneti D. Review on Heterosis and Combining Ability Study for Yield and Morphological Characters of Coffee (Coffea arabica L) in Ethiopia. 2019;9. https://doi.org/10.7176/JEES/9-12-03

49. Cruz CD. Princípios de Genética Quantitativa. 1st ed. UFV, editor. Viçosa; 2005.

50. de Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Silva FFE, et al. Genomic prediction of leaf rust resistance to arabica coffee using machine learning algorithms. Sci Agric. 2020; 78: 1–8. https://doi.org/10.1590/1678-992x-2020-0021

51. Azevedo C, Nascimento M, Silva F, Resende M, Lopes P, Guimarães S, et al. Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. Genet Mol Res. 2015; 14: 12217–12227. https://doi.org/10.4238/2015.October.9.10 PMID: 26505370

52. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2ed ed. New York: Springer; 2009.

53. de Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Silva FFE, et al. Genomic prediction of leaf rust resistance to arabica coffee using machine learning algorithms. Sci Agric. 2020;78. https://doi.org/10.1590/1678-992x-2020-0021