



# Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models

Luís Felipe Ventorim Ferrão<sup>1</sup> · Romário Gava Ferrão<sup>2</sup> · Maria Amélia Gava Ferrão<sup>2,3</sup> · Aymbiré Fonseca<sup>2,3</sup> · Peter Carbonetto<sup>4,5</sup> · Matthew Stephens<sup>4,6</sup> · Antonio Augusto Franco Garcia<sup>1</sup>

Received: 15 January 2018 / Revised: 23 May 2018 / Accepted: 30 May 2018 / Published online: 25 June 2018  
© The Genetics Society 2018

## Abstract

Genomic selection has been proposed as the standard method to predict breeding values in animal and plant breeding. Although some crops have benefited from this methodology, studies in *Coffea* are still emerging. To date, there have been no studies describing how well genomic prediction models work across populations and environments for different complex traits in coffee. Considering that predictive models are based on biological and statistical assumptions, it is expected that their performance vary depending on how well these assumptions align with the true genetic architecture of the phenotype. To investigate this, we used data from two recurrent selection populations of *Coffea canephora*, evaluated in two locations, and single nucleotide polymorphisms identified by Genotyping-by-Sequencing. In particular, we evaluated the performance of 13 statistical approaches to predict three important traits in the coffee—production of coffee beans, leaf rust incidence and yield of green beans. Analyses were performed for predictions within-environment, across locations and across populations to assess the reliability of genomic selection. Overall, differences in the prediction accuracy of the competing models were small, although the Bayesian methods showed a modest improvement over other methods, at the cost of more computation time. As expected, predictive accuracy for within-environment analysis, on average, were higher than predictions across locations and across populations. Our results support the potential of genomic selection to reshape traditional plant breeding schemes. In practice, we expect to increase the genetic gain per unit of time by reducing the length cycle of recurrent selection in coffee.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1038/s41437-018-0105-y>) contains supplementary material, which is available to authorized users.

---

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

<sup>1</sup> Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, SP 13400-970, Brazil

<sup>2</sup> Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural - Incaper, Vitória, ES 29052-010, Brazil

<sup>3</sup> Empresa Brasileira de Pesquisa Agropecuária—Embrapa Café, Brasília, Brazil

<sup>4</sup> Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

<sup>5</sup> Research Computing Center, University of Chicago, Chicago, IL 60637, USA

<sup>6</sup> Department of Statistics, University of Chicago, Chicago, IL 60637, USA

## Introduction

Plant and animal breeders have used quantitative genetics effectively to increase mean phenotypic performance in selected populations. Traditionally, genetic progress has been achieved by combining phenotypic evaluations with the pedigree record, which involves visual evaluation and trait screening over several successive generations (Goddard and Hayes 2007). These approaches have brought significant gains in recent decades. However, it is important to take into account the effort required to achieve these gains; for the majority of perennial crops, this approach is costly and time-consuming, particularly for traits expressed late in a plant's life-cycle.

The advent of molecular markers has provided an opportunity to achieve faster genetic gains (Lande and Thompson 1990). Meuwissen et al. (2001) first proposed to use all available molecular markers to predict quantitative traits in breeding programs. Known as genomic selection (GS), the methodology has become widely adopted in the

animal and plant breeding communities because of its potential to increase genetic gains and shorten the breeding cycle. The rationale behind this approach is that, whenever marker density is high enough, most quantitative trait loci (QTLs) will be in linkage disequilibrium (LD) with some markers and hence estimates of marker effects will lead to accurate predictions of genetic merit for a complex trait (Goddard and Hayes 2007).

When confronting the problem of modeling the relationship between genetic variation and variation in the observed traits, an important question is what statistical method might better describe this relationship. Several analytical approaches have been proposed for genome-based prediction of genetic values, such as penalized and Bayesian estimation procedures, as well as nonparametric regression and dimension reduction methods (Gianola et al. 2009; Kärkkäinen and Sillanpää 2012; Gianola 2013; de Los Campos et al. 2013). A common feature of all these methods is that they were designed to handle highly-dimensional data, with a particular focus on producing accurate estimates in settings in which the number of variables, or SNPs ( $p$ ), is larger than the number of samples ( $n$ ). Most successful approaches are based on variable selection and/or shrinkage techniques from the statistics literature (Kärkkäinen and Sillanpää 2012; Zhou et al. 2013; Garrick et al. 2014).

Comparisons between genomic prediction models have been carried out in a variety of scenarios for different species and traits (Heslot et al. 2012; Riedelsheimer et al. 2012; Daetwyler et al. 2013; Wang et al. 2015). Empirical and simulation studies have suggested that different models work better in different scenarios, since biological and technical factors affect prediction accuracy. These factors include population size, genetic architecture and differences between the training and validation data sets (de Los Campos et al. 2013; Daetwyler et al. 2013). Because of this, when considering analyses of new species and breeding scenarios, it can be helpful to compare and assess several methods before carrying out the final genomic analyses. Here we perform such an assessment for genomic prediction in coffee, an important agricultural commodity in which genomic studies are still emerging.

So far, genomic prediction accuracy has usually been evaluated within single environments (Windhausen et al. 2012a, b; Beaulieu et al. 2014; Gamal El-Dien et al. 2015). In coffee, however, breeding schemes are most commonly performed in multiple environments to measure performance of genotypes across a range of conditions. In this study, therefore, we focus on the following question: *are marker effects estimated in one set of environments useful for prediction in other environments?* This question has important practical implications for the effectiveness of GS in perennial species such as coffee. If feasible, for example, a single prediction model could be used across different

environments resulting in time and cost economy (Resende et al. 2012a).

This investigation builds on our earlier work that explored the potential of GS for production of coffee beans (Ferrão et al. 2017). In that work, we used a mixed model framework with integration of spatial and temporal variance-covariance structures. In combination, this investigation explores the potential of several statistical methods to predict three important traits—production of coffee beans, leaf rust incidence, and yield of green beans—evaluated in two recurrent selection populations of *Coffea canephora*. Our results demonstrate the usefulness of genome-base prediction for coffee breeding. We also provide guidance on implementing molecular assisted selection in practical breeding programs.

