

**ITHALO COELHO DE SOUSA**

**COMPUTATIONAL INTELLIGENCE AND STATISTICAL LEARNING APPLIED  
TO *Coffea canephora***

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria para o título de *Doctor Scientiae*.

Orientador: Moysés Nascimento

Coorientadores: Ana Carolina Campana  
Nascimento  
Camila Ferreira Azevedo  
Cosme Damião Cruz  
Isabela de Castro Sant'anna

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade  
Federal de Viçosa - Campus Viçosa**

T

S725c  
2022 Sousa, Ithalo Coelho de, 1990-  
Computational intelligence and statistical learning applied  
to *Coffea canephora* / Ithalo Coelho de Sousa. – Viçosa, MG,  
2022.

1 tese eletrônica (58 f.): il. (algumas color.).

Texto em inglês.

Orientador: Moyses Nascimento.

Tese (doutorado) - Universidade Federal de Viçosa,  
Departamento de Estatística, 2022.

Inclui bibliografia.

DOI: <https://doi.org/10.47328/ufvbbt.2022.344>

Modo de acesso: World Wide Web.

1. Marcadores genéticos - Métodos estatísticos.  
2. Aprendizado do computador. 3. Redes neurais (Computação).  
I. Nascimento, Moyses, 1979-. II. Universidade Federal de  
Viçosa. Departamento de Estatística. Programa de  
Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 576.530727


**ITHALO COELHO DE SOUSA**

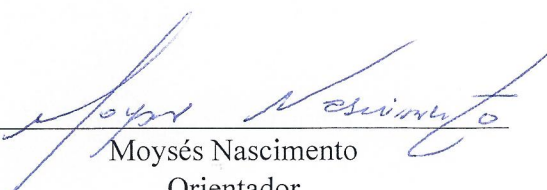
**COMPUTATIONAL INTELLIGENCE AND STATISTICAL LEARNING APPLIED  
TO *Coffea canephora***

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria para o título de *Doctor Scientiae*.

APROVADA: 02 de maio de 2022

Assentimento:

  
Ithalo Coelho de Sousa  
Autor

  
Moysés Nascimento  
Orientador

DEDICO

À minha família.

## AGRADECIMENTOS

Agradeço primeiramente a DEUS por ser meu refúgio em todos os momentos, principalmente nos mais difíceis.

Aos meus pais, Joselita Coelho Barros e Francisco Barbosa de Sousa Neto, pelo carinho, amizade, amor e por não medirem esforços para me dar a melhor educação possível.

À minha irmã Thalyta, pela companhia, conselhos, amizade, apoio e por se fazer próxima mesmo morando longe.

A toda a minha família, incluindo tios, primos e avós, que servem de exemplo para que eu continue determinado e serem compreensivos nos momentos em que não me faço presente, e por saber que estão sempre na torcida e fazem muita falta no meu cotidiano.

Aos meus grandes amigos que considero como irmãos André, Pedro, Franklin, Kaio, Eduardo e Filipe pela amizade e parceria.

Ao meu orientador Moysés Nascimento pela confiança, ensinamentos, paciência, preocupação e pelo incentivo dado. Agradeço também por ser um grande amigo, além de um exemplo como pessoa e como profissional.

Ao Nick VL Serao, por me supervisionar e contribuir com minha formação durante meu doutorado e pela amizade.

Ao Fabyano Fonseca e Silva (*In memoriam*) pelo grande exemplo de profissional, pelos ensinamentos e pela amizade.

Aos meus coorientadores Ana Carolina Campana Nascimento, Camila Ferreira Azevedo, Cosme Damião Cruz e Isabela de Castro Sant'anna por contribuírem no meu aprendizado e pelas sugestões nos trabalhos até aqui realizados.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, que sempre se empenharam e se mostraram acessíveis, dispostos a compartilhar conhecimento e suporte para com os alunos.

Ao Biocafé pela cooperação e por disponibilizar os dados para a realização deste trabalho.

Aos membros da banca por aceitarem o convite e por estarem dispostos a dar suas contribuições para este trabalho.

Aos meus mestres e amigos da Universidade Federal do Piauí, por me ensinarem os primeiros conceitos da estatística.

Aos amigos do LICAE e do laboratório de Bioinformática, pela amizade, ensinamentos e momentos de descontração que tornam o dia a dia mais fácil.

As universidades nas quais tive a oportunidade de estudar nos intercâmbios já realizados, Morgan State University, Southern Illinois University of Carbondale e Iowa State University.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

**MUITO OBRIGADO!**

*“It is a capital mistake to theorize before one has data.”*

(Sir Arthur Conan Doyle)

## RESUMO

SOUSA, Ithalo Coelho de, D.Sc., Universidade Federal de Viçosa, maio de 2022. **Inteligência computacional e aprendizado estatístico aplicados ao *Coffea canephora*.** Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento, Camila Ferreira Azevedo, Cosme Damião Cruz e Isabela de Castro Sant'anna

A predição genômica no melhoramento de café tem mostrado um grande potencial na capacidade preditiva (CP), da predição dos valores genômicos, ganhos genéticos e redução no tempo do ciclo de seleção. Várias metodologias são utilizadas para predizer o mérito genético dos indivíduos, porém algumas metodologias necessitam da informação a priori de efeitos de dominância e epistático, uma vez que seus efeitos devem ser inseridos no modelo utilizado. Redes Neurais Artificiais (RNA) possuem a vantagem de não precisar inserir a priori os efeitos de dominância e epistático, permitindo lidar com diferentes tipos de efeitos não aditivos, sem a necessidade de saber a priori se tais efeitos existem ou não na população estudada. Apesar desta vantagem, a capacidade de estimar parâmetros genéticos através das RNA ainda são limitadas. No presente projeto de pesquisa, duas questões foram formuladas. A primeira questão se trata da possibilidade de estimar parâmetros genéticos utilizando RNA e a segunda questão da possibilidade em reduzir a densidade de painéis de marcadores sem que haja redução na CP. Para responder estas perguntas, foi desenvolvido dois artigos. No primeiro artigo, o objetivo foi estimar a herdabilidade e os efeitos dos marcadores por meio de RNA para duas características morfológicas de interesse agrônomo de café canéfora (produção e resistência à ferrugem) com arquitetura genética aditiva-dominante e comparar com os resultados obtidos por meio do *Genomic Best Linear Unbiased Prediction* (GBLUP). No segundo artigo, o objetivo foi avaliar o equilíbrio entre a densidade dos painéis de marcadores utilizada e a CP obtida para oito características agrônomicas de café canéfora utilizando algoritmos de *Machine Learning* (*bagging* e *Random Forest*). Os dados foram comparados com os resultados obtidos pela metodologia BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*). O conjunto de dados, utilizado em ambos artigos, consiste em 165 plantas de café da espécie *Coffea canephora* (café canéfora) genotipados com 14.387 marcadores SNP (*Single Nucleotide Polymorphisms*), após o controle de qualidade. No primeiro artigo, as duas características fenotípicas avaliadas foram, resistência à ferrugem e produtividade. No segundo artigo, os dados fenotípicos consistem em vigor vegetativo, resistência à ferrugem, incidência de cercosporiose, tempo de maturação do fruto, tamanho do fruto, altura da planta, diâmetro da projeção da copa e produção. No primeiro artigo, a



dimensionalidade dos dados foi reduzida utilizando o *bagging* e em seguida avaliou-se 64.000 redes neurais para cada característica. Foi selecionada a RNA que obteve a maior CP para, para através das informações obtidas por esta RNA se estimar a herdabilidade, obtendo resultados compatíveis com os encontrados na literatura. No segundo artigo, foram utilizados 12 densidade de painéis de marcadores diferentes para avaliar a relação entre a densidade do painel de marcador e a CP. É observado que à medida que o número de marcadores aumenta dentro de um intervalo de 25 até 500/1000 marcadores, a CP também aumenta, no entanto acima dessa quantidade de marcadores, quanto maior for o número de marcadores utilizados menor é a CP obtida. No geral, a CP possui menores valores quando utilizado todos os marcadores. Os resultados indicam que a redução da densidade até um certo nível no painel de marcadores pode melhorar a seleção de indivíduos com um menor custo. Diante do exposto, os métodos de *computational intelligence* provam ser ferramentas poderosas para predição de valores genéticos, estimação de parâmetros genéticos e seleção de marcadores.

Palavras-chave: GBLUP. BLASSO. BAGGING. Random forest. GEBV. Marker effect. Heritability.

## ABSTRACT

SOUSA, Ithalo Coelho de, D.Sc., Universidade Federal de Viçosa, May 2022. **Computational intelligence and statistical learning applied to *Coffea canephora***. Adviser: Moysés Nasciementno. Co-advisers: Ana Carolina Campana Nascimento, Camila Ferreira Azevedo, Cosme Damião Cruz and Isabela de Castro Sant'anna.

Genomic prediction in *Coffea* breeding has shown good potential in predictive ability (PA), genetic gains and reduction of the selection cycle time. Many methodologies are used to predict the genetic merit, 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. In the elaboration of this research project, two basic questions were formulated. The first question, is it possible to estimate genetic parameters using ANNs? The second, is it possible to reduce the panel marker size with no penalty in predictive ability? For this, the analyzes were divided into two articles. In the first article, the aim 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). In the second article, the aim was to evaluate the trade-off between density marker panels size and the PA for eight agronomic traits in *Coffea canephora* using machine learning (bagging and random forest) algorithms and comparing them with BLASSO (Bayesian Least Absolute Shrinkage and Selection Operator) method. For both article, the data set consisted of 165 genotypes of *Coffea canephora* genotyped for 14,387 snp markers, after quality control analysis. For the first article the phenotypic data used was rust (Rus) and yield (Y). For the second article the phenotypic data is composed by vegetative vigor (Vig), rust (Rus) and cercosporiose incidence (Cer), fruit maturation time (Mat), fruit size (FS), plant height (PH), diameter of the canopy projection (DC) and yield (Y). In the first article we reduced the dimensionality of the data using bagging decision tree and then run 64,000 neural networks for each trait selecting the best architecture based on predictive ability for estimating the heritability, obtained results compatibles with those in literature. In the second article, 12 different density market panels were used to evaluate the effect of dimensionality reduction in PA. The common trend observed in the analysis shows an increase of the PA as the number of markers decreases, having a peak in most of the cases when used between 500 and 1,000 markers. In general, the

worst results were obtained when used the full SNP panel density. The results of the second article indicate that the reduction of the number of markers can improve the selection of individuals at a lower cost. Computational Intelligence methods prove to be powerful tools for predicting genetic values, to estimate genetic parameters and to select markers.

**Keywords:** GBLUP. BLASSO. BAGGING. Random forest. GEBV. Marker effect. Heritability.

## SUMÁRIO

1. GENERAL INTRODUCTION .....	13
REFERENCES .....	14
2. RESEARCH PAPER 1: Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in <i>Coffea canephora</i> .....	16
2.1. Abstract .....	17
2.2. Introduction .....	17
2.3. Material and Methods.....	19
2.3.1. Phenotypic data .....	19
2.3.2. SNP genotyping .....	19
2.3.3. Phenotypic data analysis .....	20
2.3.4. Genomic analyses.....	20
2.3.4.1. Genomic BLUP (GBLUP).....	20
The additive dominance model for the REML/GBLUP (restricted maximum likelihood/genomic linear unbiased predictor) method is given by: .....	20
2.3.4.2. Artificial Neural Network.....	22
2.4. Results .....	24
2.5. Discussion.....	30
2.6. Conclusions .....	31
References.....	31
2.7. Supporting information.....	36
3. ARTIGO 2: The trade-off between density marker panels size and predictive ability of genomic prediction for agronomic traits in <i>Coffea canephora</i> .....	37
3.1. Abstract .....	37
3.2. Introduction .....	38
3.3. Materials and Methods.....	38
3.3.1. Genotypes .....	38
3.3.2. Phenotypic evaluations .....	39
3.3.3. Phenotypic data .....	39
3.3.4. SNPs Markers .....	40
3.3.5. Marker selection and Genomic Prediction.....	40
3.3.6. Bayesian Generalized Linear Regression.....	41
3.3.7. Regression tree.....	42
3.3.8. Training and Validation Sets.....	43
3.3.9. Concordance analysis.....	43
3.3.10. Computational Aspects .....	43

3.4.	Results .....	44
3.4.1.	Trait Summary .....	44
3.4.2.	Reduced SNP Panel Densities Improve the Prediction Ability of the traits .....	44
3.4.3.	Comparison among methodologies .....	47
3.5.	Discussion.....	48
3.6.	Conclusions .....	49
	References.....	50
3.7.	Supporting Information.....	53
4.	GENERAL CONCLUSION .....	58

## 1. GENERAL INTRODUCTION

Coffee is the world's favorite drink, the most important commercial crop-plant, and the second most valuable international commodity after oil (KEWSCIENCE, 2022). Brazil is an important player worldwide in the production of coffee, being responsible in 2021/2022 for product 56,300 thousand 60-kilogram bags (35,000 Arabica and 21,300 Robusta) which corresponds to 33,62% of the world production (USDA, 2021).

Coffee production has been increasing with time. The worldwide coffee production increased by 77.04% from 1990/1991 to 2019/2020, and the Brazilian coffee production increased by 113.34% in the same period (INTERNATIONAL COFFEE ORGANIZATION, 2022). One way to help the goal of increasing food production is through plant breeding.

