Acta Scientiarum
http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v47i1.69698

Low-density marker panels for genomic prediction in *Coffea arabica* L.

Edilaine Silva Arcanjo[1], Moysés Nascimento[1], Camila Ferreira Azevedo[1], Eveline Teixeira Caixeta[2,3], Antônio Carlos Baião de Oliveira[3,4], Antonio Alves Pereira[4] and Ana Carolina Campana Nascimento[1*]

[1]Laboratório de Inteligência Computacional e Aprendizado Estatístico, Departamento de Estatística, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Instituto de Biotecnologia Aplicada à Agropecuária, Viçosa, Minas Gerais, Brasil. [3]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Café, Brasília, Distrito Federal, Brazil. [4]Empresa de Pesquisa Agropecuária de Minas Gerais, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E- mail: ana.campana@ufv.br

ABSTRACT. Developing new cultivars, particularly in perennial species like *Coffea arabica*, can be a time-consuming process. Employing molecular markers in genome-wide selection (GWS) for predicting genetic values offers an alternative to accelerate this process. However, implementing GWS typically involves genotyping many markers for both training and candidate individuals, which can increase the total genotyping cost for the breeding program. Therefore, this study aimed to assess the feasibility of using low-density marker panels to predict the genetic merit of *C. arabica* for a range of desirable agronomic traits. For this purpose, GWS analyses were performed using the G-BLUP method with panels of varying marker densities, selected based on marker effect magnitude. The results indicate that employing lower-density panels might be advantageous for this species' improvement. Models based on these panels yielded accurate predictions for various traits and demonstrated high agreement in terms of selected individuals compared to more complex models.

Keywords: genetic improvement; coffee; genomic selection; G-BLUP.

Received on September 20, 2023.
Accepted on January 24, 2024.

Introduction

Coffee remains a major commodity in the international market, accounting for over 170 million 60 kg bags in production during 2022 (ICO, 2023). For coffee-growing regions worldwide, cultivation plays a vital economic and social role (Partelli et al., 2014), driving research into genetic improvement programs and technological advancements. This has led to the development and recommendation of numerous cultivars with enhanced traits such as uniform maturation, resistance or tolerance to pests and diseases, high productivity, and beverage quality (Carvalho, 2008; Santana et al., 2021).

However, one of the biggest challenges in plant genetic breeding, especially for perennial species like coffee, is the lengthy and costly process involved in obtaining a new cultivar. And, the molecular markers can offer valuable assistance to breeding programs by overcoming this limitation. DNA markers linked to genes of interest are essential tools for selecting superior genotypes and play a crucial role in the selection process (Caixeta, Pestana, & Pestana, 2015).

Meuwissen, Hayes, and Goddard (2001) introduced Genomic Wide Selection (GWS), suggesting that genetic variation originates from all segments of the genome, with varying contributions. Each segment is in high Linkage Disequilibrium (LD) with at least one known genetic marker. GWS relies on molecular genetic markers of the Single Nucleotide Polymorphism (SNP) type, abundantly distributed throughout the genome. The genetic effects of these markers, in linkage disequilibrium with Quantitative Trait Loci (QTLs), can be used to identify candidate individuals for selection, thereby increasing the accuracy of genetic evaluation (Goddard & Hayes, 2007; Resende, Silva, Lopes, & Azevedo, 2012). GWS has proven to be a valuable tool for estimating the genetic values of plants and animals in a shorter period (Zhang, Yin, Wang, Yuan, & Liu, 2019).

While high-density SNP marker panels provide better genome coverage, requiring the genotyping of a large number of individuals is common for genomic prediction, leading to increased costs for breeding programs (Sousa et al., 2019a; Resende et al., 2012). One alternative to minimize these costs is to genotype selection candidates with a reduced-density panel of SNPs. This allows for genotyping of more individuals

using fewer markers. Studies in diverse fields, including aquaculture data (Kriaridou, Tsairidou, Houston, & Robledo, 2020), rice and maize (Sousa et al., 2019b), dairy cattle (Aliloo et al., 2018), soybeans (Ma et al., 2016), and pigs (Wellmann et al., 2013), have demonstrated the success of selecting markers with relevant effects, increasing both accuracy and precision from a low-density and more affordable SNP panel for implementing genomic selection in breeding programs.

