Tree Genetics & Genomes (2017) 13: 95
DOI 10.1007/s11295-017-1171-7

ORIGINAL ARTICLE

A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*

Luís Felipe Ventorim Ferrão[1] · Romário Gava Ferrão[2] · Maria Amélia Gava Ferrão[3] · Aymbiré Francisco[3] · Antonio Augusto Franco Garcia[1]

Received: 6 February 2017 / Revised: 23 June 2017 / Accepted: 4 July 2017 / Published online: 14 August 2017
© Springer-Verlag GmbH Germany 2017

Abstract Genomic selection (GS) has been studied in several crops to increase the rates of genetic gain and reduce the length of breeding cycles. Despite its relevance, there are only a modest number of reports applied to the genus *Coffea*. Effective implementation depends on the ability to consider genomic models, which correctly represent breeding scenario in which the species are inserted. Coffee experimentation, in general, is represented by evaluations in multiple locations and harvests to understand the interaction and predict the performance of untested genotypes. Therefore, the main objective of this study was to investigate GS models suitable for use in *Coffea canephora*. An expansion of traditional GBLUP was considered and genomic analysis was performed using a genotyping-by-sequencing (GBS) approach, showed good potential to be used in coffee breeding programs. Interactions were modeled using the multiplicative mixed model theory, which is commonly used in multi-environment trials (MET) analysis in perennial crops. The effectiveness of the method used was compared with other genetic models in terms of goodness-of-fit statistics and prediction accuracy. Different scenarios that mimic coffee breeding were used in the cross-validation process. The method used had the lowest AIC and BIC values and, consequently, the best fit. In terms of predictive ability, the incorporation of the MET modeling showed higher accuracy (on average 10–17% higher) and lower prediction errors than traditional GBLUP. The results may be used as basis for additional studies into the genus *Coffea* and can be expanded for similar perennial crops.

Keywords Genomic selection · Genotyping-by-sequencing(GBS) · GBLUP · Multi-environment trials (MET) · Perennial crops

Communicated by: J. Beaulieu

This article is part of the Topical Collection on *Breeding*

Key Message: First insights into the Genotyping-by-Sequencing (GBS) in *Coffea canephora* and a genomic prediction model considering the theory about multiplicative mixed model to accommodate the interaction effects across sites and harvests

✉ Antonio Augusto Franco Garcia
augusto.garcia@usp.br

[1] Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), Av. Padua Dias, 11, CP 83, CEP 13400-970, Piracicaba, SP, Brazil

[2] Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper), Rua Afonso Sarlo, 160, Bento Ferreira, CEP 29052-010, Vitória, ES, Brazil

[3] Embrapa Café/Incaper, Rua Afonso Sarlo, 160, Bento Ferreira, CEP 29052-010, Vitória, ES, Brazil

Introduction

Coffee is one of the most important global crops in terms of economic and social implications. Brazil is responsible for about a third of the world's production making it the world's largest producer. Brazil has held this position for the last 150 years (IOC 2016). The *Coffea* genus comprises hundreds of tropical species and the beverage popularly known as coffee is produced from grains of two species: *Coffea arabica*, which contributes to the aroma and sweet flavor; and *Coffea canephora*, with higher amounts of caffeine and soluble solids (Tran et al. 2016). Global efforts have

been made to increase production and quality of the final product. Thus, breeding programs have a key role in improving agronomic traits associated with grain production (Ferrão et al. 2015)

C. canephora is a good starting point for studies on the *Coffea* genus for economic and genetic reasons including the ploidy (2n = 2x) and wide genetic variability (Tran et al. 2016). Both features make the process of genotyping and statistical modeling more feasible than in *C. arabica*, which is allotetraploid and has a narrow genetic base. The economic motivation is based on grain production and crop cultivation. *C. canephora* is responsible for 40% of the world coffee production, and its grain is the main source of raw materials for soluble coffee. Further, the species has better adaptability to various environmental stresses, which makes cultivation easier and cheaper (Ferrão et al. 2007).

Traditionally, evaluation of genetic progress has been performed via phenotype data collected in field trials coupled with a long testing phase, which results in low gains per unit of time. The advent of molecular markers opened a new perspective for their use in marker-assisted selection (MAS). Meuwissen et al. (2001) suggested the use of all available molecular markers as covariates in linear regression models to predict genetic value in quantitative traits. Called genomic selection (GS), this methodology has the potential to redirect resources and activities in breeding programs and to reduce breeding cycles and increase genetic gains per time unit, especially in animal and plant breeding (de los Campos et al. 2009).

Although GS is promising for breeders, studies in coffee are still emerging in contrast to other crops. Implementing GS poses several statistical challenges such as the ability to consider genomic models that represent the breeding scenario to which the species is submitted. Typically, coffee trials are tested in multiple locations and harvests to evaluate interactions and predict the performance of untested genotypes. Such experiments are collectively referred to as multi-environment trials (MET) and are not restricted to coffee but are also used in many perennial crops (Smith et al. 2001; Kelly et al. 2009; Malosetti et al. 2014).

Numerous statistical models have been developed to evaluate interactions in MET studies. In a modern framework, the genotypic performance across environments has been modeled as correlated traits. Thus, structured and unstructured covariance functions have been utilized in a mixed model context (Smith et al. 2005; Kelly et al. 2009; Pastina et al. 2012; Malosetti et al. 2014). A main advantage is the flexible way in which these functions can be tested to describe the interactions and the residual term (Smith et al. 2001). Furthermore, when genetic effects are assumed to be random, the pedigree information can be incorporated and more accurate breeding values may be computed using best linear unbiased prediction (BLUP) (Kelly et al. 2009).

BLUP methodology relies on pedigree information to estimate the covariance between known relatives. However, this covariance can also be estimated using genomic information rather than an expected value based on the pedigree record. A matrix built with genomic information is named the genomic relationship matrix, and its combination with the BLUP theory resulted in the so-called Genomic Best Linear Unbiased Prediction (GBLUP) (VanRaden 2008). This is the current gold standard GS method used in animal and plant breeding (de los Campos et al. 2013). One of the first ideas to accommodate the interaction in GS models was described by Burgueño et al. (2012). For this purpose, the traditional GBLUP was extended to accommodate covariance functions in a multiple environment context. Among the theoretical and practical advantages, this approach used a consolidated theory about mixed models as well as straightforward implementation using existing software. More recent studies have been advanced to incorporate modern information about environmental covariates (Jarquín et al. 2014a; Heslot et al. 2014). Other studies have reported the explicit modeling between markers and environment (Schulz-Streeck et al. 2013; Lopez-Cruz et al. 2015). Recently, an in-depth description of issues related to interactions on GS studies was presented by Malosetti et al. (2016). All of these authors showed that models including the interaction resulted in substantial gains in prediction accuracy.