Plant breeding started when the humans stopped to be nomadic, and began domesticate plants which were selected based in visual analysis without any scientific methodology (LUIS CARLOS S BUENO; ANTÔNIO NAZARENO GUIMARÃES MENDES; SAMUEL PEREIRA DE CARVALHO, 2006). With the development of science, the plant breeding began to use plant crossing to create variability and select the best individuals for crossing again in the news cycle. However, the gain selection decreases as long as more cycles are done.

To improve the genetic gain, (MEUWISSEN; HAYES; GODDARD, 2001) proposed using genomic information at the DNA level for selecting the best individuals. It is done through molecular markers for predicting the genomic estimated breeding value (GEBV) and is known as Genomic Wide Selection (GWS). The high-throughput genotyping and the development of Single Nucleotide Polymorphism (SNP) markers helped GWS evolve (BERNARDO; YU, 2007).

GWS has some advantages as to reduces the number of individuals that need to be phenotyped in the selection cycle (FUGERAY-SCARBEL et al., 2021), and accelerates the breeding process (DE SOUSA et al., 2020), however GWS emerged out of a desire to exploit high density panel (MEUWISSEN; HAYES; GODDARD, 2001), and a consequence of it is that more predictors effects, (p), need to be estimated than the number, (n), of available observations (LORENZ et al., 2011).

Due the predictor be bigger than the number of observations, is not possible to estimate the marker effects using least square methodology, besides problems as multicollinearity and high dimensionality. Some methodologies are used to overcome

problems caused by multicollinearity and high dimensionality, e.g., to model the marker effects as random effects, use reduced-dimension regression methods (Partial least regression, principal component regression), but the accuracy gets lower.(SOLBERG et al., 2009)

Computational Intelligence methodologies have been used for predicting GEBV with efficient results (DE SOUSA et al., 2020), but the biological interpretability from marker effects and genetic parameters are limited to the best of our knowledge and has been criticized.(GLÓRIA et al., 2016), estimated marker effects and heritability using a Bayesian regularized artificial neural network but considering only additive effects.

A problem with genotyping is the cost be prohibitive for many species due the high density marker panel used (HAPP et al., 2019; SENTHILVEL et al., 2019; TSAIRIDOU et al., 2020). A strategy to reduce the number of markers is to select a subset of markers that can be done with stepwise regression or machine learning algorithms.

Due to all exposed before, this work aims to estimate marker effects and heritability using Artificial Neural Network considering additive-dominance effects and comparing the results with those obtained by GBLUP. To reduce the costs with genotyping we used bagging and Random Forest to select markers aiming to keep a good predictive ability and comparing the results with those obtained by BLASSO. To all aims above, we used a *Coffea canephora* data set.

## REFERENCES

- BERNARDO, Rex; YU, Jianming. Prospects for Genomewide Selection for Quantitative Traits in Maize. **Crop Science**, [S. l.], v. 47, n. 3, p. 1082–1090, 2007. DOI: 10.2135/CROPSCI2006.11.0690. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2135/cropsci2006.11.0690>. Acesso em: 14 abr. 2022.
- DE SOUSA, I. C. et al. Genomic prediction of leaf rust resistance to arabica coffee using machine learning algorithms. **Scientia Agricola**, [S. l.], v. 78, n. 4, 2020. DOI: 10.1590/1678-992x-2020-0021.
- FUGERAY-SCARBEL, Aline; BASTIEN, Catherine; DUPONT-NIVET, Mathilde; LEMARIÉ, Stéphane. Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience. **Frontiers in Genetics**, [S. l.], v. 12, p. 1185, 2021. DOI: 10.3389/FGENE.2021.629737/BIBTEX. Acesso em: 14 abr. 2022.
- GLÓRIA, Leonardo Siqueira; CRUZ, Cosme Damião; VIEIRA, Ricardo Augusto Mendonça; DE RESENDE, Marcos Deon Vilela; LOPES, Paulo Sávio; DE SIQUEIRA, Otávio H. G. B. Dias; FONSECA E SILVA, Fabyano. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, [S. l.], v. 191, p. 91–96, 2016. DOI: 10.1016/j.livsci.2016.07.015. Acesso em: 19 out. 2020.

HAPP, Mary M.; WANG, Haichuan; GRAEF, George L.; HYTEN, David L. Generating High Density, Low Cost Genotype Data in Soybean [ *Glycine max* (L.) Merr.]. **G3 (Bethesda, Md.)**, [S. l.], v. 9, n. 7, p. 2153–2160, 2019. DOI: 10.1534/G3.119.400093. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/31072870/>. Acesso em: 16 abr. 2022.

INTERNATIONAL COFFEE ORGANIZATION. **Historical Data on the Global Coffee Trade**. 2022.

KEWSCIENCE. **Plants of the World online: Coffea arabica L**. 2022.

LORENZ, Aaron J.; CHAO, Shiaoman; ASORO, Franco G.; HEFFNER, Elliot L.; HAYASHI, Takeshi; IWATA, Hiroyoshi; SMITH, Kevin P.; SORRELLS, Mark E.; JANNINK, Jean Luc. Genomic Selection in Plant Breeding: Knowledge and Prospects. **Advances in Agronomy**, [S. l.], v. 110, n. C, p. 77–123, 2011. DOI: 10.1016/B978-0-12-385531-2.00002-5. Acesso em: 14 abr. 2022.

LUIS CARLOS S BUENO; ANTÔNIO NAZARENO GUIMARÃES MENDES; SAMUEL PEREIRA DE CARVALHO. **Melhoramento Genético de Plantas: Princípios e Procedimentos**. 2. ed. Lavras: Editora Lavras, 2006.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, [S. l.], v. 157, n. 4, p. 1819–1829, 2001. DOI: 10.1093/GENETICS/157.4.1819. Disponível em: <https://academic.oup.com/genetics/article/157/4/1819/6048353>. Acesso em: 11 abr. 2022.

SENTHILVEL, S.; GHOSH, Arpita; SHAIK, Mobeen; SHAW, Ranjan K.; BAGALI, Prashanth G. Development and validation of an SNP genotyping array and construction of a high-density linkage map in castor. **Scientific Reports 2019 9:1**, [S. l.], v. 9, n. 1, p. 1–10, 2019. DOI: 10.1038/s41598-019-39967-9. Disponível em: <https://www.nature.com/articles/s41598-019-39967-9>. Acesso em: 17 jan. 2022.

SOLBERG, Trygve R.; SONESSON, Anna K.; WOOLLIAMS, John A.; MEUWISSEN, Theo He. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, [S. l.], v. 41, n. 1, p. 1–8, 2009. DOI: 10.1186/1297-9686-41-29/TABLES/4. Disponível em: <https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-41-29>. Acesso em: 15 abr. 2022.

TSAIRIDOU, Smaragda; HAMILTON, Alastair; ROBLEDO, Diego; BRON, James E.; HOUSTON, Ross D. Optimizing Low-Cost Genotyping and Imputation Strategies for Genomic Selection in Atlantic Salmon. **G3 Genes | Genomes | Genetics**, [S. l.], v. 10, n. 2, p. 581–590, 2020. DOI: 10.1534/G3.119.400800. Disponível em: <https://academic.oup.com/g3journal/article/10/2/581/6026251>. Acesso em: 17 jan. 2022.

USDA. **Coffee: World Markets and Trade**. [s.l.: s.n.].

WHO. **Sustainable Development Goals (SDGs)**. 2022.

WORLD RESOURCES INSTITUTE. **Creating a sustainable food future: a menu of solutions to feed nearly 10 billion people by 2050**. [s.l.: s.n.].



## **2. RESEARCH PAPER 1: Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in *Coffea canephora***

PlosOne: Doi: 10.1371/journal.pone.0262055

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

<sup>1</sup>Department of Animal Science, Iowa State University, Ames, Iowa, USA

<sup>2</sup>Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

<sup>3</sup>Rubber Tree and Agroforestry Systems Research Center, Campinas Agronomy Institute (IAC), Votuporanga, São Paulo, Brazil

<sup>4</sup>Brazilian Agricultural Research Corporation, Embrapa Coffee, Brasília, DF- Brazil.

<sup>5</sup>Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

<sup>6</sup>Department of Plant Science, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

<sup>7</sup>Federal University of Triangulo Mineiro, Iturama, Minas Gerais, Brazil

\*Corresponding author

E-mail: serao@iastate.edu (NVLS)

## 2.1. 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 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.

## 2.2. Introduction

The interest in semi- and non-parametric statistical methods for genome-enabled prediction is increasing [1]. Methodologies based on computational intelligence, 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 canephora*. 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.

## **2.3. Material and Methods**

### **2.3.1. 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.

### **2.3.2. 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 methodology [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.

### 2.3.3. 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;  $\mu$  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  $\mu, c, f, m, s,$  and  $b$ , respectively. From this, adjusted phenotypes ( $Y^*$ ) 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].

### 2.3.4. Genomic analyses

#### 2.3.4.1. Genomic BLUP (GBLUP)

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

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

where  $Y^*$  is the vector of adjusted phenotypic observations obtained in Eq. (1),  $b$  is the vector of fixed effects,  $\mu_a$  is the vector of random of additive marker effects,  $\mu_d$  is the vector of

random of dominance marker effects,  $\mathbf{e}$  refers to the vector of random errors; and  $\mathbf{X}$ ,  $\mathbf{Z}$ , are the design matrix. The variance structure is given by:

$$\begin{bmatrix} \mu_a \\ \mu_d \\ \mathbf{e} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G}_a \sigma_{\mu_a}^2 & 0 & 0 \\ 0 & \mathbf{G}_d \sigma_{\mu_d}^2 & 0 \\ 0 & 0 & \mathbf{I} \sigma_e^2 \end{bmatrix} \right)$$

where  $\mathbf{G}_a$  and  $\mathbf{G}_d$  are the genomic relationship matrices for the additive and dominance effects, respectively, and  $\mathbf{I}$  is the identity matrix.

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

$$\mathbf{Y}^* = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{U}\mathbf{m}_a + \mathbf{Z}\mathbf{S}\mathbf{m}_d + \mathbf{e}, \quad (3)$$

where:  $\boldsymbol{\mu}_a = \mathbf{U}\mathbf{m}_a$ ;  $\text{Var}(\mathbf{U}\mathbf{m}_a) = \mathbf{U}\mathbf{I}\sigma_{m_a}^2\mathbf{U}' = \mathbf{U}\mathbf{U}'\sigma_{m_a}^2$ ;  $\boldsymbol{\mu}_d = \mathbf{S}\mathbf{m}_d$ ;  $\text{Var}(\mathbf{S}\mathbf{m}_d) = \mathbf{S}\mathbf{I}\sigma_{m_d}^2\mathbf{S}' = \mathbf{S}\mathbf{S}'\sigma_{m_d}^2$ ;  $\mathbf{X}$  is the design matrix for the vector  $\mathbf{b}$  and  $\mathbf{Z}$  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  $\sigma_{m_a}^2$  and  $\sigma_{m_d}^2$ , respectively. The quantity  $m_a$  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  $\mathbf{U}$  and  $\mathbf{S}$  are defined based on the values 0, 1 and 2 for the number of one of the alleles at the  $i^{\text{th}}$  marker locus in a diploid individual. The correct parameterization of  $\mathbf{U}$  and  $\mathbf{S}$  is as follows, according to the marker genotypes at a locus  $m$ .

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

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

The covariance matrix for the additive effects is given by  $\mathbf{G}_a \sigma_a^2 = \text{Var}(\mathbf{U}\mathbf{m}_a) = \mathbf{U}\mathbf{U}'\sigma_{m_a}^2$ , which leads to:  $\mathbf{G}_a = \mathbf{U}\mathbf{U}' / (\sigma_{m_a}^2 / \sigma_a^2) = \mathbf{U}\mathbf{U}' / \sum_{i=1}^n [2p_i(1 - p_i)]$ , as  $\sigma_a^2 = \sum_{i=1}^n [2p_i(1 - p_i)]\sigma_{m_a}^2$ . The covariance matrix for the dominance effects is given by  $\mathbf{G}_d =$

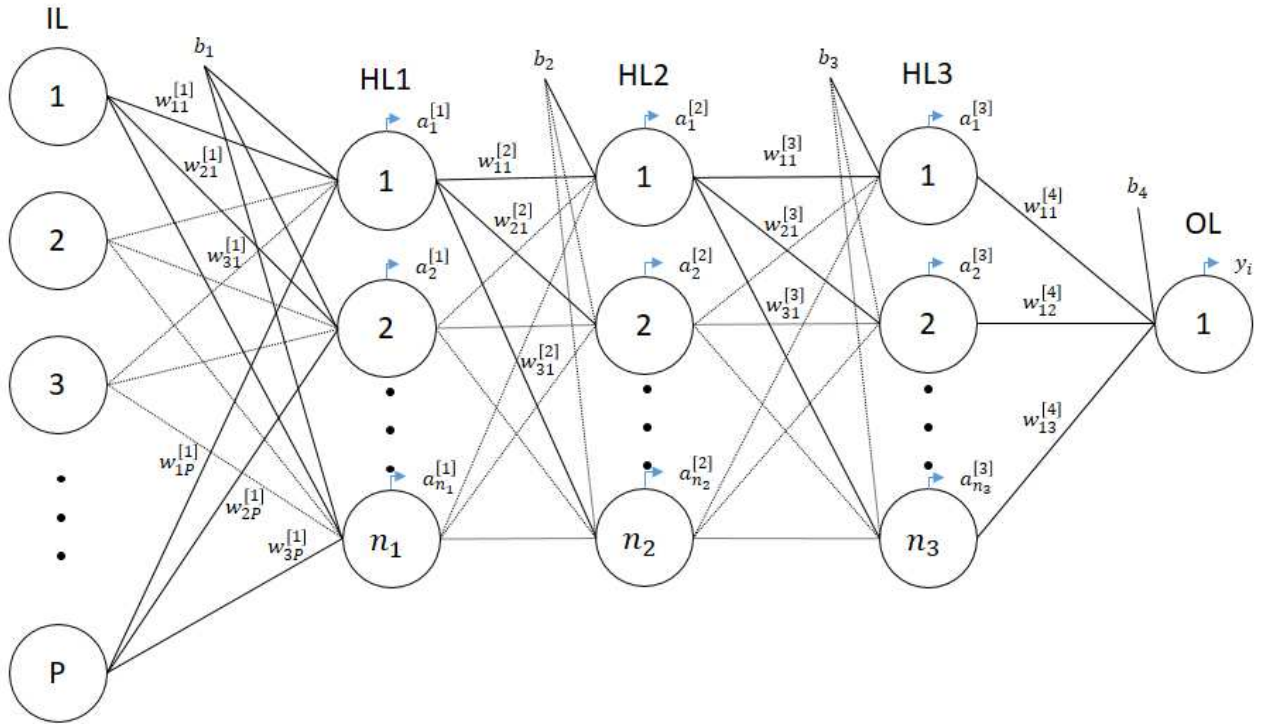
$Var(\mathbf{S}m_d) \mathbf{S}\mathbf{S}'\sigma_{m_d}^2$ . Thus,  $\mathbf{G}_d\sigma_d^2 = \mathbf{S}\mathbf{S}'/(\sigma_{m_d}^2/\sigma_d^2) = \mathbf{S}\mathbf{S}'/\sum_{i=1}^n[2p_i(1-p_i)]$  as  $\sigma_d^2 = \sum_{i=1}^n[2p_i(1-p_i)]\sigma_{m_d}^2$ . The additive (i.e., narrow-sense) heritability was calculated as  $\hat{h}_\alpha^2 = \hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2)$  and the dominant heritability as  $\hat{h}_d^2 = \hat{\sigma}_d^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2)$ . The additive-dominance GBLUP method was fitted using GenomicLand software [25] via REML through mixed model equations.