In this context, our work aims to evaluate the feasibility of using low-density marker panels to predict the genetic value of economically important traits in *Coffea arabica*. Our goal is to develop a more parsimonious prediction model and ultimately reduce genotyping costs in coffee breeding programs.

Material and methods

Field trials

The experiment was conducted at the experimental station of the Department of Plant Pathology at the Federal University of Viçosa (UFV) in Viçosa, Minas Gerais State, Brazil (20º44'25" S; 42º50'52" W), starting on February 11, 2011. Phenotypic data was collected in the years 2014, 2015, and 2016. In this study, eight agronomically important traits were analyzed: branch length (BL), number of vegetative nodes (NVN), total number of fruits (NF), canopy diameter (CD), ripening fruit size (RFS), cercosporiosis incidence (Cer), leaf miner infestation (LM), and vegetative vigor (Vig).

For the genetic material, six genotypes from two contrasting groups based on coffee leaf rust resistance were selected for crossing: three from the Catuaí group, susceptible to *Hemileia vastatrix*, and three from the 'Híbrido de Timor' (HdT) group, resistant to *H. vastatrix*. These crosses resulted in 13 *Coffea arabica* progenies from the Epamig/UFV/Embrapa breeding program, with each progeny containing 15 genotypes. This yielded a total of 195 individuals belonging to generations of resistant backcrosses (BCr), susceptible backcrosses (BCs), and F_2 (Table 1).

Table 1. *Coffea arabica* progenies from the Epamig/UFV/Embrapa breeding program.

Progeny	Genotypes	Parental 1	Parental 2
BCr1	1-15	H 419-1 c-17	UFV 445-46
BCs2	16-30	H 419-1 c-17	UFV 2143-235
BCr3	31-45	H 514-8 c-387	UFV 440-10
BCs4	46-60	H 514-8 c-387	UFV 2154-344
BCr5	61-75	H 514-7 c-364	UFV 440-10
BCs6	76-90	H 514-7 c-364	UFV 2154-344
BCr7	91-105	H 419-10 c-214	UFV 445-46
BCs8	106-120	H 419-10 c-214	UFV 2143-235
BCs9	121-135	H 513-5 c-14	UFV 2148-57
F_2 10	136-150	H 514-8 c-387	-
F_2 11	151-165	H 514-7 c-364	-
F_2 12	166-180	H 419-10 c-214	-
F_2 13	181-195	H 513-5 c-14	-

In addition to phenotyping, the 195 individuals were genotyped using 40,000 polymorphic probes, yielding 21,211 SNPs. These markers underwent quality control to remove those failing to meet minimal criteria: Call Rate (CR) ⩾ 90% and Minor Allele Frequency (MAF) ⩾ 5%. For MAF, the critical level was established by the equation: $MAF = \frac{1}{\sqrt{2N}}$, where N is the number of individuals evaluated. Additionally, markers lacking variance within the population were eliminated to avoid false SNPs (Vidal et al., 2010). Following this analysis, 20,477 SNPs were retained, representing a 3.46% reduction from the initial set.

Phenotypic data analysis

Individual analysis of the eight agronomic traits was performed to estimate genetic parameters, and the following statistical model was adjusted:

$$y = \mathrm{Xu} + \mathrm{Zg} + \mathrm{Wp} + \mathrm{Vr} + \mathrm{Tb} + \mathrm{Ri} + \mathrm{e} \quad (1)$$

where: y is the vector of phenotypes; u is the general average of each evaluation year; g is the vector of progeny effects (random effect); p is the permanent variance between plants (random effect); r is the variance

between population types (random effect); b is the variance between plots (random effect); i is the variance of the interaction progenies × years (random effect); and e is the vector of errors (random effect). In the model, uppercase letters correspond to incidence matrices for the effects. Analyses were performed using the Selegen-REML/BLUP software (Resende, 2016). To account for variations due to year, plot, and progeny × year interactions, the phenotypic data was adjusted.