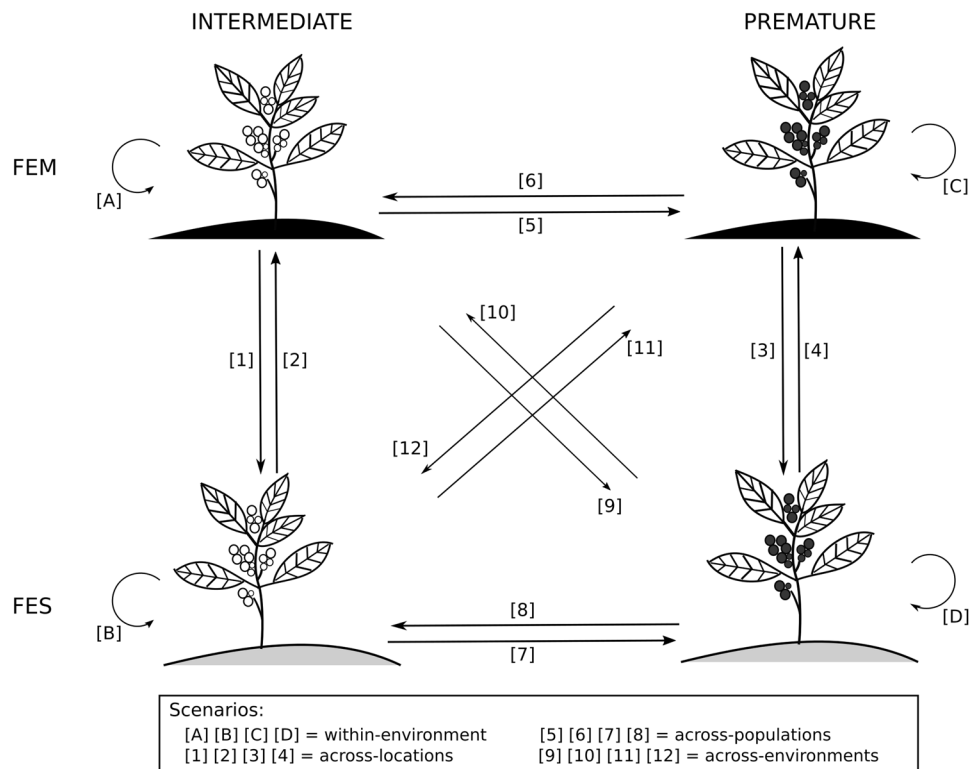
## Materials and methods

The description of the Materials and methods is organized as follows. In the Sections “Plant material” and “Experimental design”, we describe the development of the population used in the experiments, collection of the phenotype data, and breeding scenarios proposed for GS investigation. We take additional steps, described in the Section “Phenotypic model”, to prepare the phenotype data for analysis with whole-genome prediction models. At this stage of the analysis, no marker information was included in the model. The protocols for DNA sequencing and calling SNP genotypes are described in the Section “Genotypic data”. Finally, in the Sections “Genomic prediction methods” and “Evaluation of genomic predictions” we describe the genomic prediction methods used in our experiments and explain how these methods were compared.

### Plant material

We consider an experimental population designed by the *Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper)*, ES State, Brazil. Phenotype measurements were collected from two recurrent selection populations formed from the recombination of superior *C. canephora* clones. Clones were selected by the *Incaper* as progenitors (founders) based on high production of coffee beans and similar stages of fruit maturity. The latter is an important feature of new coffee varieties because it allows harvests to be standardized. Based on the maturity group, coffee populations were designated as “Intermediate” or “Premature”. The Intermediate population, on average, had fruits that started to ripen in March and April, and then were harvested in June. The fruits of the Premature population started ripening, and were harvested one month earlier, on average.

**Fig. 1** Genomic selection experimental scenarios. Here, “environment” is defined as a combination of location (FEM, FES) and population (Intermediate, Premature). Scenarios **a–d** assess genomic selection within the same environment; Scenarios 1–4 compare GS performance across-locations; Scenarios 5–8 evaluate GS performance across populations; and Scenarios 9–12 assess both across-location and cross-population predictions. Direction of the arrows indicate differences in the training and test data sets



In 1997, the original Intermediate and Premature populations were derived from crosses of 16 and 9 progenitors, respectively. Each population was planted in an isolated field under open pollination conditions. In 2000, after one cycle of recombination, seeds were derived from each maternal plant, which were then used to develop a new population. At this point, the same number of seeds per maternal plant were harvested in order to preserve genetic diversity. After four consecutive harvest-production years (2002–2005), 103 progenies from the Intermediate population and 87 progenies from the Premature population were selected based on their high performance in terms of production of coffee beans and tolerance to biotic and abiotic stress over these four years. In 2006, founders and the selected progenies belonging to both populations were cloned and assigned to randomized complete blocks with three replications and five plants per plot. The average measurement per plot was used as the phenotype for all subsequent analyses in this study.

Both populations were established in two locations, chosen to be representative of Brazilian coffee production: Marilândia Experimental Farm, or FEM (latitude 19°24' south, longitude 40°31' west, 70 m altitude); and Sooretama Experimental Farm, or FES (latitude 15°47' south, longitude 43°18' west, 40 m altitude).

The complete experiment used phenotype measurements from 3570 coffee trees in the Intermediate population, and from 2880 coffee trees in the Premature population.

Measures were recorded over four consecutive harvest-production years (2008–2011) for three traits: production of coffee beans (mature coffee fruit in the “cherries” stage, in 60-kg bags per hectare); natural infection of coffee leaf rust, caused by the *Hemileia vastatrix* fungus (levels ranging from 1 to 9, according to visual sporulation intensity evaluated in field); and yield of green beans post-harvest trait (ripened beans, in g, after processing by dry methods to remove dried husks in samples of 2 kg of coffee fruit in the cherries stage).

## Experimental design

To investigate potential for GS, we considered two aspects of plant breeding: (i) prediction accuracy for different traits, and (ii) prediction accuracy within and across environments. Here we define “environment” as a specific combination of location and population. Figure 1 summarizes our experiments.

For within-environment experiments (Scenarios A–D), predictions were evaluated using a Replicated Training-Testing evaluation (Crossa et al. 2013). In each replication, 80% of the individuals were assigned randomly to the training set (TRN), while the remaining 20% were assigned to the test set (TST). This division was repeated 30 times with different random assignments to TRN and TST. Models were fitted to the training data and prediction accuracy was evaluated in the test data.