### 2.3.4.2. 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),  $j = 1, 2, \dots, J$  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  $n_1, n_2, \dots, n_j$  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  $\mathbf{W}^i$ , 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  $a_i^{[1]} = f(\sum_{t=1}^P w_{1t}^{[1]} x_{ti} + b_1)$ , in the second HL (HL2), the outputs of the neurons is given by a linear combination of the outputs from HL1:  $a_i^{[2]} = g(\sum_{t=1}^{n_1} w_{1t}^{[2]} a_t^{[1]} + b_2)$ . 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  $y_i = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} a_t^{[3]} + b_4) = y_i = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} a_t^{[2]} + b_3) + b_4) = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} g(\sum_{t=1}^{n_1} w_{1t}^{[2]} a_t^{[1]} + b_2) + b_3) + b_4) = z(\sum_{t=1}^{n_3} w_{1t}^{[4]} h(\sum_{t=1}^{n_2} w_{1t}^{[3]} g(\sum_{t=1}^{n_1} w_{1t}^{[2]} f(\sum_{t=1}^P w_{1t}^{[1]} x_{ti} + b_1) + b_2) + b_3) + b_4)$ .



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

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 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

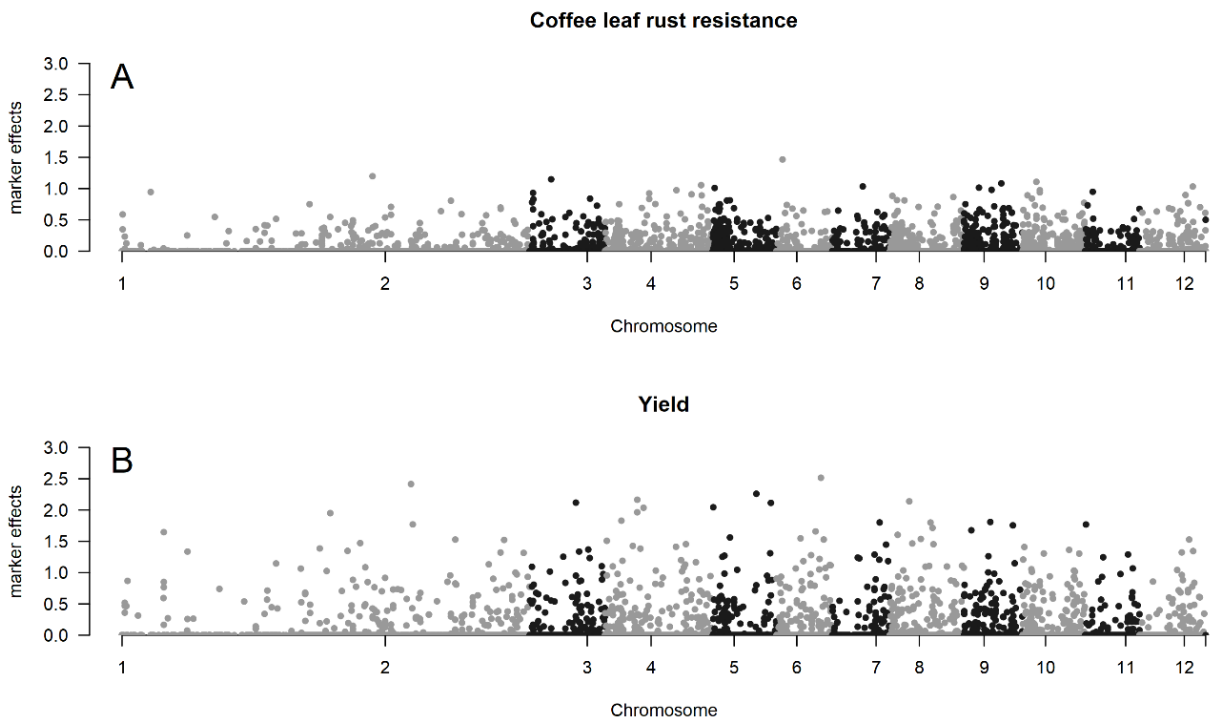


even when the ANN has multiple hidden layers to obtain the RI. To calculate the vector of RI of all markers, the connection weights matrices were multiplied. Considering  $\mathbf{W}^{[i-1]}$  as the matrix of estimated weights connecting the  $(j-1)^{th}$  layer to the  $j^{th}$  layer where  $j$  is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

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

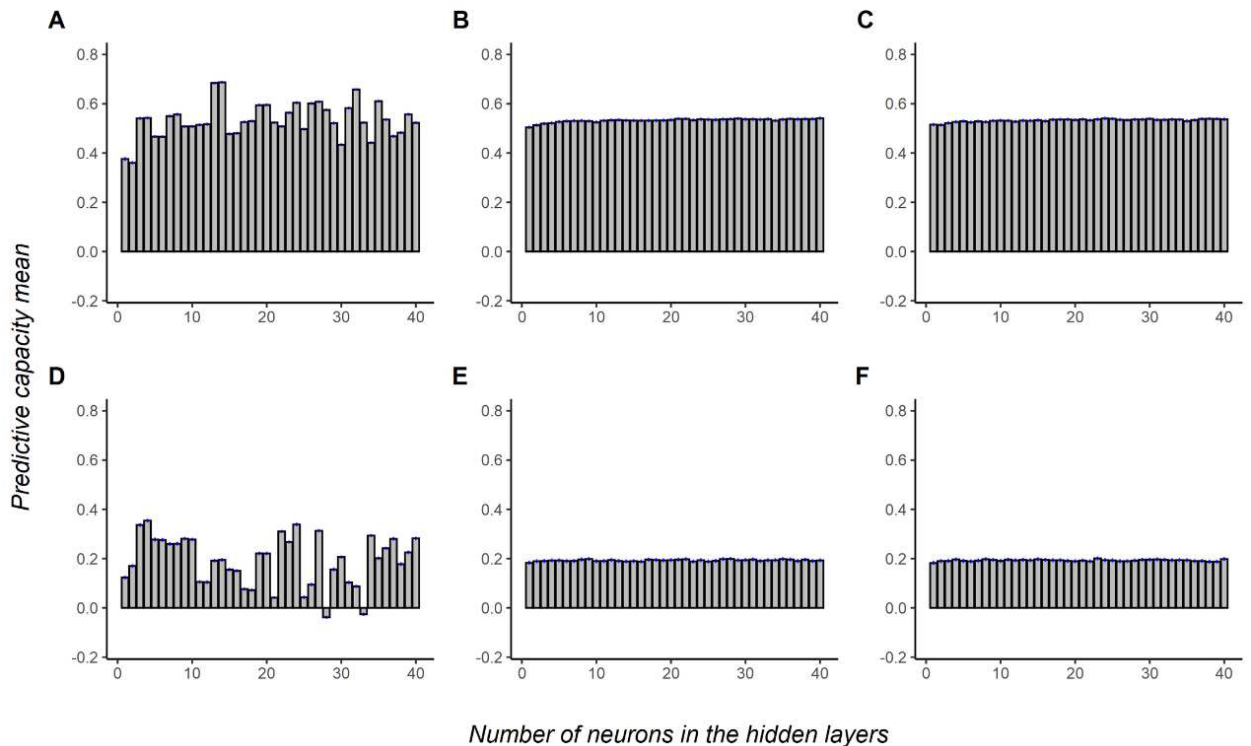
## 2.4. Results

The input layer (IL) was composed of a genotype matrix  $\mathbf{X}$  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 Figure 2, we can observe the map of each trait with the effects (in absolute terms) of each marker estimated by the ANNs cited above.



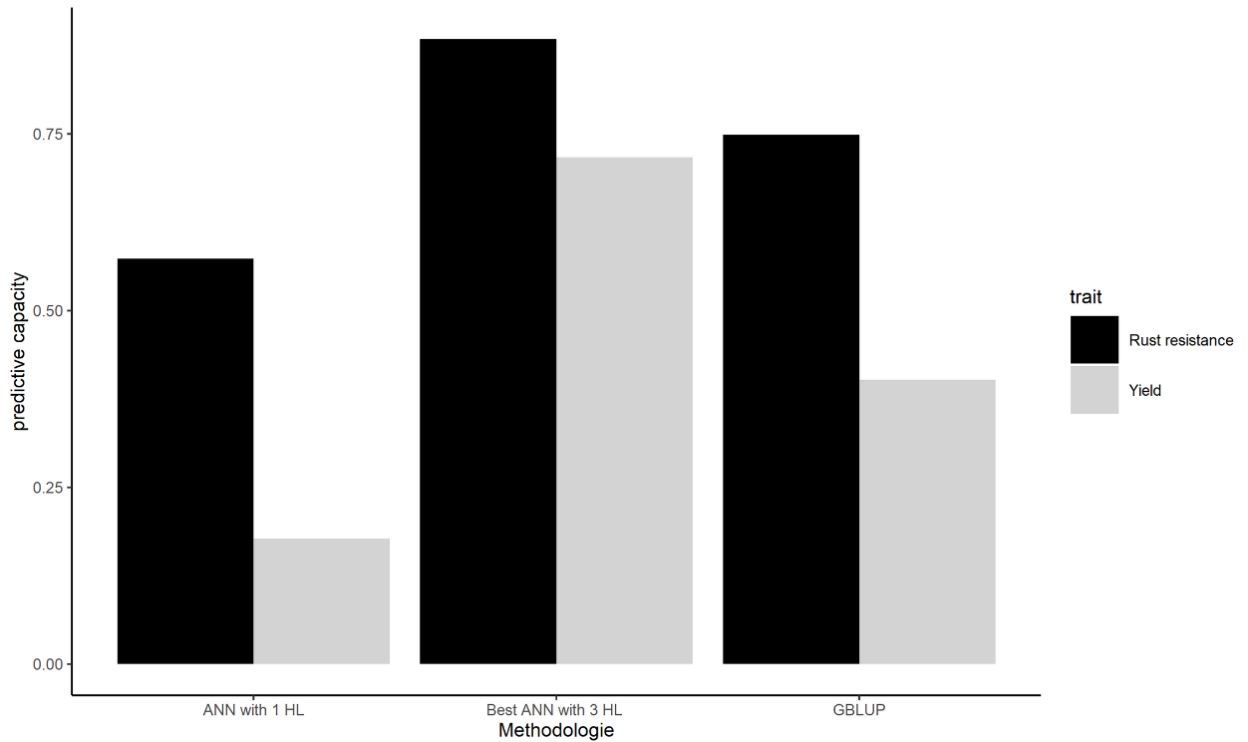
**Fig 2. Manhattan plot.** A, Manhattan plot showing the effects (in absolute terms) of each marker for coffee leaf rust resistance according to the chromosome position. B, Manhattan plot showing the effects (in absolute terms) of each marker for yield according to the chromosome position.

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.



**Fig 3. Average predictive capacity of the neural networks according to the numbers of neurons in each hidden layer.** 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 average when varying the number of neurons in the first, second, and third hidden layers, respectively, for yield in coffee *Canephora*.

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 epistasis [1,10,32]. It occurs because the interactions between the markers are implicit in the neuron's outputs.