Genomic selection and cross-validation

Genomic predictions were performed using the G-BLUP (Genomic Best Linear Unbiased Prediction) method. This involved first fitting a mixed linear model to estimate additive individual genetic effects based on corrected phenotypic data (Resende et al., 2012), as follows:

$$y^* = Xb + Zg + e \quad (2)$$

where: y^* is the vector of adjusted phenotypic observations for the effects of years, plots, and interaction progenies × years; b is the vector of the mean fixed effects; g is the vector of additive individual genetic effect $[g \sim N(0, G\sigma_g^2)]$, with G being the additive kinship matrix and σ_g^2 the additive genetic variance; X and Z are the incidence matrices of the fixed and random effects, respectively; e is a normal random error vector, with $e \sim N(0, I\sigma_e^2)$ and σ_e^2 being the residual variance. The vector g was predicted using the following mixed model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix},$$

where: $G = \frac{WW'}{2\sum_i^n p_i(1-p_i)}$ is the additive genomic relatedness matrix, W is the incidence matrix of markers coded according to Vitezica, Varona, and Legarra (2013), p_i is the allele frequency of A of the i^{-th} marker, and n is the number of markers. Variance components were estimated via REML (Restricted maximum likelihood). Therefore, based on the prediction of genomic genetic values ($\hat{g}$) through the G-BLUP, the estimates of marker effects ($\hat{m}$) were obtained as follows: $\hat{m} = (W'W)^{-1}W'\hat{g}$.

The population consisted of 195 coffee trees, which were divided into k = 5 folds. Within these folds, 156 individuals were used for estimating and training the predictive models, while the remaining 39 individuals were reserved for validation purposes. This process was repeated five times, ensuring that each group served as a validation set once. After these iterations, the predictive ability ($r_{y\hat{y}}$) was estimated. Subsequently, Cohen's Kappa coefficient was employed to assess the agreement among individuals selected using the genomic estimated breeding values (GEBV) derived from the various adjusted models. This coefficient takes into consideration the probability of random agreement, rendering it a more precise index. The calculation of this coefficient can be expressed as follows:

$$\hat{k} = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where: $Pr(a) - Pr(e)$ is the proportion of observed agreements that occurred with a magnitude greater than what would be expected, and $1 - Pr(e)$ represents the proportion of observed disagreements. The coefficient $\hat{k}$ ranges from 0 to 1, where higher values indicate a greater level of agreement among the groups. According to Landis and Koch (1977), $\hat{k}$ values greater than 0.4 are considered satisfactory.

Different densities of SNP markers were used to assess the feasibility of employing low-density panels for predicting the GEBV of the eight analyzed traits. This feasibility was determined by evaluating how the reduction in the number of markers affected the predictive ability of the model $(r_{y\hat{y}})$, using the G-BLUP method. Additionally, concordance between individuals selected using GEBV derived from all 20,477 markers and those selected from models adjusted for various densities was examined.

Initially, models were adjusted for each trait of interest with different marker densities, namely: 2,047; 4,095; 6,143; 8,191; 10,238; 12,286; 14,334; 16,382; 18,429; and 20,477 SNPs, selected based on the magnitude of their effects (largest effects in absolute value). These densities correspond to 10, 20, and so on up to 100% of the 20,477 SNPs.

From the initial analyses, the model with the best predictive ability estimates and the highest concordance with the full model (20,477 SNPs) was selected for each trait. Based on these best models, further evaluations were conducted using SNPs corresponding to the intersection and union of those previously selected within the three groups of traits evaluated in this study.

The formation of these groups aimed to identify SNPs associated with different traits collectively, as the breeding process considers all traits simultaneously. Group 1 included traits associated with Arabica coffee morphology (CD, BL, NVN, and Vig), Group 2 included traits related to *C. arabica* fruits (RFS and NF), and Group 3 comprised traits associated with diseases and pests affecting *C. arabica* (Cer and LM).