Although promising, all these methods do not address an important aspect of perennial crops: having data from multiple harvests and a short sequence of repeated measurements. Longitudinal data of this nature are common not only in coffee but also in other crops such as sugarcane (Pastina et al. 2012; Margarido et al. 2015), forage grass (Smith and Casler 2004), and cereal (Kelly et al. 2009). In this context, effective implementation of GS methods depends on the ability of the model to predict real conditions in breeding programs. Statistical challenges create more complex scenarios. Hence, in MET analysis, the main challenge is to properly consider the genetic and the environment effects, because it involves a multidimensional space with a variation that is defined by the effects of locations, years and their interactions with genotypes (Malosetti et al. 2016).

In addition to statistical challenges, the modest number of reports considering high-throughput genotyping also hampers genomic studies in coffee. Genotyping-by-sequencing (GBS) is representative of this new class of molecular markers, which combines the reduction in genomic complexity with next-generation sequencing (NGS) (Elshire et al. 2011). A single sequencing run on an NGS platform can generate data on the gigabase-pair levels. This usually contains hundreds of thousands of SNPs. Therefore, in a one-step approach, GBS can make it possible to discover new markers and genotype entire populations. It is

rapid, flexible, and perfectly suited for GS compared to traditional molecular markers. Research using GBS is common in many crops (Poland et al. 2012; Crossa et al. 2013; Jarquín et al. 2014b), but there is a still an important gap in the coffee literature.

The main objective of this research was to consider a genomic selection model suitable for use in *C. canephora* and other crops with similar experimental design. This model addresses issues related to the breeding strategy used for that species, including sources of interaction. We present aspects related to the applicability of genotyping-by-sequencing (GBS) as well as future perspectives.

Material and methods

Phenotypic data

The experimental population was developed and evaluated by the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper); ES State, Brazil. It consisted of a recurrent selection population formed from the recombination of 16 superior clones of *C. canephora*. Of the thousands of genotypes maintained in Incaper, these superior clones were selected as progenitor due to the high production and the similar grain maturity date. The latter is an important trait for new coffee varieties because it allows for harvest standardization.

After one cycle of recombination (open-pollination), 103 progenies and the 16 progenitors were cloned and evaluated in randomized complete blocks with three repetitions and five plants per plot. The population was installed in two representative environments (locations) for the Brazilian production of *C. canephora*: Marilândia Experimental Farm (FEM) - latitude 19°24′ south, longitude 40°31′ west, 70 m altitude; and Sooretama Experimental Farm (FES) - latitude 15°47′ south, longitude 43°18′ west, 40 m altitude. The complete experiment was made of 3570 coffee trees and total grain production (kilograms of mature coffee fruit in the cherries stages) of each progeny was evaluated over four consecutive harvest-production years (2008, 2009, 2010, and 2011).

Genotypic data

The GBS protocol followed that from the Genomic Diversity Facility, Cornell University (http://www.biotech.cornell.edu/brc/genomics-facility). Leaves of each of the 103 progenies and 16 progenitors were collected and lyophilized. DNA extraction was made using Qiagen DNeasy Plant and the genomic libraries were prepared following (Elshire et al. 2011). DNA samples were digested using the *ApeKI* restriction enzyme, and 96 samples were multiplexed per Illumina flow cell for sequencing.

The GBS analysis pipeline implemented in TASSEL-GBS (v.4.3.7) (Glaubitz et al. 2014) was used to analyze sequence data. Sequenced tags were aligned against the *C. canephora* reference genome sequence (Denoeud et al. 2014). The raw Variant Call Format (VCF file) was filtered manually considering the following cutoff: (i) triallelic SNPs were removed; (ii) minimum minor allele frequency (0.01 MAF); (iii) SNPs that are present in less than $< 50\%$ of the samples were eliminated; (iv) minimal depth coverage of $10\times$ (the mean number of sequence reads per locus averaged across all individuals) was considered.

All filtering and SNP manipulation was carried out using VCFtools package (Danecek et al. 2011) and customized scripts in R (R Core Team 2013) and bash (GNU 2007). GBS markers which had up to 50% missing data were imputed using the mean value for each marker. The graphical analyzes were performed using the OmicCircos (Hu et al. 2014).

Phenotypic models

Phenotypic data were analyzed with the following model, which uses a notation presented by Pastina et al. (2012). The statistical model in which the underlined terms indicate a random variable is:

$$\underline{y}_{ijkr} = \mu + L_j + B|L_{rj} + H_k + LH_{jk} + \underline{G}_{ijk} + \underline{e}_{ijkr} \quad (1)$$

Here, $\underline{y}_{ijkr}$ is the phenotype of the r^{th} block (r=1,2,3) of the i^{th} individual (i = 1,2...,n), of the j^{th} location (j=1,2) and k^{th} harvest (k = 1,2,3,4). Term μ is the overall mean; L_j is the effect of location; $B|L_{rj}$ is the block effect nested within location; H_k is the harvest effect; LH_{jk} is the location by harvest interaction; $\underline{G}_{ijk}$ is a random genetic effect of individual i, at harvests k and location j; and $\underline{\epsilon}_{ijkr}$ is the random non-genetic residual error term.

For the genetic effects, we assumed a multivariate normal distribution with a zero mean vector and a variance-covariance (VCOV) matrix indexed by three factors (harvest, location and genotype) written as the Kronecker product ($\otimes$) of matrices as follows: $G = G_H^{k\times k} \otimes G_L^{j\times j} \otimes \Sigma_g^{n\times n}$ in which G_H and G_L are VCOV and relate to harvest and location. The diagonal element of these matrices represents the genetic variance within the k^{th} harvest and the genetic variance within the j^{th} location, respectively. The VCOV structures for these matrices are represented in Table 1. For G_L, the reduced number of locations (two) restricted the search in three VCOV structures (ID, DIAG and UNS), while for G_H all the VCOV structures cited were tested. Two important points deserve comments: (i) each structure has different assumptions about the heterogeneity of variance and may be used to quantify the interactions and (ii)

the number of estimated parameters represents the variation in the degree of complexity.

The term Σ_g is used here as a generic form to highlight the different assumptions that can be assumed for the genetic term. The off-diagonal elements are the genetic covariance (Σ_g). An Identity matrix (I_g) is used when it is reasonable to assume that the genotypes are not related to each other (same variance and lack of covariance between individuals). The Identity assumption ensures that the breeding values of each genotype will be predicted only by the value of the empirical responses of the genotype itself. This is an assumption often used in family studies in the absence of pedigree information. However, information about the genetic relationship may be incorporated in the presence of pedigree record or molecular information. Variations in these genetic assumptions and the interaction accommodation were the central point of this study. This will be presented in the next section.

The residual term was factored in similarly to genetic effects. It is assumed to be a multivariate normal distribution implying a zero mean and VCOV matrix indexed by four factor (harvest, location, block and genotype) written using the Kronecker product as follows: $R = R_H^{k\times k} \otimes R_L^{j\times j} \otimes R_B^{r\times r} \otimes I_g^{n\times n}$, in which R_H, R_L and R_B are VCOV tested for harvest, location and block, respectively. The I_g is an Identity residual (co)variance matrix. In principle, all the structures mentioned in Table 1 were tested for the residual term. In addition, spatial adjustments were tested, in order to adjust for possible trends in the field trial data. An autoregressive (AR1) structure that allows correlations between the residual values in neighboring plots (both within rows and within columns) was considered.