For across-environment experiments (Scenarios 1–12), we subdivide the experiments as follows (see Fig. 1): (i) Scenarios 1–4 capture across-location predictions, in which the test set contains samples collected from a different location than the training set, while the source population is kept the same; (ii) Scenarios 5–8 consider across-population predictions, in which the training and test sets contain samples from different populations, while the location is kept the same; (iii) Scenarios 9–12 capture both across-location and across-population predictions, in which the test set has samples from a different population and a different location than the training set. In these experiments we did not use the Replicated Training-Testing design; for example, in Scenario 1 the model was trained with all Intermediate individuals from one location (e.g., FEM) and validated using all Intermediate individuals from another location (e.g., FES).

## Phenotypic model

The phenotypes were adjusted for linear effects of environmental covariates, and other experimental covariates. In particular, in our experiments we collected longitudinal data across multiple harvest-years. Different variance-covariance structures were tested to describe this temporal variation across harvest, and therefore improve the estimation of genetic effects. Using a similar notation to Pastina et al. (2012), we considered the following statistical model (underlined terms indicate random variables):

$$y_{ijk} = \mu + B_j + H_k + \underline{G}_{ij} + \varepsilon_{ijk} \quad (1)$$

where  $y_{ijk}$  is the phenotype measured in individual  $i \in \{1, \dots, n\}$  from block  $j \in \{1, \dots, r\}$  and harvest  $k \in \{1, \dots, K\}$ ;  $\mu$  is the intercept;  $H_k$  is the fixed harvest effect;  $B_j$  is the fixed block effect;  $\underline{G}_{ij}$  is a random genetic effect of genotype  $i$  at harvest  $k$ ; and  $\varepsilon_{ijk}$  is a random non-genetic residual error term. Here,  $r = 3$  (the number of blocks) and  $K = 4$  (the number of harvests).

To model the random genetic effects, we assumed a multivariate normal distribution with a zero mean and a variance-covariance matrix  $\mathbf{G}$ . We formulated  $\mathbf{G}$  as the Kronecker product  $\mathbf{G} = \sum_H^{K \times K} \otimes I_g^{n \times n}$ , in which  $I_g^{n \times n}$  is the  $n \times n$  identity matrix. Four structures  $\sum_H^{K \times K}$  different levels of complexity (i.e., number of model parameters to be estimated) were investigated (see Supplementary Table S1).

Similarly, for the residual error, we assumed a multivariate normal distribution with a zero mean and variance-covariance matrix  $\mathbf{R}$  defined as  $\mathbf{R} = R_H^{K \times K} \otimes I_B^{r \times r} \otimes I_g^{n \times n}$ . The term  $I_B^{r \times r}$  is an Identity of dimension equal to the number of blocks,  $r$ . For the term  $R_H^{K \times K}$ , the “Ident” and “Diag” variance-covariance structures were considered (Supplementary Table S1). Our previous study showed no improvements in the goodness-of-fit values when spatial

trends were evaluated (see Ferrão et al. 2017); therefore, we did not consider spatial analysis here.

The final model choices were based on AIC (Akaike Information Criterion) (Akaike 1974) and BIC (Bayesian Information Criterion) (Schwarz 1978). Since calculation of heritability in complex linear mixed models is not straightforward (Cullis et al. 2006; Oakey et al. 2016), broad-sense heritability ( $h^2$ ) was estimated from the simplest phenotypic model—that is, identity structure for the genetic and residual matrices—as:  $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2/rK}$ ; where  $\sigma_g^2$  is the estimated variance of the genotype component,  $\sigma_e^2$  is the estimated variance of the residual component, and  $r$  and  $K$  are the number of blocks and harvests, respectively.

All analyses described in this section were performed using the R package nlme (Pinheiro et al. 2013).

## Genotypic data

The Intermediate and Premature populations were genotyped using the Genotype-by-Sequencing (GBS) approach that was first developed by Elshire et al. (2011). We followed the GBS protocol used by the Genomic Diversity Facility at Cornell University.

Leaves were collected and lyophilized. DNA was extracted using Qiagen DNeasy Plant, and the genomic libraries were prepared following Elshire et al. (2011). The DNA samples were digested using the *ApeKI* restriction enzyme, and 96 samples were multiplexed per Illumina flow cell for sequencing.

The GBS analysis pipeline was implemented with the TASSEL-GBS software, version 4.3.7 (Glaubitz et al. 2014). Sequenced tags were aligned against the *C. canephora* genome assembly (Denoeud et al. 2014). SNPs were extracted from the raw Variant Call Format (VCF file) and filtered manually as follows: (i) triallelic SNPs were removed; (ii) SNPs with minor allele frequency (MAF) less than 1% were removed; and (iii) SNPs with genotypes that were called in less than 70% of the samples were discarded.

To ensure that all genotypes were called consistently, we used a Bayesian approach that incorporates genetic background information, similar to Chan et al. (2016), to call genotypes with low coverage (which we defined as genotypes with less than or equal to five sequenced reads). Specifically, we used a two-step approach: first, the maximum-likelihood estimates of genotypes were computed following Chan et al. (2016) (assuming a uniform genotype prior); second, the inferred parental genotypes were provided as prior information for inferring the genotypes of the progenies. To improve accuracy of the parental genotype estimates, we increased the sequencing coverage of the parents to 3× the coverage of the progenies. To call the genotypes, we retained the maximum-probability genotype, encoded as reference allele counts (0, 1 or 2) in our files.

All SNP manipulation and genotype calling (aside from genotypes of low-coverage samples) were carried out using VCFtools (Danecek et al. 2011). We used R (R Core Team 2013) to implement the Bayesian genotype calling incorporating parental genotype estimates. SNP density plots were created using the synbreed R package (Wimmer et al. 2012).

## Genomic prediction methods

We compared 13 different methods for genomic prediction of coffee traits. Most of the genomic prediction approaches included in our experiments are based on a linear regression in which the outcome of interest  $y$  is modeled as a linear combination of the SNP markers:

$$y = \mu 1_n + X\beta + \varepsilon \quad (2)$$

Here,  $y$  is an  $n$ -vector of phenotypes measured on  $n$  individuals, after adjusting for linear effects of environmental factors and other experimental factors, as explained in Sec. 2.3;  $X$  is an  $n \times p$  matrix of genotypes measured at  $p$  SNPs;  $\beta$  is a  $p$ -vector of SNP effects to be estimated;  $1_n$  is an  $n$ -vector of 1's;  $\mu$  is the intercept, and  $\varepsilon$  is an  $n$ -vector of normally distributed residuals,  $\varepsilon \sim N(0, \sigma_e^2 I_{n \times n})$ .