The set intersection of SNPs represents those markers that are common to the traits within their respective group. Conversely, the SNP union sets result from markers that are relevant to at least one of the traits in each group, including markers selected for all traits. Subsequently, new densities were obtained, and the analyses were repeated to estimate predictive ability and agreement.

Results and discussion

This study aimed to assess the potential of low-density marker panels for predicting the GEBVs associated with economically significant traits in *Coffea arabica*. To this end, the G-BLUP genomic selection method was employed. This method involved adjusting models for each trait of interest at various marker densities, ranging from 2,047 to 20,477 SNPs. These models were selected based on the strength of their effects, prioritizing those markers with the most effects in absolute terms.

After adjusting the model for different marker densities, the goal was to identify the subset of markers that would provide the most accurate GEBV values for the evaluated traits. To achieve this, we estimated predictive capacities ($r_{y\hat{y}}$), along with their standard errors (Table 2), and assessed the concordance between individuals selected based on GEBV derived from the complete set of 20,477 SNPs and those chosen from models fitted with different marker densities (Table 3).

Predictive capabilities ranged from 0.03 to 0.39 across different marker densities for the eight evaluated traits. Overall, these estimates remained relatively stable, except for the LM trait, which exhibited an increase in predictive ability from 0.03 for the complete model to 0.12 for the model with the lowest density evaluated (Table 2). However, when considering the standard error, they remained constant for each trait. This outcome suggests that genomic prediction can be effectively conducted with lower marker densities, without sacrificing predictive ability.

Likewise, Sousa et al. (2019a) used the G-BLUP method with the RKHS (Reproducing Kernel Hilbert Spaces) procedure, incorporating a Bayesian algorithm and using 20,477 SNPs for a GWS research on *C. arabica* traits. They reported a predictive ability of 0.40 for CD, 0.32 for BL, 0.21 for Vig, 0.23 for RFS, and 0.31 for Cer. These values closely align with ours, where a lower number of SNPs were analyzed.

Table 2. Mean predictive abilities values ($r_{y\hat{y}}$) and their standard errors (between parentheses) using different SNP marker densities.

Traits	Number of SNPs									
	2047	4095	6143	8191	10238	12286	14334	16382	18429	20477
CD	0.36	0.36	0.39	0.39	0.39	0.38	0.37	0.37	0.36	0.36
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
BL	0.36	0.35	0.35	0.35	0.35	0.35	0.34	0.34	0.34	0.33
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
NVN	0.26	0.26	0.29	0.30	0.30	0.30	0.30	0.29	0.28	0.27
	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)	(0.06)	(0.05)
Vig	0.30	0.32	0.33	0.33	0.32	0.32	0.30	0.29	0.28	0.27
	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)
RFS	0.28	0.27	0.25	0.24	0.24	0.24	0.24	0.23	0.23	0.22
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
NF	0.29	0.26	0.25	0.23	0.21	0.20	0.18	0.16	0.14	0.11
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
Cer	0.33	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.31	0.31
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
LM	0.12	0.09	0.09	0.09	0.09	0.08	0.08	0.06	0.05	0.03
	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)

CD = Canopy diameter; BL = Branch length; NVN = Number of vegetative nodes; Vig = Vegetative Vigor; RFS = Ripening fruit size; NF = Total number of fruits; Cer = Cercosporiosis incidence; LM = Leaf miner infestation.

The agreement between individuals selected based on GEBV using all 20,477 markers and those chosen from models with lower marker densities (Table 3) ranged from 0.62 to 1, falling within the substantial to almost perfect range according to Landis and Koch (1977). While decreasing marker density led to a slight reduction in agreement, the selected individuals remained similar in their predictions.

Table 3. Cohen's Kappa coefficient of agreement between individuals selected based on genomic estimated breeding values (GEBV) using all (20,477) vs. reduced-density markers (selected by fitted models).