GBLUP version for multiple harvest-location trials (MET-GBLUP)

The aforementioned Model 1 was used to test for the presence of interactions (Genotype × Location- G×L and Genotype × Harvest- G×H) and the inclusion of molecular information in prediction models. Thus, different assumptions about the random effects distribution were tested. Two classes of models were defined in accordance with the inclusion of interaction terms (MET modeling) (Table 2).

The first class of methods ignored the MET modeling and simple structures for genetic and residual random effects were assumed. Initially, the absence of genetic relationship across individuals was assumed (**Id** method). The **BLUP** method considered the additive relationship matrix (A_p) as genetic covariance between individuals, while the **GBLUP** method considered the realized kinship (A_m). The A_p matrix was based on the numerator relationship matrix, which was computed from the coefficient of co-ancestry (termed as θ_{xy}) between genotypes x and y as $A_p = \{2\theta_{xy}\}$. This assumed that relatives are not inbred (Falconer and Mackay 1996). The A_m matrix, often called realized genomic relationship matrix or G-matrix, was computed using molecular marker information. To illustrate, let $\mathbf{X} \in \{0, 1, 2\}^{n\times m}$ be the genotype matrix for n individuals and m biallelic SNP markers with alleles designed A and B and marker scores coded: AA=0, AB=1 and BB=2. Let the frequency of the B allele at locus k be p_k. The $\mathbf{Z}$ matrix denotes the centered genotype matrix constructed by subtracting the marker mean from each data point: $Z_{ik} = X_{ik} - 2p_k$. Realized relationship matrix (A_m) can be obtained by VanRaden (2008): $A_m = \frac{Z'Z}{2\sum p_k(1-p_k)}$. Division by $2\sum p_i(1-p_i)$ scales the A_m matrix to be analogous to A_p matrix.

The second class of methods considered the MET modeling. Here, the genetic and residual matrices were modeled considering the structures cited in Table 1 as well as variations of the genetic covariances (Σ_g matrix). The **MET** method regarded the interactions, but had no correlation imposed by the pedigree. The **MET.BLUP** refers to an expansion of the BLUP model but accommodates MET modeling. **MET.GBLUP** is simultaneous MET modeling with the use of molecular markers to estimate the relationship matrix (A_m). The last approach was termed as "GBLUP version to multiple harvest-location trials" and refers to

Table 1 Variance and covariance structures examined for the random effects in model 1

Model	Num.Par[a]	Description
ID	1	Identical variation
DIAG	M	Heterogeneous variations
CS	2	Compound symmetry with homogeneous variance
CS_Het	M+1	Compound symmetry with heterogeneous variance
FA1	2M	First order factor analytic model
AR1	M+1	First order autoregressive model
UNS	M(M+1)/2	Unstructured model

[a]The number of parameters for the models follows from the sum of the parameters for the component matrices minus the number of identification constraints. M = J or K, where J is the number of locations and K is the number of harvests

Table 2 Summary of the tested models and the assumption on the variance and covariance structure related to the random effects specified in the Model 1 description. MET prefix on the name of each method indicates models where the interaction is explicitly modeled, testing covariance structures for location and harvest

Method	G[a]	R[a]
Id[1]	$I_G^{n\times n}$	$I_G^{n\times n}$
BLUP[1]	$A_p^{n\times n}$	$I_G^{n\times n}$
GBLUP[1]	$A_m^{n\times n}$	$I_G^{n\times n}$
MET[2]	$G_L^{j\times j}\otimes G_H^{k\times k}\otimes I_G^{n\times n}$	$R_L^{j\times j}\otimes R_H^{k\times k}\otimes R_B^{r\times r}\otimes I_G^{n\times n}$
MET.BLUP[2]	$G_L^{j\times j}\otimes G_H^{k\times k}\otimes A_p^{n\times n}$	$R_L^{j\times j}\otimes R_H^{k\times k}\otimes R_B^{r\times r}\otimes I_G^{n\times n}$
MET.GBLUP[2]	$G_L^{j\times j}\otimes G_H^{k\times k}\otimes A_m^{n\times n}$	$R_L^{j\times j}\otimes R_H^{k\times k}\otimes R_B^{r\times r}\otimes I_G^{n\times n}$

[a]Variance and covariance structures tested for the random effects specified in the Model 1. The I_G, A_p and A_m represent a Identify matrix, additive relationship matrix and genomic relationship matrix, respectively.
[1] First class of methods, that ignored the Multi-Environment Trials (MET) modeling; [2] Second class of methods, that considered the MET modeling

the idea of accommodating the G×L and G×H interactions using MET theory and a genomic selection model (GBLUP).

All the fitted models were performed in Genstat 14th edition (Payne et al. 2011) using Restricted Maximum Likelihood (REML). Additive relationship matrix (A_p) was computed using the pedigreem R package (Bates and Vazquez 2014). Realized genomic relationship (A_m) was computed using customized scripts in R (R Core Team 2013).

Comparison of models

Two criteria were used to compare the models (Table 2): (i) goodness-of-fit statistics, via AIC (Akaike 1974) and BIC (Schwarz 1978) and (ii) predictive ability measured by cross-validation. Three cross-validation schemes were considered. Scenario 1 (CV1) aims to evaluate the predictive ability for genotypes that have not undergone field evaluation (i.e., mimic situations with genotype that were not evaluated in any block, location and harvest). Scenario 2 (CV2) aims to make predictions for one specific location and scenario 3 (CV3) for one specific harvest. The simulated scenarios ranged in complexity, the largest number of predictions was being made in CV1 followed by CV2 and CV3.

The predictive ability were assessed using a Replicated Training-Testing evaluation. In each replication, 90% of the individuals (107 genotypes) were assigned randomly for training data set (TRN), while the remaining 10% were assigned for testing data set (TST). This division was replicated 10 times with independent random assignments into TRN and TST. A similar scheme was used by Crossa et al. (2013). The predictive capacity was measured using the average predictive ability and the mean squared prediction error (MSPE) across the 10 repetitions. The predictive ability was computed via the Pearson correlation between predicted ($\hat{y}_i$) and observed values (y_i). The MSPE was computed by the formula: $MSPE = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$, where n is the number of individuals that predicted in the TST.

Results

Phenotypic data

The lowest AIC and BIC values were observed for the combination of UNS form for location (G_L) and harvest (G_H) (Table 3). The values of ID combinations highlight the poor quality of the goodness-of-fit value when traditional ANOVA assumptions are considered—even when homogeneous variances across locations and harvests are applied.