**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).

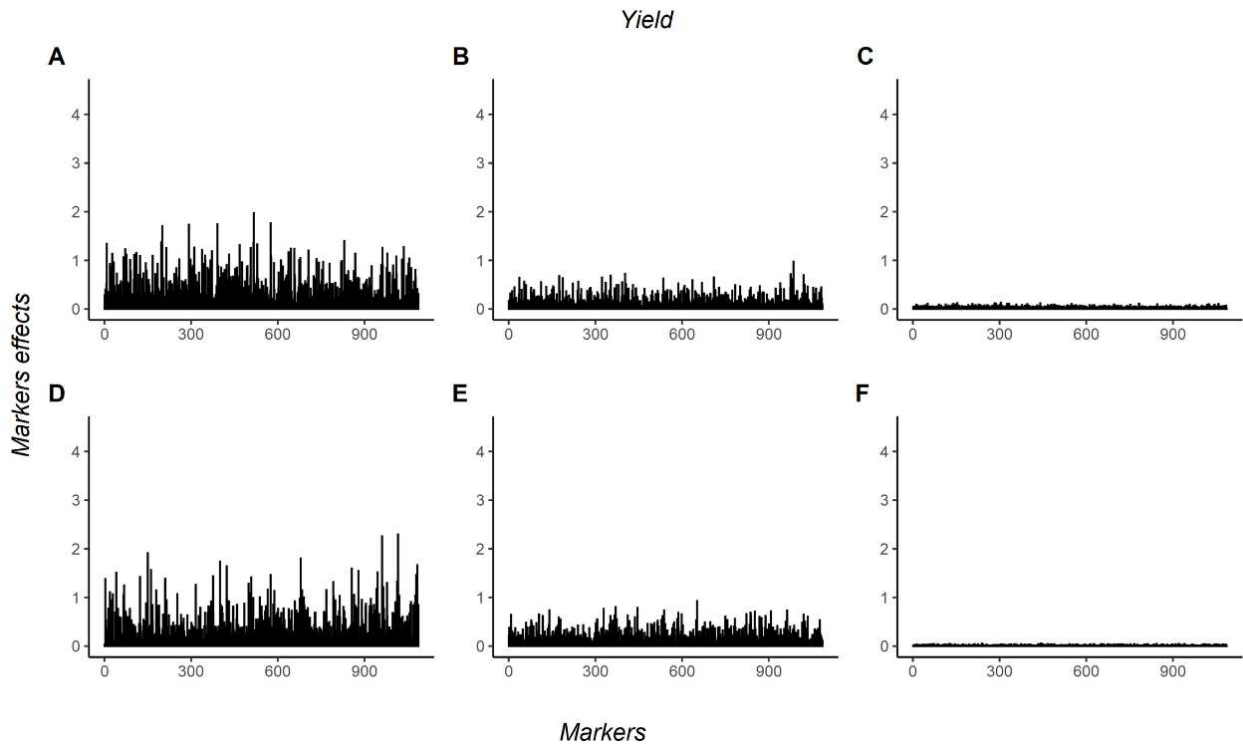
For both traits, the additive and dominance heritabilities captured by ANN with 3HL (ANN/3HL) were similar to those obtained by GBLUP (Table 1). The ANN with 1HL (ANN/1HL) showed only additive heritability from coffee leaf rust resistance was similar to the other methodologies.

**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

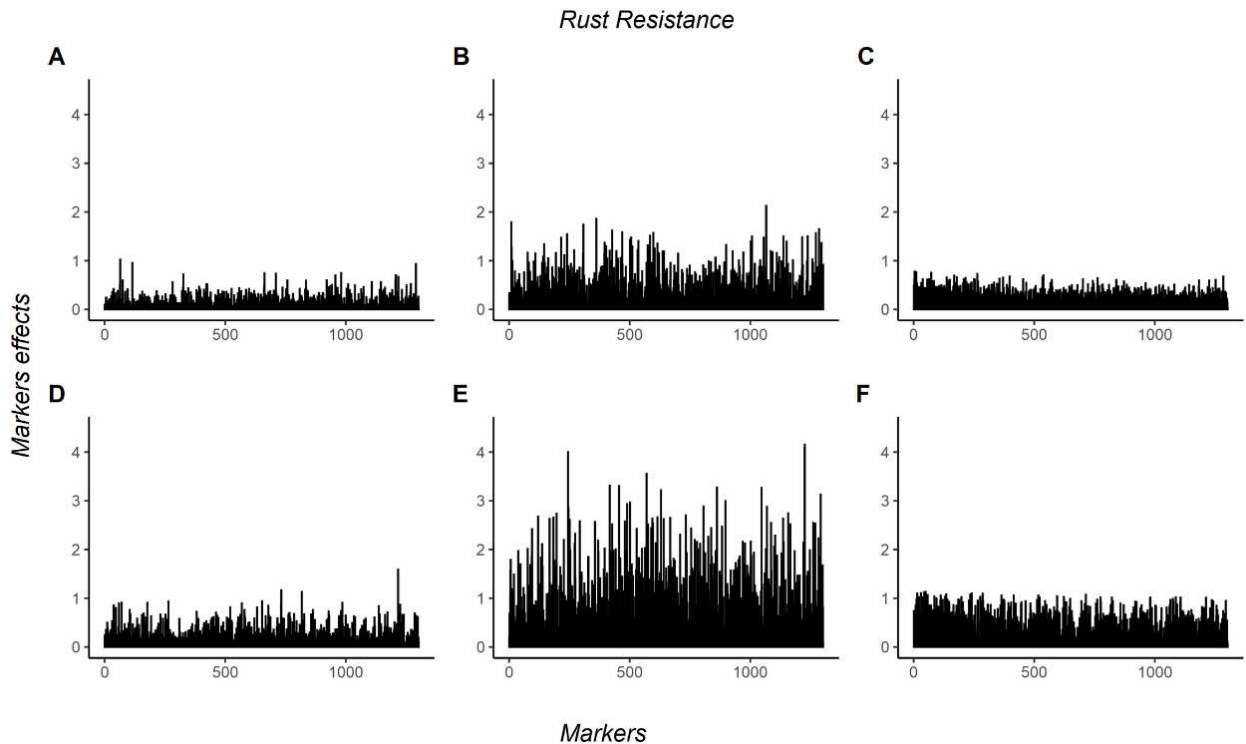
ANN/1HL, an artificial neural network with one hidden layers; ANN/3HL, an artificial neural network with three hidden layer; GBLUP, genomic best linear unbiased predictor;  $h_a^2$ , additive heritability;  $h_d^2$ , dominance heritability.

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.



**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.

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.



**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.

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).

**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.

## 2.5. 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 [33] and Maldonado [7]. Indeed, the better result was expected since the ANN allows to estimate the functional relationships between the variables using nonlinear functions [34]. Thus, the ANN allows great flexibility to handle different types of complex non-additive effects such as dominance and epistasis [35]. 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 [11,36], ANN does not require a priori assumptions about the model relationships allowing to infer the trait architecture directly from the data set [1,11,37].

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[38]). For coffee leaf rust resistance, the estimate was close to that reported by Alkimin et al.[38] (0.37).

Glória et al [1] considering only additive effects showed that it is possible to obtain estimates from heritabilities through fitting an ANN composed by one layer, one neuron, and identity activation function. However, for some species, for example maize [39,40], eucalyptus [41,42], cotton [43,44], rice [45,46], pinus [47,48] and coffee [49,50], 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 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 [51], when the trait is polygenic, and there is none or 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 [52]. 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 [53]. Other approaches to select markers used in this work are based on machine learning [54]. Sousa et al. [55] 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.

## 2.6. 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.

## 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. *Livestock Science*. 2016;191: 91–96. doi: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. *Genetics Selection Evolution*. 2015;47: 22. doi:10.1186/s12711-015-0097-5



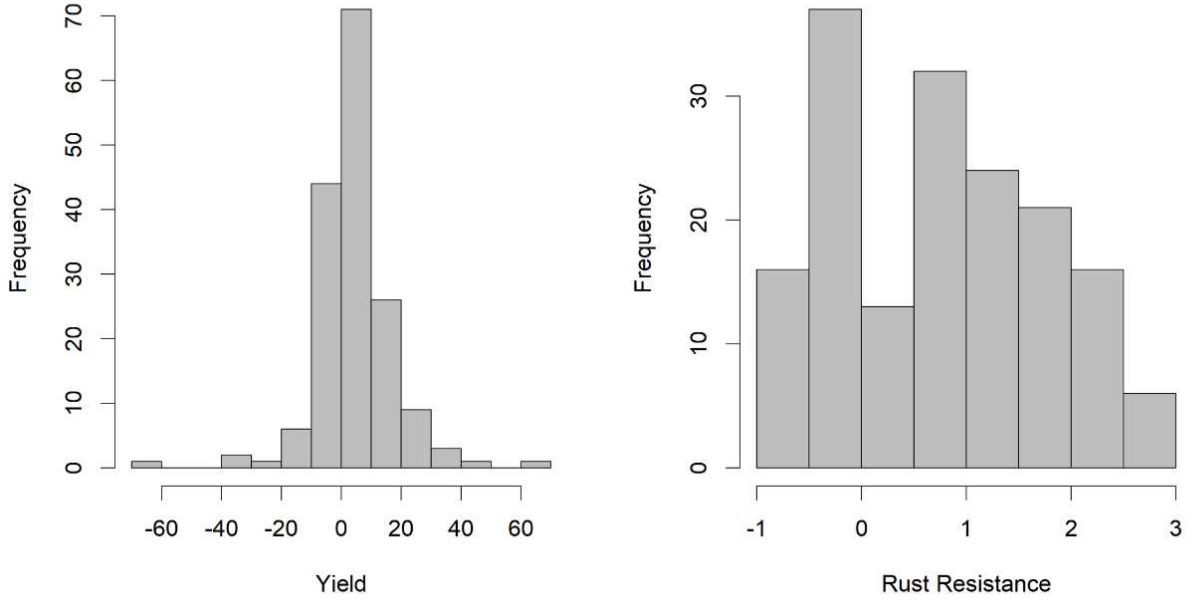
3. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*. 2020;52: 1–15. doi:10.1186/s12711-020-00531-z
4. González-Camacho JM, Crossa J, Pérez-Rodríguez P, Ornella L, Gianola D. Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics*. 2016;17: 1–16. doi:10.1186/s12864-016-2553-1
5. Khaki S, Wang L. Crop Yield Prediction Using Deep Neural Networks. 2019; 139–147. doi: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. doi:10.1016/j.tplants.2017.08.011
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. *Frontiers in Plant Science*. 2020;11: 1808. doi:10.3389/FPLS.2020.593897/BIBTEX
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. *Frontiers in Genetics*. 2019;10: 1091. doi:10.3389/FGENE.2019.01091/BIBTEX
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 Genetics*. 2011;12: 1–14. doi:10.1186/1471-2156-12-87/FIGURES/5
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 Genetics*. 2014;15: 1–10. doi:10.1186/s12863-014-0149-9
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, Genetics*. 2014;4: 1027–1046. doi:10.1534/g3.114.010298
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. *Molecular Breeding*. 2012;29: 297–311. doi: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. *Theoretical and Applied Genetics*. 2012;125: 1181–1194. doi:10.1007/s00122-012-1905-8
14. Denis M, Bouvet JM. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. *Tree Genetics and Genomes*. 2013;9: 37–51. doi: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. doi:10.1371/journal.pone.0143548

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. doi:10.1038/hdy.2016.23
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. *Frontiers in Plant Science*. 2019;9: 1934. doi:10.3389/fpls.2018.01934
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. *Molecular Breeding*. 2017;37: 6. doi: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. doi:10.1007/s10681-018-2126-2
20. Diniz LEC, Sakiyama NS, Lashermes P, Caixeta ET, Oliveira ACB, Zambolim EM, et al. Analysis of AFLP markers associated to the Mex-1 resistance locus in Icatu progenies. *Crop Breeding and Applied Biotechnology*. 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. doi:10.1007/s10681-018-2126-2
22. Resende MDV de. Software Selegen-REML/BLUP: a useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*. 2016;16: 330–339. doi:10.1590/1984-70332016v16n4a49
23. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria; 2019.
24. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 2007;177: 2389–2397. doi:10.1534/genetics.107.081190
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 Scientiarum - Agronomy*. 2019;41. doi: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. *Nature Genetics*. 2019;51: 12–18. doi:10.1038/s41588-018-0295-5
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. doi: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. *Ecological Modelling*. 2004;178: 389–397. doi: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. *Genetics Research*. 2012;94: 73–83. doi:10.1017/S0016672312000274
32. 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, Genetics*. 2014;4: 1027–1046. doi:10.1534/g3.114.010298
33. Waldmann P. Approximate Bayesian neural networks in genomic prediction. *Genetics Selection Evolution*. 2018;50: 70. doi:10.1186/s12711-018-0439-1
34. 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. doi:10.1186/S12864-020-07319-X/TABLES/5
35. 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 Scientiarum Agronomy*. 2020;43: e46307. doi:10.4025/actasciagron.v43i1.46307
36. Long N, Gianola D, Rosa GJM, Weigel KA. Marker-assisted prediction of non-additive genetic values. *Genetica*. 2011;139: 843–854. doi:10.1007/s10709-011-9588-7
37. 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. *Functional Plant Breeding Journal*. 2020;1.
38. 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 Genetics and Genomes*. 2020;16: 1–11. doi:10.1007/s11295-020-01433-3
39. Ferrão LF V., Marinho CD, Munoz PR, Resende Jr MFR. Improvement of predictive ability in maize hybrids by including dominance effects and marker × environment models. *Crop Science*. 2020;60: 666–677. doi:10.1002/csc2.20096
40. 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. doi:10.1534/genetics.120.303025
41. 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. doi:10.1038/hdy.2017.37
42. Tan B, Grattapaglia D, Wu HX, Ingvarsson PK. Genomic relationships reveal significant dominance effects for growth in hybrid *Eucalyptus*. *Plant Science*. 2018;267: 84–93. doi:10.1016/j.plantsci.2017.11.011
43. Shang L, Liang Q, Wang Y, Zhao Y, Wang K, Hua J. Epistasis together with partial dominance, over-dominance and QTL by environment interactions contribute to yield heterosis in upland cotton. *Theoretical and Applied Genetics*. 2016;129: 1429–1446. doi:10.1007/s00122-016-2714-2