	Number of SNPs									
Traits	2047	4095	6143	8191	10238	12286	14334	16382	18429	20477
CD	0.69	0.72	0.84	0.81	0.87	0.94	0.94	0.97	0.97	1
BL	0.87	0.94	0.94	0.97	1	1	1	1	1	1
NVN	0.65	0.81	0.91	0.94	0.97	0.97	0.97	1	1	1
Vig	0.65	0.84	0.91	0.94	0.94	0.94	0.97	1	1	1
RFS	0.62	0.84	0.87	0.91	0.91	0.94	0.94	0.97	1	1
NF	0.78	0.84	0.94	0.94	0.97	0.97	0.97	0.97	0.97	1
Cer	0.69	0.75	0.81	0.87	0.97	0.97	0.97	0.97	0.97	1
LM	0.65	0.75	0.84	0.97	0.94	0.91	0.91	0.94	0.97	1

CD = Canopy diameter; BL = Branch length; NVN = Number of vegetative nodes; Vig = Vegetative Vigor; RFS = Ripening fruit size; NF = Total number of fruits; Cer = Cercosporiosis incidence; LM = Leaf miner infestation.

Our results indicate that genomic prediction in *C. arabica* can effectively be carried out with fewer than 20,477 SNPs. The models' predictive abilities remained stable, and agreement between individuals selected by different models remained strong (Tables 2 and 3). This finding aligns with previous research. For instance, Habier, Fernando, and Dekkers (2009) demonstrated, using simulated data, that a subset of low-density markers can effectively represent high-density SNP genotypes. Moreover, Wellmann et al. (2013), in a study involving a breeding line of pigs, showed that imputing low-density marker panels is a promising strategy, even when the low-density panel contains fewer than 1,000 markers. Likewise, Kriaridou et al. (2020) examined SNP panels of varying densities in a dataset featuring different aquaculture species, revealing that a low-density SNP panel can provide near-maximal prediction accuracy for most polygenic traits in aquaculture stock.

Starting with a panel of 2,047 markers, representing 10% of the total available SNPs, we explored the feasibility of utilizing even lower marker densities without compromising prediction quality. Subsequently, within the formed groups, we aimed to identify subsets by intersecting and uniting the 2,047 SNPs with the highest significance for each trait within their respective groups. We then evaluated these subsets in terms of predictive ability $(r_{y\hat{y}})$, standard errors, and agreement between models based on these subsets.

In Group 1 (comprising CD, BL, NVN, and Vig traits), the intersection of the 2,047 markers yielded 655 SNPs, equivalent to 32% of the original 2,047 and about 3.2% of the total 20,477 evaluated SNPs. The predictive ability estimates $(r_{y\hat{y}})$ or the traits within this Group (Table 4) remained consistent when compared to the model with 2,047 SNPs. This observation suggests that prediction quality remained unaffected despite the reduction in marker density.

In Group 2 (RFS and NF), the intersection of the 2,047 markers resulted in 890 SNPs, representing approximately 43.5% of the original 2,047 markers and about 4.3% of the total 20,477 SNPs. Within this group, the predictive ability of the model containing 890 SNPs remained consistent when compared to the model with 2,047 SNPs for the RFS trait (Table 4). Notably, the predictive ability for RFS in this study exceeded that reported for the same trait in *Coffea canephora* (Alkimim et al., 2017), where it was recorded as $r_{y\hat{y}} = 0.00$. However, for the NF trait, there was a slight decrease in predictive ability (Table 4).

Table 4. Mean predictive abilities of models with 2,047 SNPs, intersection sets, and union sets of SNPs.