All the structures mentioned in Table 1 were also tested for the residuals. Convergence problems and negative variance components were however observed when more complex models were tested (results not shown). Therefore, the DIAG form was assumed for each factor in the residual matrix. The option of a simple structure was based on reducing the complexity and number of estimated parameters. This is because our main focus was the genetic part. Although this structure may not be the most suitable for representing residuals, this model is more realistic than the assumptions assumed in the traditional ANOVA that consider an ID structure for each factor, and consequently, homogeneity between locations, harvests, and blocks (Smith et al. 2001). In addition, spatial adjustment was tested to correct for possible trends in the field trial data. No improvements on the AIC and BIC criterion were observed when data were adjusted for neighboring plots (results not shown).

Figure 1 presents the phenotypic dispersion across the harvests and the variance component magnitude. The dispersion of the phenotypic observations shows that the FES location was more productive (on average) than FEM. There was more variation in the FES. Evidence of G×L was first observed via this production difference and confirmed via heterogeneity of variance across locations. There was an important pattern observed across the harvests: a lack of annual production stability. The boxplot highlights cyclical production including highly productive years (2008 and

Table 3 Goodness-of-fit statistics considering the AIC and BIC criteria evaluated for grain production in a *Coffea canephora* population. The genetic matrix was factored by location (G_L) and harvest (G_H). A Identify matrix was considered for the residual random effect

	G_L					
	ID		DIAG		UNS	
G_H	AIC	BIC	AIC	BIC	AIC	BIC
ID	20027.5	20039.37	20014.6	20032.42	19962.09	19985.84
DIAG	20010.39	20040.08	19997.04	20032.66	19948.37	19989.93
AR1	19974.45	19992.26	19966.07	19989.83	19922.81	19952.5
FA1	19885.61	19939.05	19878.38	19937.75	19844.49	19909.81
CS	19939.21	19957.02	19932.66	19956.41	19892.13	19921.82
CS_het	19920.23	19955.85	19913.37	19954.93	19876.05	19923.55
UNS	19854.45	19919.77	19848.4	19919.66	**19811.65**	**19888.84**

ID:Identical variation; DIAG: Heterogeneous variations; CS: compound symmetry with homogeneous variance; CS_het: compound symmetry with heterogeneous variance; FA1: first order factor analytic; AR1: first order autoregressive; UNS: unstructured model.
Bold numbers represent the smallest AIC and BIC values, indicating the best fitted phenotypic model

2010) and low production years (2009 and 2011). Lack of stability and, consequently, evidence of G×H interactions were quantified via the UNS form fitted for G_H. This is represented by low genetic correlations between subsequent years. These results are clear indications of the importance of MET modeling for subsequent GS models.

Genotypic data

A total of 5,198,498 unique 64-bp sequence tags were identified in the *C. canephora* libraries; 32.1% were uniquely aligned to the reference genome, 7% were aligned to multiple positions, and 60.9% could not be aligned. Of this total, 449,467 raw SNPs were identified in the unfiltered VCF file.

We noted a predominance of SNPs with low percentages of missing data (0-10%). SNPs in chromosomes with more than 80% missing data were unusual. The number of SNPs per chromosome ranged from 24497 to 77635 (Table 4). An abrupt decrease was observed for the MAF cutoff and when the depth coverage increases. The SNP density before and after filtration was 449,467 and 13,117 SNPs, respectively. This represented 2.91% of the unfiltered SNP, but 15× is an extremely conservative value for cutoff in depth coverage. Therefore, for subsequent genomic studies, a security coverage of 10x was assumed (18,586 SNPs selected).

A summary of GBS results is presented in layers (Fig. 2). The first (from outer to inner layers) represents each chromosome with a specific color. The scale is proportional to the reference genome size. For better representation, all parameters in the subsequent layers were computed considering the average in a window of 400,000 base pairs (bp). The second layer is the number of raw SNPs per window. Unique tag counts were higher in the chromosome ends versus to pericentromic regions. The third layer is the depth coverage per window and ranged from 1 to 38 reads. The fourth and fifth layer are the percentage of SNPs per window with Minor Allele Frequency (MAF) lower than or

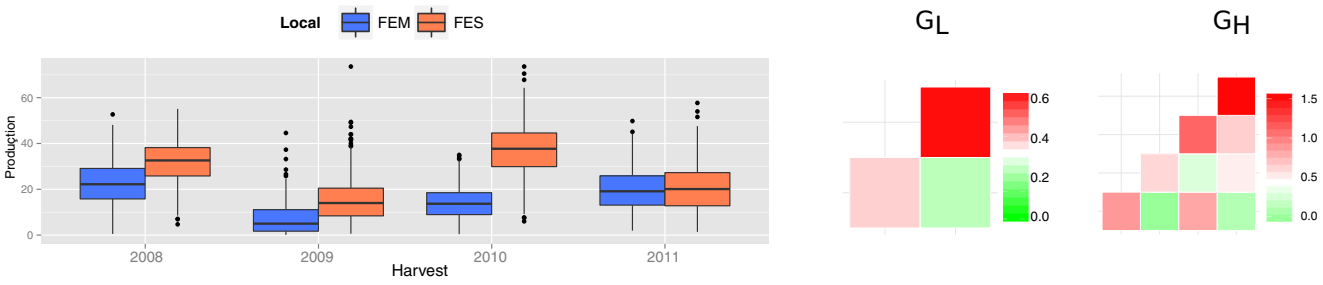

Fig. 1 Boxplot of grain production (kilograms of mature coffee fruit in the cherries stages) across the locations (FEM and FES) and harvests (2008, 2009, 2010 and 2011), and a heatmap representing the unstructured form estimated for locations (G_L) and harvests (G_H)

Table 4 Number of SNP markers per chromosome (Chr) before and after filter in *Coffea canephora* GBS libraries. A sequential filtering was considered: i) Triall: removing all triallelic SNPs; ii) MAF: removing SNPs with MAF < 0.01, plus Triall filter; iii) MD: removing SNPs that are present in less than < 50% of the samples, plus Triall and MAF filters; iv) Depth Coverage: removing SNPs with mean number of sequence reads per locus averaged across all individuals less than 1x, 5x, 10x and 15x, plus Triall, MAF and MD filters

Chr	Raw[a]	Triall	MAF	MD	1x	5x	10x	15x
Chr 1	46897	43692	16810	8679	8296	3359	1987	1400
Chr 2	77635	72150	26621	13805	13133	5470	3094	2164
Chr 3	31799	29728	12127	67901	6460	2771	1572	1131
Chr 4	34713	32368	11153	5870	5576	2252	1329	953
Chr 5	34140	31842	13263	6700	6361	2674	1509	1047
Chr 6	48775	45417	15822	8157	7686	2984	1722	1263
Chr 7	44370	41160	15197	8011	7594	2981	1728	1235
Chr 8	34554	32229	11678	5864	5612	2411	1373	965
Chr 9	24497	22859	8174	4332	4153	1752	1017	726
Chr 10	34158	31847	11786	6305	6014	2567	1563	1075
Chr 11	37929	35397	14990	7917	7553	2944	1692	1158
Total	449467	418689	158621	82431	78438	32165	18586	13117
(%)[b]	100	93.15	35.3	18.34	17.45	7.15	4.13	2.91

[a]Raw SNPs: original number of SNP markers per chromosome

[b]Percentage of SNPs remaining after the sequential filtering

equal to 5 and 1%, respectively. The sixth layer indicates the percentage of missing data. This ranged from 3 to 63% of missing data across the chromosome. The last layer is the SNP density after filtering and is composed of two colors, the gray background is the number of unfiltered SNPs, and the blue bars are the density after filtering.