For comparing genomic prediction approaches, we defined three classes of methods: (1) fixed multiple regression, (2) Bayesian methods, and (3) a third class of methods based on techniques originally developed in machine learning (which don't already fit into the first two categories).

### Fixed multiple regression

This class of method builds on standard statistical association analysis approaches used in genome-wide association studies (GWAS) which test each SNP, one at a time, for association with the phenotype (we refer to this as "single-SNP" analyses). We implement a fixed regression procedure following Spindel et al. (2015), using a subset of markers identified from a single-SNP analysis. For each replication in the cross-validation scheme (Replicated Training-Testing evaluation), single-marker regression was applied to all SNPs, and association  $p$ -values were computed using an  $F$ -test. A linear regression model was then fitted to the data using the 100 most significant markers. We refer to this method as "fixedMRL".

### Machine learning methods

We consider three approaches from the machine learning literature: regularized regression, dimension reduction, and random forests.

1. **Regularized regression:** This method fits a regression model with all  $p$  SNPs, shrinking all coefficients toward zero. Regression coefficients are fitted by solving an optimization problem that balances goodness-of-fit against model complexity (de Los Campos et al. 2013; James et al. 2013). Several regularized approaches have been proposed, and they differ in the choice of penalty function. Ridge regression (RR) and LASSO are the two most prominent approaches. RR shrinks all coefficients toward zero, with a penalty applied to the  $\ell_2$ -norm of the coefficients. In contrast, LASSO uses the  $\ell_1$ -norm (James et al. 2013). The RR-BLUP is a version of RR that implements best linear unbiased prediction (BLUP) using a mixed model approach (Endelman 2011). The RR-BLUP and LASSO approaches were implemented using, respectively, the rrBLUP (Endelman 2011) and glmnet (Friedman et al. 2010) R packages. The LASSO penalty strength was chosen via cross-validation, following Silva et al. (2011).
2. **Partial least squares regression (PLSR):** This is a dimension-reduction approach that transforms the variables (SNPs), then fits a model with the transformed variables. PLSR is similar to principal component regression (PCR); both methods construct a matrix of latent components as a linear transformation (James et al. 2013). PLSR was implemented using the pls R package (Wehrens and Mevik 2007) with the default settings.
3. **Random forest (RF):** A random forest is a collection of regression trees, in which a subset of SNPs is used to define the best split at each node (James et al. 2013). Different variables are used at each split in different trees. The RF prediction for an observation is obtained by averaging the predictions over trees. One feature of the RF approach is that it allows for non-linear relationships between genotype and phenotype. RF was implemented in our study using the RandomForest R package (Liaw and Wiener 2002) with the default settings.

### Bayesian methods

All Bayesian approaches are based on a hierarchical linear regression method, building on (2), and differ primarily in the priors placed on the regression coefficients and other model parameters (Gianola 2013). Using notation similar to de Los Campos et al. (2013), the posterior distribution of the model parameters  $\mu, \beta, \sigma^2$  given the hyperparameters  $\omega$



**Table 1** Summary of effect size distributions used in selected genomic prediction methods, adapted from Zhou et al. (2013)

Name	$p(\beta \omega)^{**}$	Keyword	Software
t	$\beta_j \sim t(0, v, \sigma_a^2)$	bayesA <sup>a</sup>	BGLR <sup>h</sup>
Point-t	$\beta_j \sim \pi t(0, v, \sigma_a^2) + (1 - \pi)\delta_0$	bayesB <sup>a</sup>	BGLR <sup>h</sup>
Point-normal	$\beta_j \sim \pi N(0, \sigma_a^2) + (1 - \pi)\delta_0$	bayesC <sup>b</sup> , bayesVS <sup>c</sup>	BGLR <sup>h</sup> , varbvs <sup>c</sup>
Point-normal-mixture	$\beta_j \sim \pi_1 N(0, \sigma_a^2) + \pi_2 N(0, 0.1\sigma_a^2) + \pi_3 N(0, 0.01\sigma_a^2) + (1 - \pi_1 - \pi_3 - \pi_3)\delta_0$	bayesR <sup>d</sup>	BayesR <sup>d</sup>
Normal	$\beta_j \sim \pi N(0, \sigma_a^2)$	bayesRR <sup>e</sup>	BGLR <sup>h</sup>
Normal-mixture	$\beta_j \sim \pi N(0, \sigma_a^2 + \sigma_b^2) + (1 - \pi)N(0, \sigma_b^2)$	bslmm <sup>f</sup>	GEMMA <sup>f</sup>
Double-exponential	$\beta_j \sim DE(0, \theta)$	bayesLASSO <sup>g</sup>	BGLR <sup>h</sup>

<sup>a</sup>Meuwissen et al. (2001)<sup>b</sup>Habier et al. (2011)<sup>c</sup>Carbonetto et al. (2017)<sup>d</sup>Erbe et al. (2012)<sup>e</sup>Whittaker et al. (2000)<sup>f</sup>Zhou et al. (2013)<sup>g</sup>Pérez and de los Campos (2013)<sup>h</sup>Park and Casella 2008

\*\*Notation used in Zhou et al. 2013. DE = “double-exponential” distribution. For the scaled *t*-distribution, *v* and  $\sigma_a^2$  are the number of degrees-of-freedom and scale parameter, respectively. In the DE distribution,  $\theta$  is the scale parameter.  $\delta_0$  denotes a point mass at zero. *Notes*: Some applications of these methods combine a particular effect size distribution with a random effects term, with covariance matrix *K*, to capture sample covariance structure (“cryptic relatedness”); if  $K \propto XX^T$ , it can be shown that this is equivalent to assuming a normal distribution for the effect sizes; this is one motivation for the effect size distributions used in the many of the methods summarized in the above table. In some papers, the “Keyword” column may refer to fitting techniques rather than the assumed effect size distributions. More details in Zhou et al. 2013

is expressed as:

$$p(\mu, \beta, \sigma^2 | y, \omega) \propto p(y | \mu, \beta, \sigma^2) p(\mu, \beta, \sigma^2 | \omega) \quad (3)$$

where  $p(\mu, \beta, \sigma^2 | y, \omega)$  is the posterior density of model parameters  $\mu, \beta, \sigma^2$  given the data (*y*) and the hyperparameters  $\omega$ ,  $p(y | \mu, \beta, \sigma^2)$  is the regression likelihood based on (2), and  $p(\mu, \beta, \sigma^2 | \omega)$  is the prior density of model parameters. Table 1 summarizes the Bayesian methods evaluated in our experiments.