Group	Traits	Initial Model *	SNPs of intersection	SNPs of the union of traits by group	SNPs of the union of all traits
1	CD	0.36 (0.02) [2047]	0.31 (0.02) [655]	0.35 (0.03) [3830]	0.35 (0.03) [5970]
	BL	0.36 (0.03) [2047]	0.32 (0.02) [655]	0.35 (0.02) [3830]	0.34 (0.02) [5970]
	NVN	0.26 (0.07) [2047]	0.18 (0.05) [655]	0.30 (0.08) [3830]	0.31 (0.07) [5970]
	Vig	0.30 (0.07) [2047]	0.22 (0.07) [655]	0.30 (0.08) [3830]	0.30 (0.08) [5970]
2	RFS	0.28 (0.07) [2047]	0.28 (0.07) [890]	0.23 (0.07) [3204]	0.20 (0.08) [5970]
	NF	0.29 (0.04) [2047]	0.18 (0.02) [890]	0.30 (0.03) [3204]	0.17 (0.03) [5970]
3	Cer	0.33 (0.08) [2047]	0.31 (0.08) [270]	0.31 (0.08) [3824]	0.31 (0.09) [5970]
	LM	0.12 (0.06) [2047]	0.18 (0.06) [270]	0.02 (0.08) [3824]	0.10 (0.07) [5970]

CD = Canopy diameter; BL = Branch length; NVN = Number of vegetative nodes; Vig = Vegetative Vigor; RFS = Ripening fruit size; NF = Total number of fruits; Cer = Cercosporiosis incidence; LM = Leaf miner infestation. *Model containing 10% of the markers with the best accuracy and concordance. (*) standard error of the estimates; and [*] number of SNPs considered in the model.

In Group 3 (Cer and LM), the intersection of the 2,047 SNPs resulted in 270 SNPs, constituting nearly 13% of the original 2,047 markers and just over 1% of the total 20,477 SNPs (Table 4). Regarding the Cer trait, there

was no statistically significant change in predictive ability when comparing the model containing 270 SNPs to the one with 2,047 SNPs. However, there was a slight increase in predictive ability for the LM trait (Table 4). It is worth noting that Sousa et al. (2019a) reported similar $r_{y\hat{y}}$ values for Cer (0.31) and LM (0.18) in their *C. arabica* studies.

The union of the 2,047 markers within Group 1 (CD, BL, NVN, and Vig) resulted in 3,830 SNPs, for Group 2 (RFS and NF), it yielded 3,204 SNPs, and for Group 3 (Cer and LM), it produced 3,824 SNPs. When considering all traits simultaneously (CD, BL, NVN, RFS, NF, Cer, and LM), it resulted in 5,970 SNPs (Table 4). In terms of predictive abilities, there were no statistically significant changes observed for the union of Group 1 traits when compared to the parameter estimates, indicating that prediction quality was maintained when increasing the density from 2,047 SNPs to 3,830 markers. This outcome was expected since, as shown in Table 2, marker densities ranging from 2,047 SNPs to 20,477 SNPs exhibited statistically similar values of $r_{y\hat{y}}$. This trend was also observed for the traits within Group 2.

In Group 3, the predictive ability estimated of $r_{y\hat{y}}$ for the density of 3,824 SNPs remained consistent when compared to the density of 2,047 SNPs for the Cer trait (Table 4). However, for the LM trait, there was a decrease in $r_{y\hat{y}}$ for the density of 3,824 markers compared to the density of 2,047 SNPs (Table 4).

Regarding the union of SNPs for all traits, the $r_{y\hat{y}}$ estimated of $r_{y\hat{y}}$ for the density of 5,970 SNPs did not exhibit statistically significant changes for CD, BL, NVN, Vig, Cer, and LM when compared to the density of 2,047 SNPs. However, there was a slight decrease in $r_{y\hat{y}}$ for RFS and NF concerning this parameter (Table 4).

In terms of concordances between individuals selected based on GEBV using all 20,477 SNPs and those chosen from the model fitted with the 655-marker density in the intersection of Group 1, there was substantial agreement for BL, moderate agreement for CD, and fair agreement for NVN (Table 5). Within the intersection of Group 2, the Kappa index remained at 0.62 for the RFS trait when considering the density of 890 SNPs compared to the density of 2,047 markers, indicating substantial agreement. For Group 3, the concordances were moderate for the individuals selected using GEBV based on the 20,477 SNPs and those selected from the model with 655 SNPs for both traits in this group.