MET and GS models

Models that ignored the MET modeling (**Id**, **BLUP** and **GBLUP**) showed higher AIC and BIC values and hence poor fit for grain production trait (Table 5). The inclusion of molecular information consistently improved the results based on the criteria of minimum AIC and BIC. The **MET.GBLUP** had the lowest AIC and BIC values and was the best model.

Models that included the MET modeling had better predictive ability (Table 6). In the most complex scenario (CV1), the difference in predictive ability between the **MET.GBLUP** method and traditional **GBLUP** was on the order of 10%. In CV2, this difference was higher (17%) and showed how problematic it can be to ignore the interaction to realize predictions. For CV3, a lower number of predictions was required, and the lowest differences were observed across the models (1%). In all scenarios, methods that ignored the MET modeling had very similar predictive ability implying that inclusion of molecular information could not improve the predictive ability over that obtained with pedigrees.

Another comparative criterion used during the cross-validation was the MSPE, which was held in the perception of the distance among observed and predicted values. Phenotypic metrics evaluated in field were considered the observed values, while predicted values were the adjusted means. The **MET.GBLUP** showed good results across the scenarios. For grain production, models that ignore the MET modeling generally, showed the highest MSPE values; the exception was the **GBLUP** in the CV1. The lower values of MSPE for the CV3 suggest that this scenario is less complex in terms of prediction.

Discussion

The potential of GS to accelerate crop improvement due to shorter generation times and the avoidance of phenotypic evaluation has been established and widely appreciated in plant and animal breeding (Jannink et al. 2010; de los Campos et al. 2013). In coffee, the reduction of repeated cycles of selection, breeding, and testing are our main motivation. Developing new cultivars can take decades, but this can be accelerated with the incorporation of GS concepts in breeding schemes. Good prospects have been reported in maize (Crossa et al. 2013), wheat (Poland et al. 2012), rice (Spindel et al. 2016) and forest tree species (Grattapaglia and Resende 2010). In this research, we considered an expansion of traditional GBLUP to address the conjugate

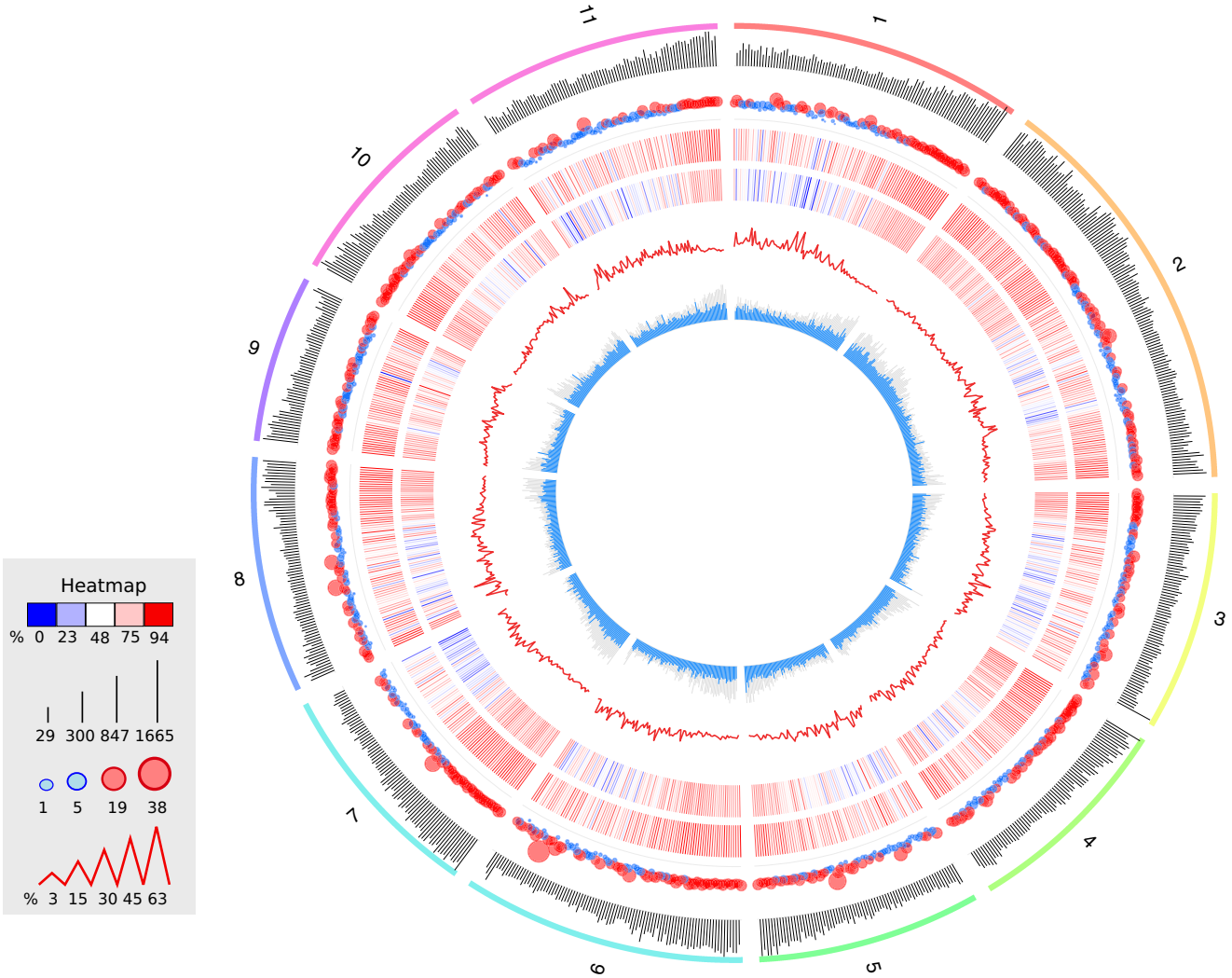

Fig. 2 Circular visualization of GBS information across the *Coffea canephora* chromosomes. From outer to inner layers, the graphic is separated in seven layers: i) Chromosomes; ii) number of raw SNPs; iii) depth coverage; iv) percentage of SNPs eliminated considering the Minor Allele Frequency (MAF) lower than 5%; v) percentage of SNPs eliminated considering the MAF lower than 1%; vi) percentage of missing data; vii) number of filtered SNPs (blues bars) in contrast with the number of raw SNPs (gray background). All these metrics were computed considering the average in a window size of 400,000 base pairs (bp). The scale, in the bottom left, aids in the perception on the magnitude of the values