44. 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. *Scientific Reports*. 2019;9: 3984. doi:10.1038/s41598-019-40611-9
45. Lin T, Zhou C, Chen G, Yu J, Wu W, Ge Y, et al. Heterosis-associated genes confer high yield in super hybrid rice. *Theoretical and Applied Genetics*. 2020;133: 3287–3297. doi:10.1007/s00122-020-03669-y
46. 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. doi:10.1186/s12284-018-0229-y
47. 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*). *Journal of Trace Elements in Medicine and Biology*. 2018;48: 190–195. doi:10.1016/j.jtemb.2018.04.009
48. de Almeida Filho JE, Guimarães JFR, e Silva FF, de Resende MD V, 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. doi:10.1038/hdy.2016.23
49. Geneti D. Progress of Coffee (*Coffea arabica* L) Hybridization Development Study in Ethiopia: A Review. 2019;92. doi:10.7176/FSQM/92-03
50. Geneti D. Review on Heterosis and Combining Ability Study for Yield and Morphological Characters of Coffee (*Coffea arabica* L) in Ethiopia. 2019;9. doi:10.7176/JEES/9-12-03
51. Cruz CD. *Princípios de Genética Quantitativa*. 1st ed. UFV, editor. Viçosa; 2005.
52. 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. *Scientia Agricola*. 2020;78: 1–8. doi:10.1590/1678-992x-2020-0021
53. 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. *Genetics and Molecular Research*. 2015;14: 12217–12227. doi:10.4238/2015.October.9.10
54. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2ed ed. New York: Springer; 2009.
55. 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. *Scientia Agricola*. 2020;78. doi:10.1590/1678-992x-2020-0021

### 2.7. Supporting information



**S1 Fig. Histogram.** Histogram of yield and rust resistance.

### 3. ARTIGO 2: The trade-off between density marker panels size and predictive ability of genomic prediction for agronomic traits in *Coffea canephora*

PlosOne: Under Review

#### 3.1. Abstract

Genomic prediction in *Coffea* breeding has shown good potential in predictive ability (PA), genetic gains, and reduction of the breeding cycle time. It is known that the cost of genotyping was prohibitive for many species, and their value is associated with the density markers panel used. Knowing which density marker panel to use is an information that may reduce the genotyping cost and improve the PA. This study, aimed to evaluate the trade-off between density marker panels size and the PA for eight agronomic traits in *Coffea canephora* using machine learning (bagging and random forest) algorithms and comparing them with BLASSO (Bayesian Least Absolute Shrinkage and Selection Operator) method. This data consisted of 165 genotypes of *Coffea canephora* genotyped for 14,387 SNP. The phenotypic data is composed of vegetative vigor (Vig), rust (Rus) and cercosporiose incidence (Cer), fruit maturation time (Mat), fruit size (FS), plant height (PH), diameter of the canopy projection (DC) and yield (Y). To evaluate the effect of the dimensionality reduction in the PA, twelve different density marker panels were used. In the analysis is observed that as the number of markers increase from 25 to 500/1000 markers, the PA increases, however, above 500/1000 markers as the number of markers increases the PA decrease. In general, the PA have lower values when used the full SNP panel density. Comparing the bigger and the lower PA for each trait, some had an improvement around of 100% (PH: 0.35-0.77; Cer: 0.40-0.84; DC: 0.39-0.82; Rus: 0.39-0.83, Vig: 0.40-0.77), the other presented an improvement more than 340% (Mat: 0.12-0.60; Y: 0.14-0.61; FS: 0.07-0.60). The current study results indicate that the reduction of the number of markers until 500/1000 markers can improve the selection of individuals at a lower cost.

### 3.2. Introduction

Genome-wide selection (GWS) was developed by Meuwissen et al. [1] and is currently used in animal and plant breeding programs [2–7]. One of its most important advantages is shorter generation intervals because selection candidates can be tested earlier in life [8].

However, one of the limitations is the cost to genotype many individuals with high density SNP (HD-SNPs) array platforms, which can be, in practice, prohibitively expensive for routine application for most breeding programs. To overcome this problem, to exploit a low-density SNP (LD-SNPs) panel may be a solution, reducing the genotyping costs.

Some approaches reducing a HD-SNPs panel has been used, [9] selected the subsets of SNPs randomly in two ways: i) selected random SNPs within each chromosome, keeping the proportionality of the number of the SNPs according to the length of each linkage group; ii) selected random SNPs across the whole genome. Habier et al. [10] and Ogawa et al. [11] studied subsets with markers evenly-spaced across the genome. These studies cited above using LD-SNPs did not obtain better results than HD-SNPs. However, Li et al. [12] used machine learning methods to select the SNPs for genomic prediction in beef cattle and obtained similar results to HD-SNPs.

Although a LD-SNPs panel may be an option to reduce costs, the optimal SNP density to use may vary depending on the species and trait of interest [9]. In this study, we used *Coffea canephora* due to its economic relevance worldwide, being responsible for approximately 40% of the world's coffee production [13].

This study aimed to evaluate the trade-off between density marker panel size and the predictive ability (PA) using machine learning (bagging and random forest) algorithms. The results were compared with those provided by the BLASSO (Bayesian Least Absolute Shrinkage and Selection Operator) method, for eight traits of *C. canephora*.

### 3.3. Materials and Methods

#### 3.3.1. Genotypes

In this work, 165 genotypes of *C. canephora* under breeding program was analyzed. The population consisted of 51 and 32 clones of the Conilon and Robusta varietal groups, respectively, and 82 intervarietal hybrids. These hybrids were originated from artificial

crosses between five genotypes of the Conilon group (male parents) and five genotypes of the Robusta group (female parents), evaluated in the interpopulational partial diallel.

The genetic material was obtained from the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper), and the Robusta material was obtained from the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE). This population composes the breeding program of the Empresa de Pesquisa Agropecuária de Minas Gerais (Epmig), in partnership with the Universidade Federal de Viçosa (UFV) and the Empresa Brasileira de Pesquisa Agropecuária—Café (Embrapa Café), located in Oratórios/MG and Viçosa/MG.

### 3.3.2. Phenotypic evaluations

Evaluations were performed as described in Alkimin et al. [14] for eight traits during three consecutive years (2014-2016). The traits evaluated were: vegetative vigor (Vig), field evaluation of rust incidence (Rus), cercosporiosis incidence (Cer), fruit maturation time (Mat), fruit size (FS), plant height (PH), diameter of the canopy projection (DC), and yield in liters per plant (Y).

### 3.3.3. Phenotypic data

The used model was:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{c} + \mathbf{Q}\mathbf{s} + \mathbf{S}\mathbf{b} + \mathbf{e},$$

where  $\mathbf{y}$  is the data vector;  $\boldsymbol{\mu}$  is the vector of year-mean effects (assumed as fixed) added to the overall mean;  $\mathbf{c}$  is the vector of specific combining ability effects between the Conilon and Robusta parents (assumed as random and distributed as  $N(0, I\sigma_c^2)$ );  $\mathbf{a}$  is the vector of additive genetic effects of individuals (assumed as random and distributed as  $N(0, A\sigma_a^2)$ );  $\mathbf{s}$  is the vector of permanent effects of individuals (assumed as random and distributed as  $N(0, I\sigma_s^2)$ );  $\mathbf{b}$  is the vector of permanent environment effects of blocks (assumed as random and distributed as  $N(0, I\sigma_b^2)$ ); and  $\mathbf{e}$  is the residual vector (assumed as random and distributed as  $N(0, I\sigma_e^2)$ ). All the effects were assumed as uncorrelated. Uppercase represents the incidence matrices for these effects. The phenotypes were corrected for environmental effects of years and blocks using the Selegen REML/BLUP software [15], given by  $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}} - \mathbf{S}\hat{\mathbf{b}}$  and are called deregressed phenotypes, which enter in the genomic analyses [3,16].



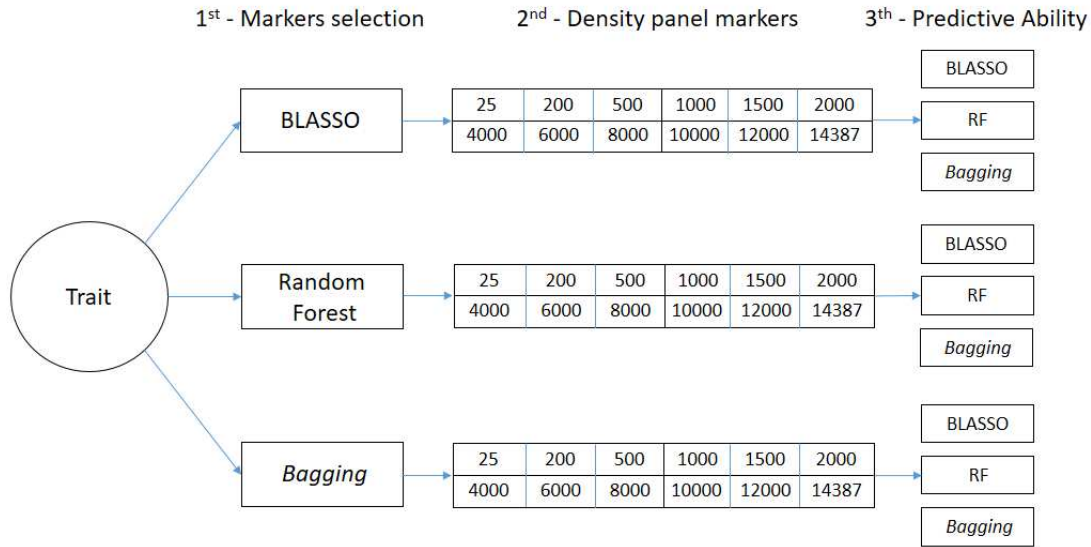
### **3.3.4. SNPs Markers**

The DNA extraction, identification and quality analysis of SNP markers were performed by Alkimim et al. [14]. The original data was reduced from 18111 SNP markers to 14387 markers considering MAF (minor allele frequency) of 0.05 and call rate of 0.90. The additive codification was used, for marker corresponding to homozygote with lower allele frequency we attribute 0, for heterozygous marker the value is 1, and for homozygote marker with bigger allele frequency we attribute 2.

### **3.3.5. Marker selection and Genomic Prediction**

BLASSO, random forest (RF), and bagging were applied to obtain the markers importance for the posterior use in prediction using lower density panel markers. The most important markers are those that have a higher influence on the studied trait, being used all individuals in the construction of the models to select such markers with greater precision. In BLASSO, the markers with the highest regression coefficients in absolute values were defined as the most important markers. In bagging and RF, we assumed as the important markers those that in mean influenced more in the reduction of the sum of the square of residuals. We choose 12 density panels size (25, 200, 500, 1000, 1500, 2000, 4000, 6000, 8000, 10000, 12000, 14387 markers) always choosing the most important markers according to each methodology.

To evaluate the dimensionality reduction in the predictive ability (PA) ( $\rho_{Y,\hat{Y}}$ ) of the genomic selection were used 108 different scenarios strategies for each trait. These scenarios



are the combination of 3 steps: 1<sup>st</sup> - to choose the methodology to select the markers (Blasso, bagging, or RF); 2<sup>nd</sup> - to choose the density panel size (25, 200, 500, 1000, 1500, 2000, 4000, 6000, 8000, 10000, 12000, 14387); 3<sup>th</sup> - to predict the PA (Blasso, bagging or RF), totaling 108 scenarios. (Fig 1).

**Fig 1. Steps to construct the 108 scenarios.** The three steps used to obtain all the 108 scenarios.

### 3.3.6. Bayesian Generalized Linear Regression

The genomic prediction method used was the Bayesian Least Absolute Shrinkage and Selection Operator [17] was also used for the prediction of GEBV (Genomic Estimated Breeding Values). The model is given by:

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

where  $\mathbf{y}$  is the vector of phenotypes with dimension  $N \times 1$  where  $N$  is the number of individuals,  $\mathbf{1}$  is a vector of the same dimension as  $\mathbf{y}$  with all elements equal to 1,  $\mu$  the intercept,  $\boldsymbol{\beta}$  is the vector of additive genetic effects of markers with dimension  $n \times 1$  with  $n$  being the number of markers,  $\mathbf{X}$  is the incidence matrix of the additive effects of the markers

and  $e$  ( $N \times 1$ ) is the error vector of the model, with  $e \sim N(0, I\sigma_e^2)$  being  $\sigma_e^2$  the error variance and  $I$  is the identity matrix.

The data distribution is denoted by  $y|\mu, \beta, \sigma_e^2 \sim N(1\mu + X\beta, I\sigma_e^2)$  and the prior distributions are:

$$\mu \sim N(0, 10^8)$$

$$\sigma_e^2 \sim v_e S_e^2 \chi^{-2}$$

$$\beta | \tau_{\beta_j}^2, \sigma_e^2 \sim N(0, I\tau_{\beta_j}^2 \sigma_e^2)$$

$$\tau_{\beta_j}^2 \sim \text{Exp}\left(\frac{\lambda^2}{2}\right)$$

$$\lambda^2 \sim \text{Gamma}(r, s).$$

where  $v_e S_e^2 \chi^{-2}$  represents the scaled inverse chi-squared distribution with the hyperparameters  $v_e$  and  $S_e^2$ , *Exp* and *Gamma* represents, respectively, exponential and gamma distribution.