Table 5. Cohen's Kappa coefficient (CK) of agreement between individuals selected based on genomic estimated breeding values (GEBV) using all (20,477) vs. reduced-density markers (selected by fitted models) with the different intersection and union densities.

Group		CK - Initial Model*	CK – intersection of traits by group	CK – union of traits by group	CK - union of all traits
1	CD	0.69 [2047]	0.53 [655]	0.69 [3830]	0.81 [5970]
	BL	0.87 [2047]	0.72 [655]	0.91 [3830]	0.94 [5970]
	NVN	0.65 [2047]	0.34 [655]	0.78 [3830]	0.84 [5970]
	Vig	0.65 [2047]	0.18 [655]	0.75 [3830]	0.87 [5970]
2	RFS	0.62 [2047]	0.62 [890]	0.75 [3204]	0.81 [5970]
	NF	0.78 [2047]	0.53[890]	0.87 [3204]	0.84 [5970]
3	Cer	0.69 [2047]	0.53 [270]	0.72 [3204]	0.81 [5970]
	LM	0.65 [2047]	0.56 [270]	0.65 [3204]	0.81 [5970]

CD = Canopy diameter; BL = Branch length; NVN = Number of vegetative nodes; Vig = Vegetative Vigor; RFS = Ripening fruit size; NF = Total number of fruits; Cer = Cercosporiosis incidence; LM = Leaf miner infestation. [*] number of SNPs considered in the model.

In the union groups, most of the traits exhibited increased Kappa index estimates compared to the group using 20,477 SNPs. Substantial agreement was observed for traits CD, NVN, and Vig in Group 1. A similar substantial agreement was found for RFS in Group 2 and for Cer and LM in Group 3. As for the union of SNPs across all traits, the Kappa index (CK) registered an overall increase in estimates for all traits (Table 5). This rise in concordances for the union groups and when considering all traits was expected because higher densities approach the 20,477 SNPs density, leading to increased concordance between individuals selected using GEBV with these SNPs and those selected from the model with varying densities.

In summary, using marker densities from the intersection of SNPs within groups, considering trait relationships, proves to be a viable strategy for enhancing *C. arabica* through GWS. Despite slight decreases in Kappa indexes for these group traits, individuals remained in agreement regarding phenotype and total SNP count.

Furthermore, employing the union of SNPs within the formed groups also appears to be a feasible strategy. It does not substantially alter predictive capabilities, and concordances between selected individuals remain noteworthy.

These strategies (SNP intersection and union) offer breeders the opportunity to reduce genotyping costs while improving *C. arabica* by selecting more productive, adapted, and high-quality beverage cultivars. Platforms allowing the use of a selected, cost-effective subset of SNPs with robust sequencing coverage can be implemented in practice.

Conclusion

Our findings highlight the effectiveness of low-density marker panels in genomic prediction, particularly when considering intersection groups and unions among markers for different traits in *Coffea arabica*. Models based on these panels consistently yield favorable predictive ability estimates and substantial concordance values among selected individuals when compared to models that use higher marker densities.

Acknowledgements

The authors are grateful for the financial support of the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* – FAPEMIG (APQ-01638-18), the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – CAPES (Code 001), the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* – CNPq (307798/2019-4 and 306772/2020-5) and the Brazilian Coffee Research and Development Consortium (*Consórcio Pesquisa Café*-CBP and D/Café).

References

Aliloo, H., Mrode, R., Okeyo, A. M., Ni, G., Goddard, M. E., & Gibson, J. P. (2018). The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. *Journal of Dairy Science*, *101*(10), 9108-9127. DOI: https://doi.org/10.3168/jds.2018-14621

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Pereira, A. A., Oliveira, A. C. B., Zambolim, L., & Sakiyama, N. S. (2017). Marker-assisted selection provides arabica coffee with genes from other *Coffea* species targeting on multiple resistance to rust and coffee berry disease. *Molecular Breeding*, *37*(6), 1-10. DOI: https://doi.org/10.1007/s11032-016-0609-1

Caixeta, E. T., Pestana, K. N., & Pestana, R. K. N. (2015). Melhoramento do cafeeiro: ênfase na aplicação dos marcadores moleculares. In G. O. Garcia, E. F. Reis, J. S. Lima, A. C. Xavier, & W. N. Rodrigues (Orgs.), *Tópicos especiais em produção vegetal V*. Alegre, ES: CAUFES.