Table 5 AIC and BIC values for models with different variance and covariance structures for the genetic and residual random effects, evaluated for grain production in a *Coffea canephora* population. MET prefix on the name of each method indicates models where the interaction is explicitly modeled, testing covariance structures for location and harvest

Method	Genetic matrix[a]	Residual matrix	AIC	BIC
Id	I_G	I_G	20753.25	20765.12
BLUP	A_p	I_G	20758.13	20770.01
GBLUP	A_m	I_G	20741.60	20753.50
MET	$G_L \otimes G_H \otimes I_G$	$R_L \otimes R_H \otimes R_B \otimes I_G$	19723.10	19835.92
MET.BLUP	$G_L \otimes G_H \otimes A_p$	$R_L \otimes R_H \otimes R_B \otimes I_G$	19705.38	19818.20
MET.GBLUP	$G_L \otimes G_H \otimes A_m$	$R_L \otimes R_H \otimes R_B \otimes I_G$	*19689.75*	*19802.56*

Italicized numbers represent the smallest AIC and BIC values, indicating the best fitted method

[a]Variance and covariance structures tested for the random effects specified in the Model 1. The I_G, A_p and A_m represent a Identify, additive relationship and realized kinship matrix, respectively

Table 6 Correlation between the predicted breeding values and the observed phenotypic values measured by the predictive ability (r) and mean squared prediction error (MSPE) considering six modeling methods for grain production in a *Coffea canephora* population. Three breeding scenarios were considered: Scenario 1 (CV1) aims to evaluate the predictive ability for genotypes that have not undergone field evaluation (i.e., mimic situations which genotype that were not evaluated in any block, location and harvest). Scenario 2 (CV2) aims to make predictions for one specific location and scenario 3 (CV3) for one specific harvest

	CV1		CV2		CV3	
Method	r	MSPE	r	MSPE	r	MSPE
Id	0.676	288.878	0.498	182.741	0.854	135.787
BLUP	0.676	266.75	0.498	180.098	0.854	140.949
GBLUP	0.676	*241.277*	0.498	171.566	0.854	140.919
MET	0.760	290.686	*0.677*	107.771	0.865	*103.843*
MET.BLUP	0.767	264.801	*0.677*	97.814	*0.866*	108.722
MET.GBLUP	*0.774*	244.864	0.670	*93.537*	0.864	111.227

Italicized numbers represent the greatest *r* values and the smallest MSPE

use of genomic information and MET modeling. A similar approach was described by Burgueño et al. (2012) and Oakey et al. (2016), although certain differences have been considered here, including the explicit G×H interaction modeling and a higher number of VCOV structures tested. It is noteworthy that our model could be considered for other perennial or annual species with a similar experimental design.

Multiplicative mixed models have been commonly used for MET analysis (Smith et al. 2001; 2005; Malosetti et al. 2014). The G matrix in MET models is a genotypic covariance matrix that is defined for the genetic random effect that was decomposed into harvest,locations and genotypes, i.e, $G = G_H^{k\times k} \otimes G_L^{j\times j} \otimes \Sigma_g^{n\times n}$. The term $\Sigma_g^{n\times n}$ can be used to include different assumptions for the genetic term. These assumptions reflected independence among genotypes (I_g) or similarities in terms of pedigree records (A_p) or DNA information (A_m). The $G_H^{k\times k}$ assumed correlation between harvests and $G_L^{j\times j}$ among locations. All these components jointly determined similarities among genetic effects across locations and harvests. Strictly speaking, the genotype and environmental interactions were modeled by considering that different genotypes do not necessarily react similarly to equal conditions. Information could be borrowed via a multidimensional genotypic space that is defined as the genotype-location-harvest combination. This offers predictions for the untested genotypes (Malosetti et al. 2016).

It is important to test for an appropriate VCOV structure in terms of harvest and location. These structures will reflect the nature of the interactions. Kelly et al. (2009) and Meyer (2009) reported that the most general form is the fully unstructured (UNS) matrix, although it often leads to estimation issues. A common solution is the factor analytic (FA) form—an intermediate structure in terms of parsimony and flexibility (Crossa et al. 2013). In this study, the reduced number of locations and harvests motivated a test of different VCOV structures to find the best biological description. Pastina et al. (2012), Margarido et al. (2015), and Oakey et al. (2016) reported a similar approach. This search was not fixed solely on the FA form. For the residual effect, we assumed a block diagonal structure (heteroscedasticity) where each location, harvest and block has its own component of residual variance. Although spatial analysis is an important alternative in data analysis of field experiments in plant, no improvement in the goodness-of-fit statistics was observed when spatial correlation was fitted (results not shown). This might be because of the experimental design, which was not a typical square or rectangular block.

Analyses based on mixed models showed important aspects about the phenotypic variation. Evidences of G×L interaction were observed both on the boxplot dispersion (given the differential behavior across the locations) and the heterogeneous variances (the fully unstructured matrix showed the best fit). Previous results about G×L interaction were reported using ordinary least squares analysis of variance (Ferrão et al. 2007). In accordance with these studies a change in the genotypic ranking was observed (results not shown). The G×H interaction in our results shows a lack of annual yield stability. Although this phenomenon has been commonly reported in *C. arabica*, some studies have shown a similar behavior in *C. canephora* (Cilas et al. 2011). Our results support this. Planned pruning can reduce the annual instability and is commonly used in Brazilian breeding programs. It is part of a series of agronomic recommendations that minimize the variations across the harvests and stabilizes the production.

The phenotypic analysis clearly showed the importance of including interaction terms in the model and their importance to a breeding program. In naive models, all environmental-specific effects (i.e., location and harvest) are assumed to come from the same distribution with the same genetic variance component. However, if genetic effects are conditional on the environment, then the genetic components should be allowed to vary across environments (Malosetti et al. 2016). From a quantitative genetics perspective, it is reasonable to expect that genotypic effects may differ across years and locations because the final state of a trait will be the cumulative result of the number of causal interactions between the genetic make-up of the genotype and the condition in which the plant developed (Malosetti et al. 2014). This agrees with MET modeling. Our study showed that it is important to consider interactions for further GS modeling.

In terms of statistical modeling, the models were compared using different criteria. Cross-validation is the standard method to compare GS models, although it might not be always a sensitive instrument for model comparison (Wang and Gelman 2014; Gelman et al. 2014). Here, we reinforce the relevance of using more than one criterion to draw conclusions. The goodness-of-fit value, commonly used in genetic studies (Kelly et al. 2009; Pastina et al. 2012; Oakey et al. 2016), was considered for this proposal. Hence, when the inclusion of the MET modeling or the pedigree record has been studied, we are essentially quantifying the plausibility of a model that considers this source over others. Although rarely discussed in GS studies, the AIC and BIC criterion were used here. More plausibility (lower AIC and BIC) was observed for methods that considered the MET modeling. This highlights its importance on model formulation.