For inference about the posterior distribution of the estimated effects of SNPs, 600,000 iterations were used for the Markov Chain Monte Carlo (MCMC) algorithms, of which 20,000 were discarded (burn-in) to guarantee the heating of the chain and selection of one in 150 iterations (thin). Convergence analysis was performed using the criterion proposed by Geweke [18].

### 3.3.7. Regression tree

To construct a regression tree, the objective is to obtain regions  $R_1, R_2, \dots, R_M$  that minimize the residual sum of squares-RSS given by [19]:

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where  $\hat{y}_{R_j}$  represents the mean response for the training observations in the  $j^{th}$  region. The RSS decreases according to the tree growth that occurs through recursive binary splitting. To increase the predictive performance of the model were used bootstrap aggregation (bagging) and random forest.

The bootstrap aggregation (bagging) consists of obtaining B samples with replacement (size equal to N) from the data set, obtaining B models ( $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ ) that will be used as individual classifiers. A new individual will be predicted as the mean of the B individual predictions. The random forest (RF) follows the same idea of the bagging, changing just the number of predictive variables ( $m < p$ ) used in each split. According to [19], RF results in the process of “decorrelating” the generated trees, improving them even more, the PA.

### 3.3.8. Training and Validation Sets

To estimate the PA the 5-fold cross-validation was used, dividing the data set into two parts: training set and validation set. We used the individuals from validation set to build the models and the individuals from validation set to estimate the PA, for each one of the 5 folds. For each fold, the training set was kept with the same individuals to model all methodologies, being these composed by 80% of each class (132 observations), taken at random, the 20% (33 observations) left were used in the validation set. In the literature, the percentages used in the training set vary between 60 and 90% as seen in [20] and [21].

### 3.3.9. Concordance analysis

We use the Cohen’s Kappa coefficient proposed by Cohen [22] to analyze the agreement between the methodologies in identifying markers (using the 12 groups defined above). The coefficient of Cohen’s Kappa is given by:

$$kappa = \frac{NAO - NAEC}{NOA - NAEC}$$

where NAO number of agreement observed, NAEC the number of agreements expected by chance and NOA number of observations analyzed [23].

Pearson's correlation [24] was used to check the pattern between the kappa’s coefficient obtained among the methodologies and the number of markers.

### 3.3.10. Computational Aspects

Data analysis was performed on a computer with 3.40GHz core i7 processor and 16GB of RAM and was used the software R 3.6.1 [25]. The BGLR function, belonging to the

BGLR package [26], was used to estimate the Bayesian generalized models. The randomForest function belonging to the randomForest package was used to construct the model of the bagging and the Random Forest [27].

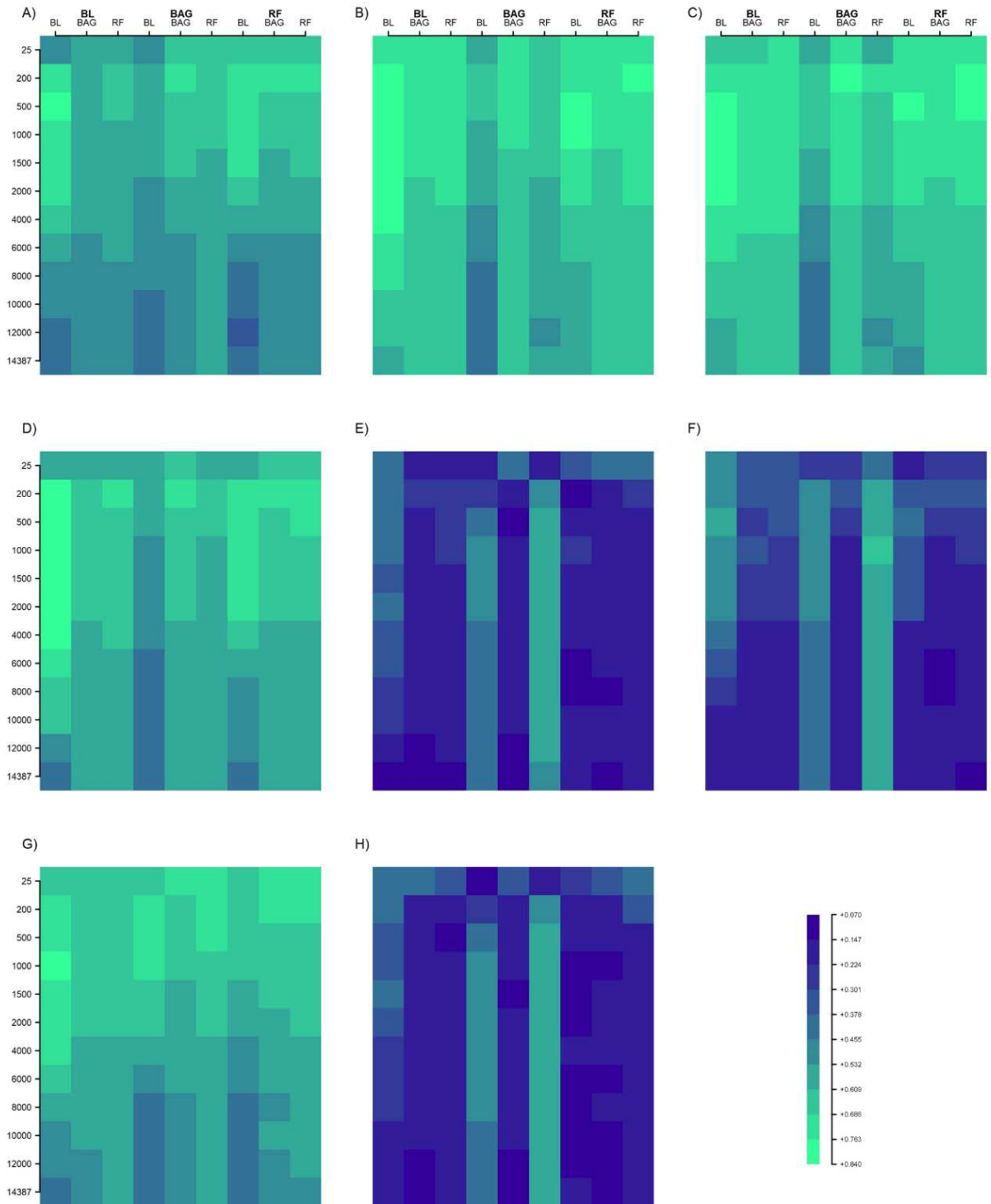
### 3.4. Results

#### 3.4.1. Trait Summary

The estimated genomic heritability and standard deviation ( $h_a^2 \pm sd$ ) values varied from  $0.15 \pm 0.04$  to  $0.53 \pm 0.03$ , for Y and DC, respectively. For disease resistance, two traits was analyzes, Rus and Cer, with the heritability of  $0.37 \pm 0.04$  and  $0.43 \pm 0.04$ , respectively. For FS, Mat, PH and Vig the genomic heritabilities were  $0.21 \pm 0.19$ ,  $0.21 \pm 0.19$ ,  $0.36 \pm 0.04$  and  $0.43 \pm 0.04$ , respectively. The mean of the traits were 6.02, 1.59, 2.08, 2.09, 2.16, 162.66, 144.54 and, 7.39 for Vig, Rus, Cer, Mat, FS, PH, DC and, Y, respectively [14].

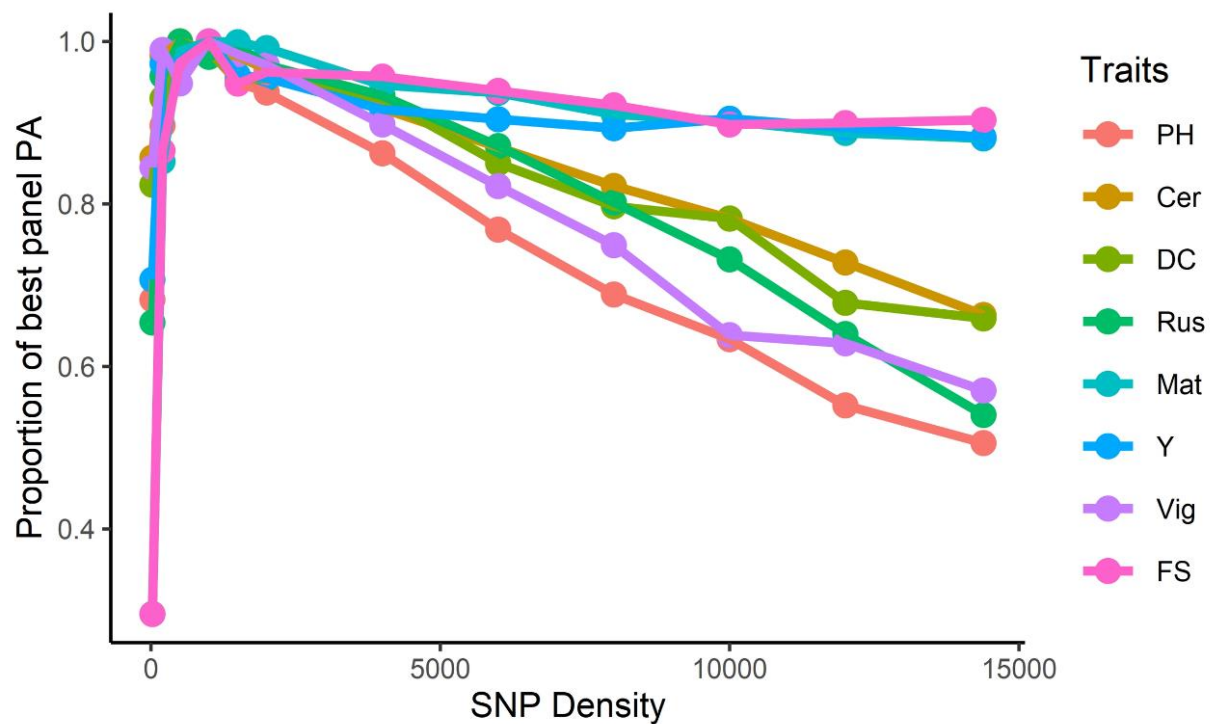
#### 3.4.2. Reduced SNP Panel Densities Improve the Prediction Ability of the traits

The predictive ability (PA) of genomic prediction was evaluated using five-fold cross-validation (training set 80%, validation set 20%) for 12 different panels' sizes. The markers' importance was obtained in each methodology and was used to evaluate the PA. For each trait, the PA was analyzed in all 108 scenarios. The common trend across the different traits and scenarios is observed by plotting the heat map (Figure 2) and shows that as the number of markers increase from 25 to 500/1000 markers, the PA increases, however, above 500/1000 markers as the number of markers increases the PA decrease. In general, the PA have lower values when used the full SNP panel density.



**Fig 2. Heatmap.** Heatmap of predictive ability considering all the 108 scenarios, created by the combination among three methodologies to select markers (in bold) with the same three methodologies to predict (no bold) the breeding value (in the top of X-axis) and 12 panel markers size (Y-axis). BL, Bayesian LASSO; BAG, Bagging; RF, Random Forest; A, Plant height; B, Cercosporiose incidence; C, Diameter of the canopy projection; D, Rust incidence; E, Fruit maturation time; F, Yield in liters per plant; G, Vegetative vigor; H, Fruit size.

In Figure 3, we selected the combination of methodologies that was used to select and to predict the PA that obtained the best results in general (BLASSO to select and to predict), using this combination the PA was calculated for each density marker panels. The density marker panel with the higher PA was choose for each trait, these PA values were considered as the based value and then was calculated the proportion of other PA values obtained in other panel markers densities. When the value plotted is equal to 1, it means that in this density panel size, the PA was higher. The average of these proportions increases from 25 markers (0.645) to 1000 markers (0.993), then it starts to decrease until full SNP panel density (0.701).



**Fig 3. Proportion of predictive ability (PA) achieved with low-density panels.** The proportion of PA achieved by each SNP density was calculated by dividing the PA at that density by the best PA of each trait, resulting the proportion being equal 1 in the best panel density. PH, Plant heigh; Cer, Cercosporiose incidence; DC, Diameter of the canopy projection; Rus, Rust incidence; Mat, Fruit maturation time; Y, Yield in liters per plant; Vig, Vegetative vigor; FS, Fruit size.

### 3.4.3. Comparison among methodologies

The best scenario varied according to the traits. For the traits PH, Cer, DC, Rus the best scenario was using BLASSO to select the markers and to obtain the PA with 500 markers. For the trait Vig, the best scenario was using BLASSO to select and to obtain the PA, however using 1000 markers. For the traits Mat, Y, and FS the best scenario was using the bagging to select the markers, and using RF to obtain the PA, with 1000 markers for all these traits. The worse scenario also varied according to the traits. For the Cer, DC, Rus and Vig the worst scenario was obtained when using bagging to select the markers, and using BLASSO to obtain the PA and used the full SNP panel size density (14387 markers). In the PH, Mat and FS traits, the worst scenario was using RF to select the markers, BLASSO to obtain the PA and 12000, 6000, 10000 markers, respectively. For Y, the worst scenario was selecting the markers through RF methodology, predicting using bagging and using 8000 markers.

A great breadth of PA was observed when compared the worse and best scenarios within the traits, some had an improvement around of 100% (PH: 0.35-0.77; Cer: 0.40-0.84; DC: 0.39-0.82; Rus: 0.39-0.83, Vig: 0.40-0.77), the other presented an improvement more than 340% (Mat: 0.12-0.60; Y: 0.14-0.61; FS: 0.07-0.60) (Sup Table 1).