Carvalho, C. H. S. (2008). *Cultivares de Café: origem, características e recomendações*. Brasília, DF: Embrapa Café.

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323-330. DOI: https://doi.org/10.1111/j.1439-0388.2007.00702.x

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2009). Genomic selection using-low density marker panels. *Genetics*, *182*(1), 343-353. DOI: https://doi.org/10.1534/genetics.108.100289

International Coffee Organization [ICO]. (2023). *Coffee Report and Outlook (CRO)*. Retrieved on Sep. 10, 2023 https://icocoffee.org/documents/cy2022-23/

Kriaridou, C., Tsairidou, S., Houston, R. D., & Robledo, D. (2020). Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. *Frontiers in Genetics*, *11*(124), 1-8. DOI: https://doi.org/10.3389/fgene.2020.00124

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. DOI: https://doi.org/10.2307/2529310

Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., ... Qiu, L. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Molecular Breeding*, *36*(113), 1-10. DOI: https://doi.org/10.1007/s11032-016-0504-9

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome wide dense marker maps. *Genetics*, *157*(4), 1819-1829. DOI: https://doi.org/10.1093/genetics/157.4.1819

Partelli, F. L., Covre, A. M., Oliveira, M. G., Alexandre, R. S., Vitória, E. L., & Silva, M.B. (2014). Root system distribution and yield of 'Conilon' coffee propagated by seeds or cuttings. *Pesquisa Agropecuária Brasileira*, *49*(5), 349-355. DOI: https://doi.org/10.1590/S0100-204X2014000500004

Resende, M. D. V. (2016). Software Selegen-REML/BLUP: a useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*, *16*(4), 330-339. DOI: https://doi.org/10.1590/1984-70332016v16n4a49

Resende, M. D. V., Silva, F. F., Lopes, O. S., & Azevedo, C. F. (2012). *Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial*. Viçosa, MG: Departamento de Estatística/UFV.

Santana, L. S., Silva Ferraz, G. A., Teodoro, A. J. S., Santana, M. S., Rossi, G., & Palchetti, E. (2021). Advances in precision coffee growing research: A bibliometric review. *Agronomy*, *11*(8), 1-16. DOI: https://doi.org/10.3390/agronomy11081557

Sousa, T. V., Caixeta, E. T. C., Alkimim, E. R., Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., ... Resende, M. D. V. (2019a). Early selection enable by the implementation of genomic selection in *Coffea arabica* Breeding. *Frontiers in Plant Science*, *9*(1934), 1-12. DOI: https://doi.org/10.3389/fpls.2018.01934

Sousa, M. B., Galli, G., Lyra, D. H., Granato, I. S. C., Matias, F. I., Alves, F. C., & Fritsche-Neto, R. (2019b). Increasing accuracy and reducing costs of genomic prediction by marker selection. *Euphytica*, *215*(18), 215-218. DOI: https://doi.org/10.1007/s10681-019-2339-z

Vidal, R. O., Mondego, J. M. C., Pot, D., Ambrósio, A. B., Andrade, A. C., Pereira, L. F. P., ... Pereira, G. A. G. (2010). A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. *Plant Physiology*, *154*(3), 1053-1066. DOI: https://doi.org/10.1104/pp.110.162438

Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, *195*(4), 1223-1230. DOI: https://doi.org/10.1534/genetics.113.155176

Wellmann, R., PreuB, S., Tholen, E., Heinkel, J., Wimmers, K., & Bennewitz, J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution*, *45*(28), 1-11. DOI: https://doi.org/10.1186/1297-9686-45-28

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, *10*(189), 1-10. DOI: https://doi.org/10.3389/fgene.2019.00189