An improvement in the goodness-of-fit value was observed when the genetic relationship was considered. This result is expected in a general context. It is more plausible to consider the existence of correlation between genotypes rather than homogeneous variances and null genetic correlations (two assumptions when a Identify matrix is assumed). While empirical results reinforce the pedigree importance (Kelly et al. 2009), a significant number of MET studies still assume independence between genotypes (Smith et al. 2001). This number is inflated in coffee because few pedigree mixed models have been reported. As pointed by Piepho et al. (2008), the assumption of independence between the genetic effects results in limited gain if additional information is not considered in the estimation process of breeding values.

The difference in performance between models that considers molecular information (A_m) and pedigree (A_p) is linked with some practical and theoretical aspects. The practical aspect refers to the way in which the pedigree was recorded. Genealogy control is typically hampered in open-pollinated crops. In this study, only seeds that were harvested on the same plant, i.e., half-sib individuals, were considered. In a theoretical context, the A_m and A_p matrices keep different levels of information. While the A_p regards information from alleles to be identical by descent (IBD), the A_m regards information from alleles to be identical by state (IBS). The empirical results in full-sibs, for example, could show a variation from 0.4 to 0.6 in the genomic relationship matrix, which is possible to be captured by the Mendelian sampling term (Mrode 2014). A fixed value of 0.5 is calculated using only the pedigree record considering the expected average genetic covariance between full-sibs. The exploitation of this level of variation usually results in better goodness-of-fit statistics for GBLUP versus traditional BLUP. Both aspects support the observed superiority of the genomic models and concur with our results.

In the GS context, we reinforce the importance to draw conclusions supported in more that one criterion. Both goodness-of-fit value and predictive ability are important comparison parameters. Cross-validation was performed in this sense and the results generally agree with the fit analyses. Models that considered the MET modeling consistently had the highest accuracy values (on the order of 10–17% versus models that ignored the MET modeling). **MET.GBLUP** was generally the best or second best performing method. The main argument in favor of this method is the possibility to recover information via the covariance matrix (Malosetti et al. 2016). It also offers the possibility to use molecular data to describe the genetic similarity and to test different VCOV structures to describe the correlation across locations and harvest. This is reflected in more plausibility and better predictive capacity. Therefore, a more realistic description of this phenomenon could be obtained and combined with good predictions.

Methods that do not consider MET modeling all had poor results. The interactions have been showing to be an important source of variation in many phenotypic studies (Crossa et al. 2006; Smith et al. 2007; Burgueño et al. 2011) as well as in GS studies (Burgueño et al. 2012; Malosetti et al. 2016; Oakey et al. 2016). To evaluate its consequence in the breeding program, these results were examined for selection decisions. The top 10% of genotypes were selected considering the **MET.GBLUP** model and the genetic gain was compared with the **Trad** model. Changes on the ranking and differential response to selection were observed between both methods. An increase of 4% in grain production is expected when genotypes were selected considering the **MET.GBLUP** model comparing to the top 10% genotypes selected using the **Trad** model as criteria. These results are in accordance to the evidences described by Kelly et al. (2009). In a breeding program, the identification of

most performant genotypes for commercial release is an important condition of success. In essence, changes in the genotypic ranking indicate instability on the selection process, and hence a potential loss of selection gain in future generations.

The low number of studies using high throughput genotyping in coffee motivates a brief discussion. The good performance of the GS method highlights the importance of this tool in the *Coffea* genus. To the best of our knowledge, molecular studies have been reported in coffee; however, most of them are still based on traditional molecular approaches (Ferrão et al. 2013; Cubry et al. 2013; Mérot-L'Anthoëne et al. 2014). Large-scale genotyping expands their utility. The GBS approach identified 449,467 SNPs in the unfiltered file. After filtering, the SNP density decreased to 18,586. While this only represents 4% of the raw SNPs, this number is still larger than in recent coffee reports. In addition to the GS application, molecular information may assist in the selection of potential individuals. Self-incompatibility is a genetic mechanisms which prevent self-fertilization and thus encourage outcrossing and allogamy. In *C. canephora* species, this phenomenon hinders parental selection since progenitors should not to be highly related. In this sense, the use of molecular tools to understand the genetic relationship between individuals is an additional benefit that can support the selection decision.

Finally, in our research context, the **MET.GBLUP** model will be considered in the selection of progenies for a new cycle of recurrent selection. For practical implementation in future, we believe that some factors discussed in this research are essentials, including (i) good phenotypic evaluations, considering a proper experimental design and reliable phenotypic measures; (ii) a selection of a suitable MET model to describe the phenotypic variation; (iii) reliable molecular informations; and (iv) a GS model considering all important sources of variation, including the interactions. In addition, increasing the sample size of the training population is another relevant point for future applications. Its importance is clear for two reasons, as pointed out by de los Campos et al. (2013). First, the accuracy of estimated marker effects increases with sample size, because bias and variance of estimates of marker effects decrease with increasing sample size. Second, an increase in sample size may also increase the extent of the genetic relationship between training and testing data sets, which is an important factor to compute the predictive ability. Imputation methods and improved on the bioinformatic steps, especially in the SNP and genotype calling, are important future trends. In terms of statistical modeling, studies focusing on the importance of non-additives effects and the use of alternative approaches, such as hierarchical Bayesian regressions, are important perspectives in coffee research (Ferrão et al. 2016).

Conclusion

In this research, an expansion of the traditional GBLUP approach to address the conjugate use of genomic information and MET modeling was discussed for genomic prediction in the context of coffee breeding. This model was called "GBLUP version to multiple harvest-location trial" (**MET.GBLUP**) and showed the best goodness-of-fit statistics and high predictive ability, compared to other competitor models. Furthermore, promising results in terms of number and SNP density across the genome suggesting that GBS can be used as an efficient genotyping method in coffee research. As a final message, GS approach is recommended as a promising and innovative approach to be applied in coffee. In practice, compared to traditional phenotypic evaluation, it is expected to accelerate the breeding cycle, maintain genetic diversity and increase the genetic gain per unit of time. For this end, this research evidenced that consider a suitable genomic prediction model and understand the breeding scenario that is attempting to address are two important features to be contemplated for future implementation.

Acknowledgments This work is partially supported by FAPESP/CAPES (São Paulo Research Foundation), grants 2014/20389-2 for L.F.V.F and A.A.F.G. Phenotypic evaluations and GBS data is supported by Fapes (Espírito Santo Research Foundation), grants 55207464/11 and 65192036/14. Additional support is provided by the Instituto Capixaba de Pesquisa, Assitência Tecnica e Extensão Rural (Incaper) and Embrapa Cafe. A.A.F.G, R.G.F, M.A.G.F and A.F have a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The author thank Livia Souza and Anete P. de Souza (CBMEG, Unicamp/Brazil) by the assistance in the DNA extraction step; and Paulo Volpi (Incaper/Brazil) by the support on the phenotypic evaluation.