For all Bayesian methods except **bayesVS**, we ran the Markov chain for 20,000 time steps, with a burn-in of 2000. The bayesVS method uses a variational approximation instead of Markov chain Monte Carlo (MCMC) (Carbonetto and Stephens 2012). The **bayesA**, **bayesB**, **bayesC**, **bayesRR** and **bayesLASSO** models are implemented in the **BGLR** package (Pérez and de los Campos 2013); the **bayesR** method is implemented in the **BayesR** package (Erbe et al. 2012); and the **BSLMM** (Bayesian Sparse Linear Mixed Model) method is implemented as part of the **GEMMA** software (Zhou et al. 2013). For all methods, we adopted the default hyperparameter and prior settings.

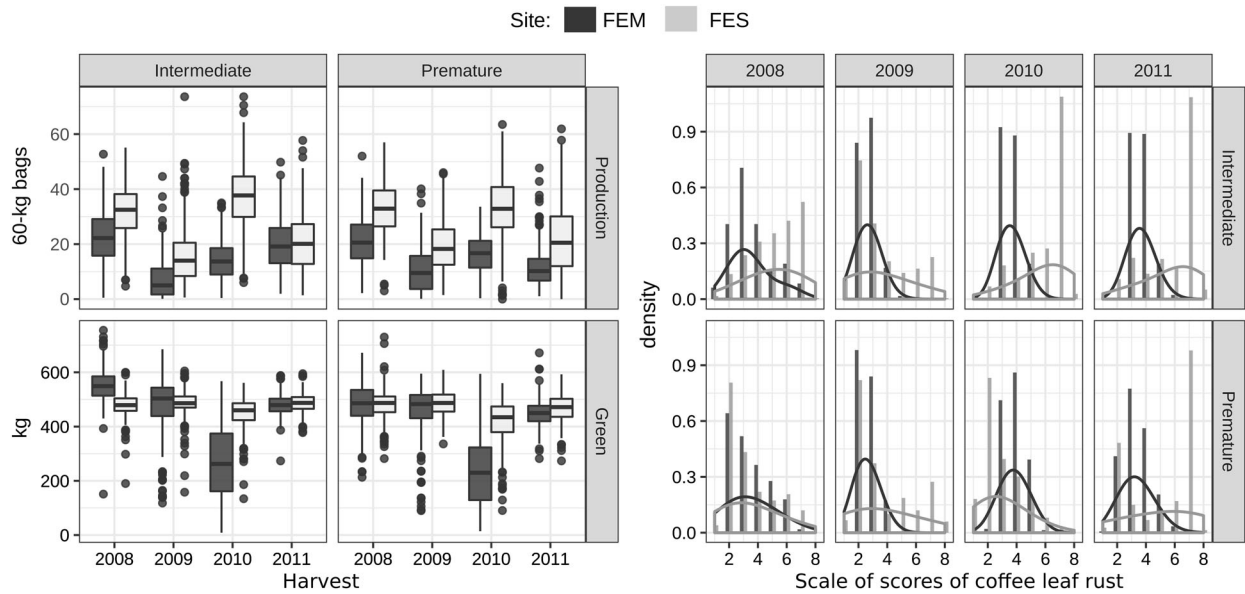
## Evaluation of genomic predictions

We applied each of the 13 methods to predict within-environment phenotypes (Scenarios A–D in Fig. 1). The best performing method was used in the others scenarios (Scenarios 1–12 in Fig. 1). To compare the models, we primarily focused on quantifying the prediction ability, commonly used in the GS literature as a measure of the prediction accuracy. To this end, we compute the Pearson correlation ( $r_{gp}$ ) between predicted ( $\hat{y}_i$ ) and adjusted phenotypes ( $y_i$ ) obtained with Eq. (1).

Following Asoro et al. (2011), we used analysis of variance (ANOVA) to investigate how different factors might be responsible for differences in accuracy among methods. We used the following model in the ANOVA:

$$r_{gp} = \mu + \text{method} + \text{trait} + \text{pop} + \text{loc} + (\text{method} \times \text{trait}) + (\text{method} \times \text{pop}) + (\text{method} \times \text{loc}) + (\text{trait} \times \text{loc}) + \text{error} \quad (4)$$

where  $\mu$  is the intercept; the levels of method are the prediction methods; the levels of trait are the three traits (production of coffee beans, coffee leaf rust, and yield of



**Fig. 2** Left-hand panel: summary of production of coffee beans (in 60-kg bags of mature coffee fruit in the cherries stage per hectare) and yield of green beans (in g, of mature beans after processing using dry methods to remove the entire dried husk in samples of 2 kg of coffee fruit in the cherries stage) evaluated in two locations (FEM, FES), four

harvests (2008–2011) and two populations (Intermediate, Premature). Right-hand panel: Summary of coffee leaf rust scores (*Hemileia vastatrix*), ranging from 1 to 9, according to sporulation observation. Curves are kernel density estimates, which are smoothed version of the histogram

green beans); the levels of pop are the two populations (Intermediate and Premature), and the levels loc are the two locations (FEM and FES) considered. Other terms in (4) correspond to double interactions among factors.

Alternatively, we also estimate the mean squared prediction error (MSPE), slope and computational time to compare the 13 models. MSPE was computed using the formula:

$MSPE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , where  $n$  is the number of samples in the test set. To compute the slope, adjusted phenotypes were linearly regressed on predicted phenotypes to express the degree of bias of the predictions, as suggested by Moser et al. (2009). Runtimes for model fitting were recorded in minutes for all methods and data sets. All computations were single threaded and performed on an Intel Core i7-3770 processor (3.40 GHz) with 8 GB of memory.

Since the degree of genetic relationships between training and test sets can impact accuracy of the predictions, the relationships of both populations were investigated using principal components analysis (PCA) and *Fst*.