When comparing the techniques among them two by two is observed that bagging and Random Forest are the methodologies that have a better agreement, obtaining in most cases an almost perfect agreement ( $Kappa = 1$ ). When comparing BLASSO with RF and with bagging, the agreement is considered poor and slight from 25 to 4000 markers. When considering all three methodologies together, the agreement among them is considered poor and slight from 25 to 6000 markers.

The correlation is considered “strong” when its absolute value is between 0.60 and 0.79 and “very strong” when its absolute value is between 0.80 and 1.00 [29]. The correlation obtained by the kappa’s coefficient coming from the agreement of bagging and RF and the number of markers, it is considered “strong” (0.71 to 0.78). Even the Kappa’s coefficient from bagging and RF being the better, the correlations with the number of markers were the worse due to the concave behavior of the kappa’s coefficient. All the other combinations presented a “very strong” correlation with the number of markers (0.93 to 1.00) (Sup Table 2).



For most of the traits the BLASSO with 500 markers was the best methodology to obtain the PA. We simulated the construction of a reduced SNP chip with the 500 most important markers for each trait, summing 3081 markers in total, once some markers appear among the most important markers for more than one trait. We compared the PA obtained when used all 3081 markers cited above with those obtained when using only the 500 most important markers for each trait. For all traits, we can see (Table 1) that using only the 500 most important markers for each trait the PA obtain higher values than using all the 3081 markers, the PA value reduced from 37% (Cer and DC) to 89% (Y).

**Table 1. Predictive ability for eight traits of coffee canephora.**

Traits	PH	Cer	DC	Rus	Mat	Y	FS	Vig
PA_500	0.77	0.84	0.82	0.83	0.41	0.55	0.37	0.73
PA_All	0.33	0.53	0.52	0.42	0.10	0.06	0.06	0.36

PH, Plant height; Cer, Cercosporiose incidence; DC, Diameter of the canopy projection; Rus, Rust incidence; Mat, Fruit maturation time; Y, Yield in liters per plant; FS, Fruit size; Vig, Vegetative vigor; PA\_500, predictive ability using 500 markers, selected and predicted by BLASSO; PA\_All, predictive ability using all markers, predicted by BLASSO.

### 3.5. Discussion

Genomic selection in *Coffea* breeding has showing good potential according to the PA, genetic gains and the reduction of the selection cycle time [30,31]. It is known that the cost of genotyping was prohibitive for many species [32–34], and their value is associated with the density markers panel used. To know what density marker panel to use is an information that may reduce the genotyping cost and still improve the PA. In this study, we used different scenarios and traits aiming to reduce the number of SNP markers and to evaluate the impacts caused by this reduction in PA. The results were consistent across the different traits, indicating that using between 500 and 1000 selected SNPs markers for each trait would be sufficient to obtain the best PA results, suggesting the use of reduced SNP panel density with selected markers for *C. canephora* studies.

Even considering different architecture from eight traits, the results were uniform. The heritability and the traits being categorical or continuous, does not show to be important for the performance of LD-SNPs panels, once the PA trends were consistent across the traits. What can explain this is the fact of having many SNPs explaining small effects, so the accuracy with which these effects are estimated are low [35].

The use of a high-density chip is not always a good strategy for selecting individuals, once we saw that the higher PA values were obtained using 500/1000 markers depending on the trait. However, to select the SNPs to use in the LD SNP chip, it is necessary to know which are the traits of interest. In this study, we reduced from 14387 to 3081 SNPS, covering all the eight traits studied, causing the cost reduction to genotype this species. According to Habier et al. [10], the use of smaller panels with SNPs can be used for genomic selection, but require separate SNPs for each trait, which was observed in this study. We have evaluated the two ways to use the reduced SNP chip for predicting the GEBV, realizing that using all the markers from complete reduced SNP chip for predicting one trait is not the best to do, once the best results were obtained by using only the 500 markers for each trait.

For this data set, BLASSO appears to be the best option. In general, this methodology demonstrated to be better for obtaining the PA, when compared with RF and bagging. The Kappa's concordance to select the markers, between BLASSO and the other methodologies had in general a slight agreement, except for disease resistance traits that had a fair agreement. The PA obtained through the reduced SNP panel density, showed to have a better performance than the full SNP panel density for *Coffea canephora*. When comparing the results obtained by GBLUP and full SNP panel density [14], with BLASSO and 500 markers, we had an increase from 41.38% to 87.80% on PA for the traits Vig, Rus, Cer, PH and DC. For Mat, FS and Y. The PA obtained by Alkimim et al. [14] were on average -0.03, 0.00 and -0.02 while using BLASSO were 40.84%, 54.79% and 37.21%, respectively.

It is expected to have similar results between RF and bagging due they be an ensemble model of decision tree [36]. Despite RF and bagging does not present better results than BLASSO, they still have better results when compared with those using GBLUP with full SNP panel density presented [14].

### 3.6. Conclusions

The results of the current study indicate that BLASSO must be used to select markers and to predict PA. The SNPs panels using between 500 and 1000 SNPs markers has better predictive ability than using full SNP panel density for the eight traits evaluated in this study of *Coffea canephora*.

## References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157: 1819–1829. doi:10.1093/GENETICS/157.4.1819
2. Sousa TV, Caixeta ET, Alkimim ER, Oliveira ACB, Pereira AA, Sakiyama NS, et al. Early selection enabled by the implementation of genomic selection in coffee arabica breeding. *Frontiers in Plant Science*. 2019;9: 1934. doi:10.3389/FPLS.2018.01934/BIBTEX
3. de Andrade LRB, Bandeira e Sousa M, Oliveira EJ, de Resende MDV, Azevedo CF. Cassava yield traits predicted by genomic selection methods. *PLOS ONE*. 2019;14: e0224920. doi:10.1371/JOURNAL.PONE.0224920
4. Oliveira GF, Nascimento ACC, Nascimento M, de Castro Sant'Anna I, Romero JV, Azevedo CF, et al. Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. *PLOS ONE*. 2021;16: e0243666. doi:10.1371/JOURNAL.PONE.0243666
5. 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. *Scientia Agricola*. 2020;78: 1–8. doi:10.1590/1678-992X-2020-0021
6. Silveira LS, Lima LP, Nascimento M, Nascimento ACC, Silva FF. Regression trees in genomic selection for carcass traits in pigs. *Genetics and Molecular Research*. 2020;19. doi:10.4238/GMR18498
7. Oliveira HR, Brito LF, Silva FF, Lourenco DAL, Jamrozik J, Schenkel FS. Genomic prediction of lactation curves for milk, fat, protein, and somatic cell score in Holstein cattle. *Journal of Dairy Science*. 2019;102: 452–463. doi:10.3168/JDS.2018-15159
8. Grossi DA, Brito LF, Jafarikia M, Schenkel FS, Feng Z. Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. *Animal*. 2018;12: 2235–2245. doi:10.1017/S175173111800085X
9. Kriaridou C, Tsairidou S, Houston RD, Robledo D. Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms. *Front Genet*. 2020;11. doi:10.3389/FGENE.2020.00124
10. Habier D, Fernando RL, Dekkers JCM. Genomic Selection Using Low-Density Marker Panels. *Genetics*. 2009;182: 343–353. doi:10.1534/GENETICS.108.100289
11. Ogawa S, Matsuda H, Taniguchi Y, Watanabe T, Nishimura S, Sugimoto Y, et al. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. *BMC Genetics*. 2014;15: 1–13. doi:10.1186/1471-2156-15-15/FIGURES/6
12. Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in Genetics*. 2018;9: 237. doi:10.3389/FGENE.2018.00237/BIBTEX

13. Coffee development report. In: International Coffee Organization [ICO] [Internet]. 2021 [cited 16 Jan 2022]. Available: [https://5aa6088a-da13-41c1-b8ad-b2244f737dfa.filesusr.com/ugd/38d76b\\_4fc7b54a15f14a548b2f4a208c2eae6d.pdf](https://5aa6088a-da13-41c1-b8ad-b2244f737dfa.filesusr.com/ugd/38d76b_4fc7b54a15f14a548b2f4a208c2eae6d.pdf)
14. 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 Genetics & Genomes* 2020 16:3. 2020;16: 1–11. doi:10.1007/S11295-020-01433-3
15. de Resende MDV. Software Selegen-REML/BLUP: a useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*. 2016;16: 330–339. doi:10.1590/1984-70332016V16N4A49
16. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*. 2009;41: 1–8. doi:10.1186/1297-9686-41-55/TABLES/1
17. Pérez P, de Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198: 483–495. doi:10.1534/GENETICS.114.164442/-/DC1
18. Geweke J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. 1991 [cited 16 Jan 2022]. doi:10.21034/SR.148
19. James G, Witten D, Hastie T, Tibshirani R. An introduction to Statistical Learning. *Curr Med Chem*. 2000;7: 995–1039. doi:10.1007/978-1-4614-7138-7
20. 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 Genetics*. 2011;12: 87. doi:10.1186/1471-2156-12-87
21. González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, et al. Genome-enabled prediction of genetic values using radial basis function neural networks. *TAG Theoretical and Applied Genetics Theoretische Und Angewandte Genetik*. 2012;125: 759. doi:10.1007/S00122-012-1868-9
22. Cohen J. A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES 1. *Educational and Psychological Measurement*. 1960;20: 37–46.
23. Resende MDV, Silva FF, Azevedo CF. *Estatística Matemática, Biométrica e Computacional*. 2014.
24. Pearson K. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*. 1895;58: 240–242. doi:10.1098/RSPL.1895.0041
25. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2019. Available: <https://www.r-project.org/>
26. Pérez P, de Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198: 483–495. doi:10.1534/GENETICS.114.164442/-/DC1
27. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2007;2: 18–22.

28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33: 159. doi:10.2307/2529310
29. Evans JD. *Straightforward statistics for the behavioral sciences*. 1996; 600.
30. Fanelli Carvalho Giovanni Galli Luís Felipe Ventrorm Ferrão Juliana Vieira Almeida Nonato Lilian Padilha Mirian Perez Maluf Márcio Fernando Ribeiro de Resende Jr Oliveira Guerreiro Filho Roberto Fritsche-Neto H, Fanelli Carvalho Á Vieira Almeida Nonato Á O Guerreiro Filho HJ, Galli Á Fritsche-Neto GR, Ventrorm Ferrão Á M F Ribeiro de Resende Jr LF, Padilha Á Perez Maluf LM. The effect of bienniality on genomic prediction of yield in arabica coffee. *Euphytica*. 216. doi:10.1007/s10681-020-02641-7
31. Ferrão LFV, Ferrão RG, Ferrão MAG, Fonseca A, Carbonetto P, Stephens M, et al. Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity* 2018 122:3. 2018;122: 261–275. doi:10.1038/s41437-018-0105-y
32. Happ MM, Wang H, Graef GL, Hyten DL. Generating High Density, Low Cost Genotype Data in Soybean [*Glycine max* (L.) Merr.]. *G3 Genes|Genomes|Genetics*. 2019;9: 2153–2160. doi:10.1534/G3.119.400093
33. Senthilvel S, Ghosh A, Shaik M, Shaw RK, Bagali PG. Development and validation of an SNP genotyping array and construction of a high-density linkage map in castor. *Scientific Reports* 2019 9:1. 2019;9: 1–10. doi:10.1038/s41598-019-39967-9
34. Tsairidou S, Hamilton A, Robledo D, Bron JE, Houston RD. Optimizing Low-Cost Genotyping and Imputation Strategies for Genomic Selection in Atlantic Salmon. *G3 Genes|Genomes|Genetics*. 2020;10: 581–590. doi:10.1534/G3.119.400800
35. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* 2013 14:7. 2013;14: 507–515. doi:10.1038/nrg3457
36. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2009 [cited 16 Jan 2022]. doi:10.1007/978-0-387-84858-7

### 3.7. Supporting Information

**S1 Table. Predictive ability.** Predictive ability obtained in 108 scenarios, considering three methodologies to select marker and to predict the breeding value, and 12 density panels size.