Author Contributions L.F.V.F, A.A.F.G, R.G.F, M.A.G.F and A.F conceived the study and designed the experiments. R.G.F, M.A.G.F and A.F installed the experimental design and collected the phenotypic data. L.F.V.F performed the DNA extraction. L.F.V.F and A.A.F.G performed the genomic prediction analysis and proposed the theoretical idea of the model. L.F.V.F wrote the paper.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Data Archiving Statement The genotypes in the study belong to the germplasm collection and breeding program of the Incaper institution (ES,Brazil). Phenotypic and genotypic data were submitted as Supplementary Material.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Bates D, Vazquez AI (2014) pedigreemm: Pedigree-based mixed-effects models. https://CRAN.R-project.org/package=pedigreemm, r package version 0.3-3

Burgueño J, Crossa J, Cotes JM, Vicente FS, Das B (2011) Prediction assessment of linear mixed models for multienvironment trials. Crop Sci 51(3):944–954

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52(2):707

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1):375–385

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2):327–45

Cilas C, Montagnon C, Bar-Hen A (2011) Yield stability in clones of coffea canephora in the short and medium term: longitudinal data analyses and measures of stability over time. Tree Genet Genomes 7(2):421–429

Crossa J, Burgueño J, Cornelius PL, McLaren G, Trethowan R, Krishnamachari A (2006) Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci 46(4):1722–1733

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E et al (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3: Genes–Genomes–Genetics 3(11):1903–1926

Cubry P, De Bellis F, Avia K, Bouchet S, Pot D, Dufour M, Legnate H, Leroy T (2013) An initial assessment of linkage disequilibrium (ld) in coffee trees: Ld patterns in groups of coffea canephora pierre using microsatellite analysis. BMC Genomics 14 (1):10

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al (2011) The variant call format and vcftools. Bioinformatics 27(15):2156–2158

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345(6201):1181–1184

Elshire RJ, Glaubitz JC, Sun Q, Poland Ja, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One 6(5):e19,379

Falconer DS, Mackay TFC (1996) Quantitative genetics. Pearson Education Limited, England

Ferrão LFV, Caixeta ET, Souza FdF, Zambolim EM, Cruz CD, Zambolim L, Sakiyama NS (2013) Comparative study of different molecular markers for classifying and establishing genetic relationships in coffea canephora. Plant Syst Evol 299(1):225–238

Ferrão LFV, Caixeta ET, Pena G, Zambolim EM, Cruz CD, Zambolim L, Ferrão MAG, Sakiyama NS (2015) New EST–SSR markers of Coffea arabica: transferability and application to studies of molecular characterization and genetic mapping. Mol Breed 35(1):1–5

Ferrão LFV, Ferrão RG, Ferrão MAG, Fonseca A, Stephens M, Garcia AAF (2016) Genomic prediction in Coffea canephora using Bayesian polygenic modeling. In: 5th international conference on quantitative genetics. WI, Madison, p 203

Ferrão RG, Ferrão MAG, Fonseca A, Pacova B (2007) Melhoramento genético de Coffea canephora. In: Ferrão R, Fonseca A, Bragança S, Ferrão M, Muner LD (eds) Cafe conilon, incaper edn Vitória-ES, pp 123–173

Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Chapman & hall/CRC Boca Raton, FL, USA

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) Tassel-gbs: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2):e90,346

GNU P (2007) Free Software Foundation. Bash (3.2.48) [Unix shell program].Retrieved from http://ftp.gnu.org/gnu/bash/bash-3.2.48.tar.gz

Grattapaglia D, Resende MDV (2010) Genomic selection in forest tree breeding. Tree Genet Genomes 7(2):241–255

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127(2):463–480

Hu Y, Yan C, Hsu CH, Chen QR, Niu K, Komatsoulis GA, Meerzaman D (2014) Omiccircos: a simple-to-use r package for the circular visualization of multidimensional omics data. Cancer Informat 13:13

IOC (2016) International Coffee Organization - Trade Statistics Tables. http://www.ico.org/

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9(2):166–177

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M et al (2014a) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127(3):595–607

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014b) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15(1):740

Kelly AM, Cullis BR, Gilmour AR, Eccleston JA, Thompson R (2009) Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet Sel Evol 41(1):1

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink JL, Singh RP, Autrique E, de los Campos G (2015) Increased prediction accuracy in wheat breeding trials using a marker x environment interaction genomic selection model. G3: Genes–Genomes–Genetics 5(4):569–82

Malosetti M, Ribaut JM, van Eeuwijk FA (2014) The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Drought phenotyping in crops: From theory to practice 4(44):53

Malosetti M, Bustos-Korts D, Boer MP, van Eeuwijk FA (2016) Predicting responses in multiple environments: issues in relation to genotype× environment interactions. Crop Sci 56(5):2210–2222

Margarido GRA, Pastina MM, Souza AP, Garcia AAF (2015) Multitrait multi-environment quantitative trait loci mapping for a sugarcane commercial cross provides insights on the inheritance of important traits. Mol Breed 35(8):175

Mérot-L'Anthoën V, Mangin B, Lefebvre-Pautigny F, Jasson S, Rigoreau M, Husson J, Lambot C, Crouzillat D (2014) Comparison of three qtl detection models on biochemical, sensory, and yield characters in coffea canephora. Tree Genet Genomes 10(6):1541–1553

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Meyer K (2009) Factor-analytic models for genotype × environment type problems and structured covariance matrices. Genet Sel Evol 41(1):21

Mrode RA (2014) Linear models for the prediction of animal breeding values. Cabi

Oakey H, Cullis B, Thompson R, Comadran J, Halpin C, Waugh R (2016) Genomic selection in multi-environment crop trials. G3: Genes–Genomes–Genetics 6(5):1313–1326

Pastina MM, Malosetti M, Gazaffi R, Mollinari M, Margarido GRA, Oliveira KM, Pinto LR, Souza AP, van Eeuwijk FA, Garcia AAF (2012) A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. Theor Appl Genet 124:835–849

Payne RW, Murray DA, Harding SA (2011) An introduction to the genstat command language (14th edn)

Piepho H, Möhring J, Melchinger A, Büchse A (2008) Blup for phenotypic selection in plant breeding and variety testing. Euphytica 161(1-2):209–228

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome J 5(3):103

R Core Team (2013) R: A Language and Environment for Statistical Computing

Schulz-Streeck T, Ogutu JO, Piepho HP (2013) Comparisons of single-stage and two-stage approaches to genomic selection. Theor Appl Genet 126(1):69–82

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Smith AB, Cullis BR, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57(4):1138–1147

Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agric Sci 143(06):449

Smith AB, Stringer JK, Wei X, Cullis BR (2007) Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. Euphytica 157(1-2):253–266

Smith KF, Casler M (2004) Spatial analysis of forage grass trials across locations, years, and harvests. Crop Sci 44(1):56–62

Spindel JE, Begum H, Akdemir D, Collard B, Redona E, Jannink Jl, Mccouch S (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395–408

Tran HTM, Lee LS, Furtado A, Smyth H, Henry RJ (2016) Advances in genomics for the improvement of quality in coffee. J Sci Food Agric 96(10):3300–3312

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Wang W, Gelman A (2014) Difficulty of selecting among multilevel models using predictive accuracy. Statistics at its Interface 7(1):1–88