In our experiments, we also investigated the effect of number of included SNPs on the predictive ability. To this end, we considered two approaches to selecting SNPs: (i) guided subsets, and (ii) random subsets. To construct the “guided” SNP subsets, we selected 10 SNPs within windows of the same length (in base-pairs) in each *C. canephora* chromosome. To construct SNP subsets of different sizes, we considered different window sizes, ranging from 5 to 900 Kb, by increments of 100 Kb. Following Spindel

et al. (2015), we selected the SNPs with highest minor allele frequencies (MAF) and best call rates within each window. This resulted in SNP subsets with the following numbers of SNPs: 35,427 (smallest windows), 20,450, 13,690, 10,189, 7989, 6577, 5559, 4780, and 4240 (largest windows). For the Premature population the number of SNPs in each subset was 40,767, 21,433, 13,969, 10,283, 8019, 6587, 5560, 4780, and 4240. To construct random SNP subsets, we used exactly the same number of SNPs as in the “guided” subsets; however, SNPs were randomly sampled in each window instead of selecting them based on MAF and call rate.

## Results

### Phenotypic data

Figure 2 summarizes the phenotypic variation in both populations and at both locations. On average, FES location was more productive than FEM location, and showed higher incidence of rust. We observed a lack of annual production stability in coffee bean production over different years in both populations. This instability was quantified in the mixed model analysis, with better goodness-of-fit values (lower AIC and BIC) when heterogeneity of residual and genetic variance were taken into account (Supplementary Table S3). Further, the boxplots in Fig. 2 highlight this cyclical production, interleaving years of high (2008, 2010) and low bean production (2009, 2011).

**Table 2** Broad-sense heritability of production (weight of mature coffee fruit in the cherries stage, in kg), incidence coffee leaf rust (1–9 scale) and yield of green beans (weight, in g, in samples of 2 kilograms of coffee fruit in the cherries stage) across two locations (FEM and FES) and two recurrent selection populations of *Coffea canephora* (Intermediate and Premature)

Trait	Intermediate		Premature	
	FEM	FES	FEM	FES
Production	0.70	0.81	0.74	0.85
Rust	0.61	0.86	0.56	0.89
Green	0.52	0.86	0.72	0.92

Heritability estimates of the three traits, in different environments, ranged from 0.56 to 0.92 (Table 2). Incidence of rust and yield of green beans showed the highest heritability values (0.89 and 0.92, respectively). On average, traits evaluated in the FES location and in the Premature population showed higher heritabilities than the FEM location and Intermediate population.

### Genotyping-by-Sequencing in *C. canephora*

After following the quality-control steps (see 'Materials and methods'), a total of 45,748 (on average, 64.4 SNPs per Mb) and 59,332 (on average, 83.5 SNPs per Mb) molecular markers (SNPs) were retained in the Intermediate and Premature populations, respectively. Among these, 38,106 SNPs (on average, 53.7 SNPs per Mb) were identified in both populations (Fig. 3b). GBS yielded good coverage of SNPs for most of the *C. canephora* genome in both populations (Fig. 3b).

Genetic similarity between training and test populations is an important factor affecting prediction accuracy (de Los Campos et al. 2013; Daetwyler et al. 2013). Based on the GBS genotypes, the Intermediate and Premature population are very similar; the *Fst* measure is 0.0158, and both populations strongly overlap in the projection of the samples onto their first two principal components (Fig. 3a).

### Comparison of methods for genomic prediction

For phenotype prediction within the same environment, we evaluated prediction accuracy using 13 previously developed genomic prediction methods for three traits. Although the methods differ in assumptions of the marker effects, most methods yield predictions at comparable levels of accuracy (Fig. 4). The one exception is the **fixedMLR** approach that consistently yielded poor predictions.

Aside from the **fixedMLR** method, average predictive ability in different traits and environments ranged from 0.17 to 0.69 (Supplementary Table S4). On average, Bayesian methods were slightly more accurate than

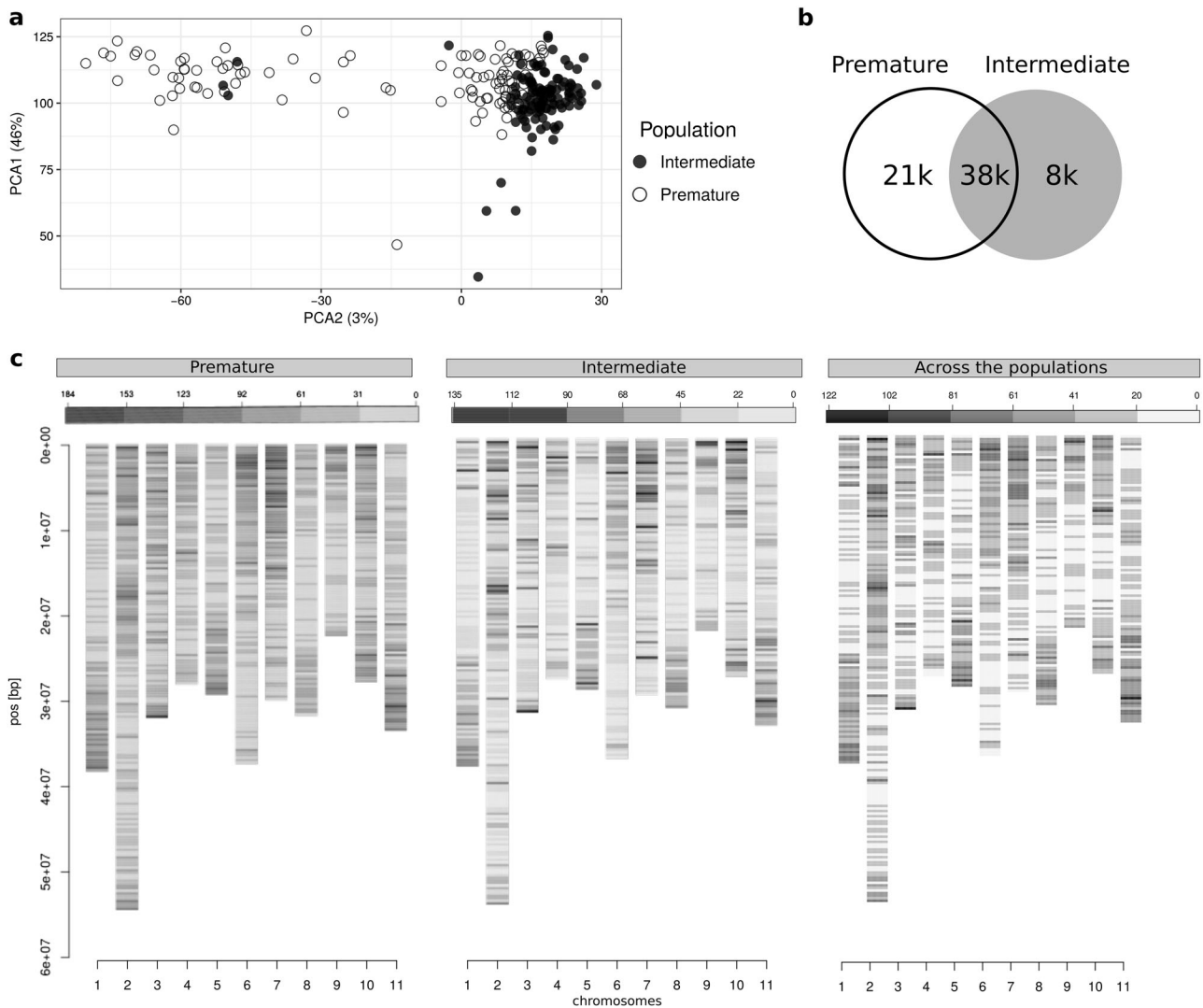
methods labeled as “machine learning” (0.47 versus 0.45). Differences in predictive ability were statistically significant ( $P < 0.05$ ) for different traits, locations and populations (Supplementary Table S2). Average predictive ability was slightly higher for incidence of leaf rust and yield of green beans (0.50 and 0.49, respectively) than production (0.38), in accordance with the heritability results (Table 2). Average prediction accuracy was slightly higher at the FEM location than FES (0.47 vs. 0.45), and predictions in the Premature population were on average more accurate than in the Intermediate population (0.54 vs. 0.39). Accordingly, similar performance across the competing models was observed for the slope and MSPE values (Supplementary Tables S5 and S6).