Number of markers	PH									Cer								
	BL			BAG			RF			BL			BAG			RF		
	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF
25	0.53	0.55	0.56	0.52	0.64	0.63	0.65	0.63	0.65	0.72	0.71	0.73	0.59	0.73	0.65	0.71	0.73	0.75
200	0.69	0.59	0.61	0.58	0.69	0.63	0.73	0.69	0.69	0.82	0.73	0.76	0.62	0.76	0.64	0.76	0.76	0.78
500	0.77	0.60	0.62	0.57	0.66	0.61	0.72	0.65	0.67	0.84	0.71	0.73	0.62	0.72	0.63	0.78	0.71	0.75
1000	0.76	0.58	0.59	0.56	0.62	0.61	0.73	0.62	0.63	0.83	0.70	0.72	0.59	0.70	0.62	0.77	0.70	0.72
1500	0.73	0.57	0.58	0.54	0.61	0.60	0.72	0.59	0.61	0.82	0.69	0.71	0.58	0.68	0.61	0.74	0.68	0.70
2000	0.72	0.56	0.58	0.52	0.59	0.60	0.68	0.58	0.60	0.81	0.67	0.69	0.56	0.67	0.59	0.72	0.67	0.69
4000	0.66	0.54	0.54	0.49	0.54	0.57	0.59	0.54	0.56	0.77	0.66	0.68	0.52	0.65	0.59	0.67	0.65	0.67
6000	0.59	0.53	0.54	0.47	0.52	0.56	0.48	0.51	0.52	0.73	0.64	0.66	0.48	0.64	0.57	0.62	0.64	0.66
8000	0.53	0.52	0.53	0.48	0.51	0.55	0.41	0.50	0.51	0.69	0.64	0.65	0.42	0.63	0.56	0.59	0.62	0.65
10000	0.49	0.50	0.50	0.42	0.50	0.55	0.40	0.49	0.50	0.65	0.62	0.64	0.44	0.62	0.55	0.58	0.63	0.63
12000	0.43	0.48	0.50	0.42	0.49	0.54	0.35	0.49	0.50	0.61	0.62	0.64	0.43	0.61	0.53	0.57	0.62	0.63
14387	0.39	0.48	0.49	0.40	0.49	0.54	0.39	0.48	0.49	0.56	0.61	0.62	0.40	0.61	0.54	0.56	0.61	0.62
Number of markers	DC									Rus								
	BL			BAG			RF			BL			BAG			RF		
	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF
25	0.68	0.67	0.70	0.60	0.76	0.60	0.69	0.75	0.76	0.54	0.58	0.60	0.54	0.68	0.56	0.56	0.65	0.67
200	0.76	0.71	0.74	0.68	0.77	0.69	0.76	0.76	0.78	0.80	0.68	0.69	0.55	0.70	0.62	0.71	0.69	0.71
500	0.82	0.71	0.72	0.67	0.74	0.66	0.78	0.74	0.77	0.83	0.65	0.66	0.55	0.68	0.61	0.75	0.67	0.69

1000	0.81	0.70	0.71	0.61	0.71	0.63	0.76	0.71	0.74	0.82	0.64	0.65	0.53	0.64	0.59	0.74	0.63	0.66
1500	0.78	0.70	0.71	0.60	0.70	0.62	0.76	0.70	0.72	0.83	0.63	0.64	0.51	0.62	0.60	0.73	0.62	0.65
2000	0.78	0.70	0.71	0.58	0.69	0.61	0.74	0.68	0.71	0.81	0.62	0.64	0.52	0.61	0.59	0.72	0.61	0.63
4000	0.76	0.69	0.70	0.49	0.66	0.58	0.67	0.66	0.68	0.78	0.60	0.61	0.48	0.58	0.59	0.64	0.58	0.60
6000	0.70	0.65	0.66	0.47	0.66	0.56	0.62	0.64	0.67	0.72	0.58	0.59	0.44	0.57	0.56	0.57	0.57	0.59
8000	0.65	0.64	0.66	0.44	0.64	0.55	0.60	0.64	0.66	0.67	0.56	0.58	0.42	0.56	0.54	0.52	0.55	0.58
10000	0.64	0.64	0.65	0.43	0.64	0.55	0.56	0.64	0.65	0.61	0.56	0.57	0.43	0.55	0.55	0.47	0.55	0.56
12000	0.56	0.64	0.66	0.42	0.64	0.53	0.55	0.64	0.65	0.53	0.54	0.56	0.44	0.55	0.55	0.47	0.55	0.56
14387	0.54	0.63	0.65	0.39	0.63	0.54	0.53	0.63	0.65	0.45	0.54	0.55	0.39	0.54	0.54	0.44	0.54	0.55
Mat										Y								
Number of markers	BL			BAG			RF			BL			BAG			RF		
	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF
25	0.39	0.17	0.21	0.15	0.40	0.18	0.35	0.40	0.41	0.49	0.34	0.33	0.26	0.27	0.43	0.20	0.25	0.23
200	0.40	0.26	0.23	0.26	0.17	0.52	0.14	0.22	0.26	0.51	0.34	0.36	0.48	0.32	0.59	0.37	0.31	0.36
500	0.41	0.20	0.23	0.43	0.14	0.60	0.21	0.16	0.17	0.55	0.30	0.31	0.53	0.26	0.60	0.39	0.26	0.30
1000	0.41	0.21	0.23	0.48	0.15	0.60	0.23	0.16	0.16	0.49	0.32	0.29	0.48	0.20	0.61	0.35	0.21	0.24
1500	0.33	0.19	0.21	0.48	0.15	0.60	0.17	0.16	0.15	0.51	0.26	0.27	0.50	0.19	0.59	0.32	0.19	0.21
2000	0.38	0.17	0.18	0.51	0.15	0.60	0.21	0.17	0.16	0.48	0.25	0.24	0.48	0.18	0.58	0.31	0.18	0.20
4000	0.34	0.17	0.19	0.44	0.15	0.57	0.16	0.17	0.16	0.39	0.20	0.21	0.45	0.16	0.56	0.17	0.16	0.17
6000	0.37	0.16	0.16	0.45	0.15	0.57	0.12	0.15	0.17	0.34	0.18	0.19	0.45	0.15	0.55	0.18	0.14	0.16
8000	0.24	0.15	0.16	0.45	0.15	0.55	0.14	0.14	0.16	0.29	0.17	0.18	0.41	0.16	0.55	0.15	0.14	0.16
10000	0.23	0.15	0.15	0.42	0.15	0.54	0.15	0.16	0.15	0.20	0.16	0.15	0.39	0.15	0.55	0.19	0.15	0.16
12000	0.16	0.14	0.15	0.41	0.13	0.54	0.16	0.15	0.15	0.21	0.15	0.16	0.41	0.16	0.55	0.20	0.15	0.15
14387	0.12	0.13	0.14	0.40	0.14	0.53	0.16	0.14	0.15	0.20	0.15	0.15	0.41	0.16	0.54	0.19	0.15	0.14
Vig										FS								

Number of markers	BL			BAG			RF			BL			BAG			RF		
	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF	BL	BAG	RF
25	0.65	0.62	0.64	0.63	0.69	0.71	0.62	0.70	0.71	0.39	0.40	0.35	0.13	0.36	0.18	0.26	0.32	0.38
200	0.76	0.63	0.66	0.69	0.68	0.72	0.67	0.69	0.71	0.39	0.16	0.18	0.30	0.18	0.52	0.18	0.22	0.31
500	0.73	0.64	0.65	0.71	0.65	0.69	0.65	0.65	0.68	0.37	0.16	0.13	0.45	0.15	0.58	0.17	0.18	0.20
1000	0.77	0.61	0.63	0.69	0.62	0.66	0.64	0.63	0.65	0.37	0.18	0.16	0.49	0.15	0.60	0.13	0.14	0.16
1500	0.75	0.61	0.63	0.65	0.59	0.64	0.60	0.61	0.63	0.42	0.19	0.16	0.48	0.14	0.57	0.11	0.16	0.19
2000	0.74	0.61	0.62	0.64	0.60	0.62	0.58	0.60	0.61	0.37	0.16	0.17	0.50	0.15	0.57	0.11	0.17	0.18
4000	0.69	0.58	0.59	0.55	0.56	0.59	0.52	0.57	0.59	0.28	0.15	0.15	0.46	0.15	0.57	0.15	0.15	0.17
6000	0.63	0.57	0.59	0.51	0.54	0.57	0.52	0.56	0.57	0.25	0.16	0.18	0.46	0.15	0.56	0.13	0.13	0.17
8000	0.57	0.56	0.57	0.45	0.53	0.55	0.43	0.53	0.56	0.24	0.15	0.18	0.47	0.16	0.55	0.13	0.16	0.17
10000	0.49	0.54	0.55	0.45	0.53	0.55	0.45	0.54	0.55	0.16	0.16	0.15	0.43	0.15	0.54	0.07	0.13	0.16
12000	0.48	0.53	0.54	0.44	0.53	0.54	0.43	0.53	0.54	0.15	0.12	0.17	0.42	0.14	0.54	0.14	0.14	0.17
14387	0.44	0.52	0.54	0.40	0.53	0.54	0.41	0.52	0.53	0.16	0.14	0.17	0.39	0.13	0.54	0.16	0.13	0.16

BL, Bayesian LASSO; BAG, Bagging; RF, Random Forest; PH, Plant height; Cer, Cercosporiose incidence; DC, Diameter of the canopy projection; Rus, Rust incidence; Mat, Fruit maturation time; Y, Yield in liters per plant; Vig, Vegetative vigor; FS, Fruit size.

**S2 Table. Kappa concordance and correlation.** Kappa's coefficient among methodologies across many numbers of markers and the correlation between Kappa's coefficient and number of markers.

Traits	Methodologies	Number of markers												Correlation
		25	200	500	1000	1500	2000	4000	6000	8000	10000	12000	14387	
Vig	Bag_RF	0.76	0.72	0.62	0.54	0.51	0.51	0.55	0.61	0.70	0.78	0.87	1.00	0.74
	RF_BL	0.04	0.07	0.13	0.18	0.23	0.27	0.36	0.48	0.59	0.71	0.84	1.00	1.00
	BAG_BL	0.00	0.06	0.14	0.19	0.22	0.25	0.36	0.47	0.59	0.72	0.84	1.00	0.99
	RF_BAG_BL	0.00	0.04	0.10	0.12	0.14	0.16	0.22	0.31	0.43	0.57	0.74	1.00	0.98



Rus	Bag_RF	0.80	0.62	0.61	0.54	0.53	0.53	0.56	0.62	0.70	0.78	0.87	1.00	0.77
	RF_BL	0.36	0.27	0.33	0.36	0.37	0.38	0.44	0.53	0.62	0.72	0.84	1.00	0.99
	BAG_BL	0.36	0.30	0.32	0.33	0.35	0.36	0.43	0.52	0.62	0.72	0.84	1.00	0.99
	RF_BAG_BL	0.36	0.23	0.24	0.24	0.24	0.25	0.29	0.36	0.46	0.58	0.74	1.00	0.93
Cer	Bag_RF	0.84	0.66	0.57	0.53	0.50	0.50	0.52	0.61	0.69	0.78	0.87	1.00	0.71
	RF_BL	0.20	0.30	0.28	0.31	0.32	0.34	0.42	0.50	0.60	0.72	0.84	1.00	0.99
	BAG_BL	0.20	0.25	0.28	0.30	0.33	0.33	0.40	0.49	0.60	0.72	0.84	1.00	0.99
	RF_BAG_BL	0.16	0.22	0.20	0.21	0.21	0.21	0.26	0.33	0.44	0.58	0.74	1.00	0.96
Mat	Bag_RF	0.92	0.24	0.28	0.47	0.51	0.52	0.57	0.63	0.71	0.79	0.87	1.00	0.72
	RF_BL	0.04	0.07	0.11	0.12	0.14	0.17	0.32	0.45	0.59	0.71	0.84	1.00	1.00
	BAG_BL	0.04	0.05	0.07	0.10	0.14	0.18	0.32	0.46	0.59	0.71	0.84	1.00	1.00
	RF_BAG_BL	0.04	0.03	0.04	0.06	0.07	0.09	0.19	0.29	0.42	0.56	0.73	1.00	0.99
FS	Bag_RF	0.84	0.20	0.35	0.55	0.57	0.56	0.56	0.63	0.70	0.78	0.87	1.00	0.75
	RF_BL	0.04	0.02	0.08	0.12	0.15	0.18	0.33	0.45	0.59	0.71	0.84	1.00	1.00
	BAG_BL	0.08	0.06	0.08	0.11	0.14	0.17	0.32	0.45	0.59	0.71	0.84	1.00	1.00
	RF_BAG_BL	0.04	0.02	0.04	0.06	0.07	0.09	0.19	0.29	0.42	0.56	0.74	1.00	0.99
PH	Bag_RF	0.84	0.61	0.54	0.52	0.52	0.52	0.56	0.61	0.69	0.78	0.87	1.00	0.75
	RF_BL	0.04	0.10	0.15	0.22	0.25	0.27	0.37	0.48	0.58	0.70	0.84	1.00	0.99
	BAG_BL	0.04	0.11	0.14	0.20	0.24	0.27	0.36	0.48	0.59	0.71	0.84	1.00	1.00
	RF_BAG_BL	0.04	0.06	0.09	0.13	0.15	0.17	0.23	0.32	0.42	0.56	0.74	1.00	0.98
DC	Bag_RF	0.76	0.67	0.55	0.50	0.49	0.49	0.53	0.61	0.69	0.78	0.87	1.00	0.78
	RF_BL	0.04	0.10	0.14	0.18	0.22	0.25	0.36	0.46	0.59	0.71	0.84	1.00	1.00
	BAG_BL	0.04	0.11	0.13	0.17	0.21	0.25	0.37	0.47	0.59	0.71	0.84	1.00	1.00
	RF_BAG_BL	0.04	0.09	0.08	0.11	0.13	0.14	0.22	0.30	0.43	0.57	0.74	1.00	0.98
Y	Bag_RF	0.84	0.64	0.63	0.54	0.54	0.55	0.57	0.63	0.70	0.78	0.87	1.00	0.73

RF_BL	0.00	0.08	0.12	0.16	0.20	0.23	0.34	0.46	0.58	0.70	0.84	1.00	1.00
BAG_BL	0.00	0.10	0.12	0.15	0.19	0.24	0.35	0.47	0.59	0.71	0.84	1.00	1.00
RF_BAG_BL	0.00	0.06	0.09	0.10	0.12	0.14	0.22	0.31	0.42	0.56	0.73	1.00	0.98

---

BL, Bayesian LASSO; BAG, Bagging; RF, Random Forest; PH, Plant heigh; Cer, Cercosporiose incidence; DC, Diameter of the canopy projection; Rus, Rust incidence; Mat, Fruit maturation time; Y, Yield in liters per plant; Vig, Vegetative vigor; FS, Fruit size.

#### **4. GENERAL CONCLUSION**

There is not a methodology that will be better than all others, it depends of data's complexity, but computational intelligence and machine learning methodologies have shown great potential in genomic analysis, predicting genomic estimated breeding values, and estimating markers effects and heritability, besides of select markers for reducing the density of markers panel and improving the predictive ability.