Although Bayesian approaches tend to yield higher predictive values, they typically come with a much greater computational cost than the alternatives methods (Fig. 5). Of the 13 methods compared, we found that the RR-BLUP and BSLMM methods achieved the best combination of high accuracy and low computational cost. Based on this result, in subsequent experiments we focused on the RR-BLUP method.

In order to check the impact of the SNP density on the predictive ability, we sampled SNPs across the chromosomes (i) guided by MAF and genotyping call rate value; and (ii) randomly sampled. Regardless to the approach taken, we observed a similar predictive ability across different SNP densities (Supplementary Figure S1). For several traits, predictive ability improved modestly as more SNPs were used, but for other traits we observed little to no improvement from using a larger number of SNPs (~35,000) versus a small subset of SNPs drawn from across the genome (e.g., ~4000 SNPs).

In Fig. 6, we summarize the prediction accuracy in different environments (Scenarios 1–12) using the RR-BLUP method. In most cases, positive prediction values were obtained (Fig. 6). When making predictions across locations (Scenarios 1–4), the predictive ability remained high for all traits. The incidence of rust and yield of green beans were consistently predicted with greater accuracy than production trait. These results suggest the potential for using GS models to make predictions in different locations. Across-population (Scenarios 5–8) analyses also yielded some positive values, but, on average, at lower magnitude than across-location predictions. Our results also indicated that production of grain beans was more impacted in across-population predictions and negative values were observed in Scenarios 6 and 8, respectively. Rust incidence and the yield of green beans yielded higher correlation values. Models trained in the Premature population and tested in the Intermediate population had lower prediction accuracy. This could be explained by the fact that the Premature population size was smaller than the Intermediate population.





**Fig. 3** **a** Principal component analysis (PCA) of the two *Coffea canephora* breeding populations; **b** Venn diagram is showing common and distinct SNPs to both population ( $k = 10^3$  SNPs) **c** SNP density

(number of SNPs per 400,000 Mb) across the 11 *C. canephora* chromosomes in Premature, Intermediate, and common to both populations

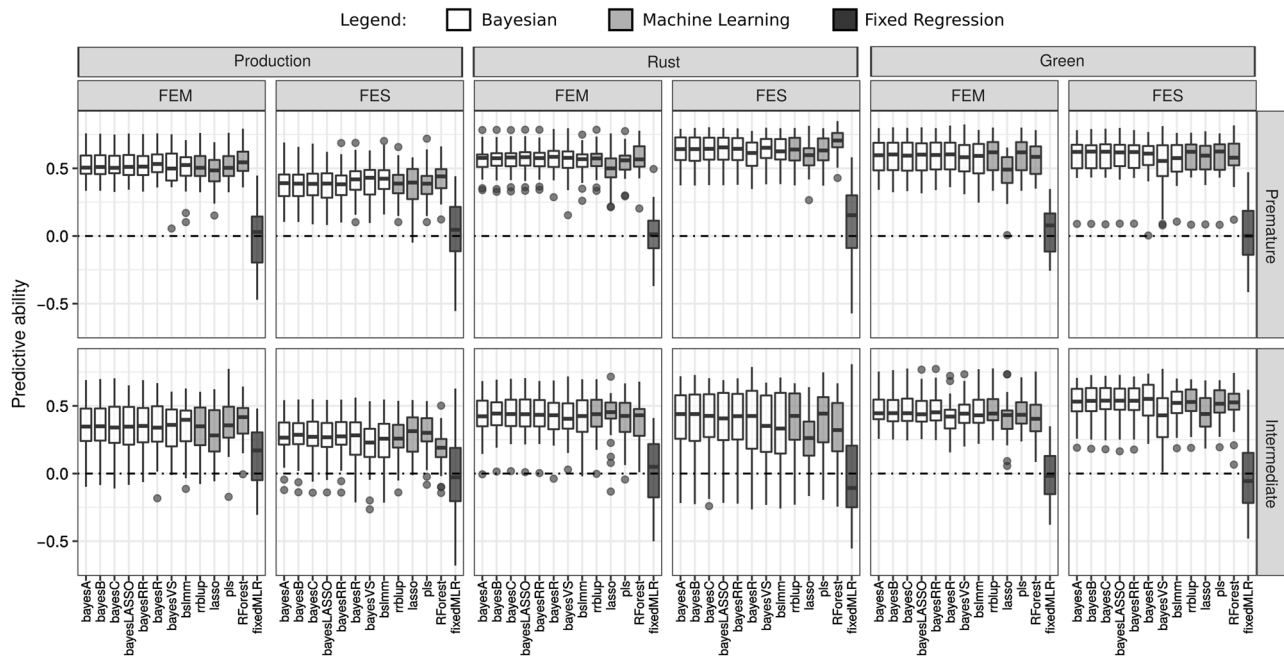
In the last set of scenarios, we evaluated prediction accuracy when both locations and populations differed between the training and test sets (Scenarios 9–12). As expected, overall lower predictive accuracy among all comparison were observed. This fact was more evidence for production traits that, once again, yielded negative values of predictive accuracy.

## Discussion

The benefits of GS compared with traditional phenotypic evaluations are well documented, and increasingly widely appreciated (Hayes et al. 2009; de Los Campos et al. 2013). Nonetheless we believe this potential remains under exploited in coffee crops. Possible reasons include: (i)

limited genomic resources available; (ii) difficulty in maintaining field experimentation given the long generation cycle, late expression of target traits and requirement of large areas for cultivation; and (iii) physiological makeup (low genetic diversity, ploidy barrier in *C. arabica*, and self-incompatibility in *C. canephora*). Herein, we have presented the potential to implement GS in conventional coffee breeding schemes. To this end, accurate phenotypic metrics, high-throughput genotyping and appropriate whole-genome statistical models are important requirements.

For studying coffee crops, complex quantitative traits are typically evaluated across multiple locations and harvests, which are collectively referred to as Multi-Environments Trials (MET) (Smith et al. 2005). Several statistical models have been proposed specifically for MET analyses, since the data collected in these setting typically violate basic



**Fig. 4** Evaluation of 13 statistical methods for predicting three coffee traits—production of coffee beans, incidence of coffee leaf rust and yield of green beans—in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two locations (FEM and FES). Refer to the “Genomic prediction methods” section for an overview of the methods compared. Prediction accuracy was computed as the

Pearson correlation between the predicted and adjusted phenotype in test samples. Cross-validation were performed using a Replicated Training-Testing. In each replication, 80% of the data were assigned to train the model and 20% remaining to validate the model. The procedure was replicated 30 times

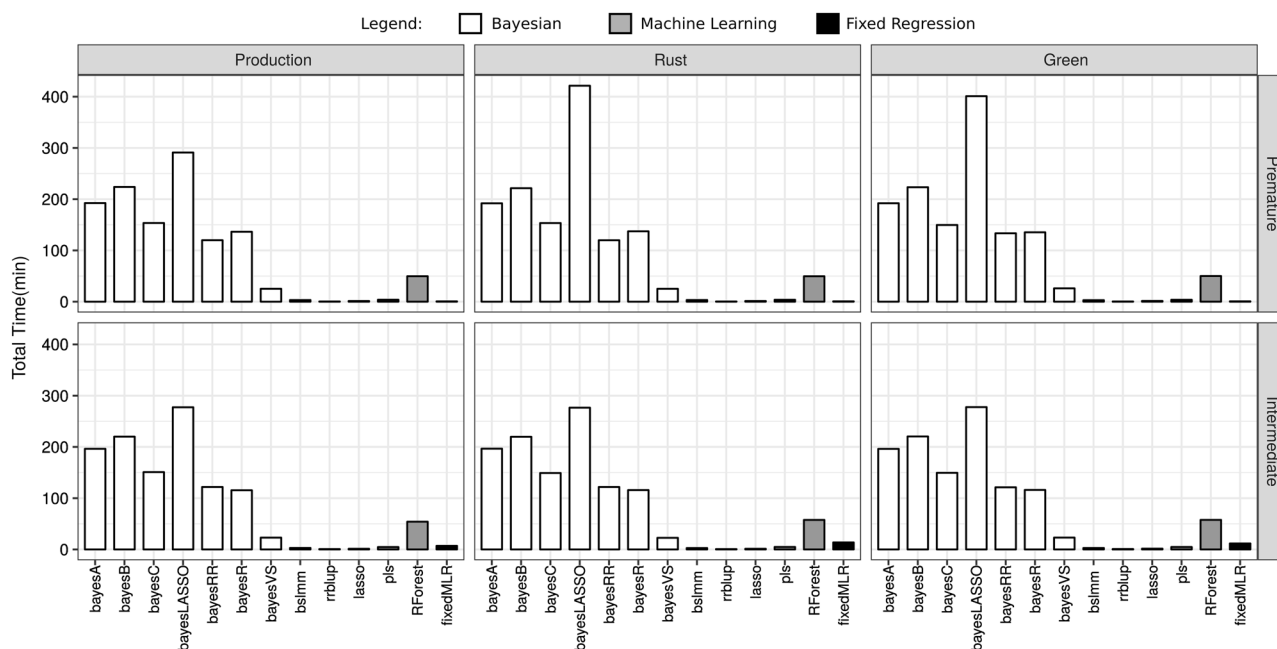
assumptions of conventional ANOVA models (e.g., homogeneity and independence of variances). As a consequence, bias can be introduced in estimation of genetic values, which may ultimately affect the predictive ability in GS studies. Guided by this previous work, we used a mixed model framework with appropriate covariance structures to account for genetic and non-genetic effects on the phenotypes. The flexibility to fit the residual and genetic variances showed better goodness-of-fit values than traditional ANOVA results.

High-throughput genotyping capacity has been increased by rapid progress in next-generation DNA sequencing (NGS). Genotyping-by-sequencing (GBS) is a product of this advance. Using GBS, we identified a total of 45,748 and 59,332 SNPs in the Intermediate and Premature populations, respectively. We emphasize that this total number of SNPs is larger than the set identified in a recent study on *C. arabica* that used a similar approach (DART methodology) (Del Moncada et al. 2015). This difference in SNP identification could be explained by the fact that *C. canephora* possesses higher genetic diversity due to its origin, reproduction method and dissemination (Ferrão et al. 2015).

For predictive analysis, we initially compared 13 predictive models on within-environment predictions. Assuming that GS models are align with the true genetic architecture of the phenotype, we were expecting a dependence between predictive ability and trait. For example,

**bayesRR** models assumes that marker effects are normally distributed with fixed variance, similar to the Fisher’s infinitesimal model proposes (Fisher 1919; Meuwissen et al. 2001). In contrast, **bayesB** assumes that most loci have no effect on the phenotypic variation, that is, traits controlled by relatively few loci whose effects vary in size (Meuwissen et al. 2001). Although conceptually different, we observed similar predicative performances of the competition models, evidencing a somewhat difference of our empirical results with previous simulation studies (Meuwissen et al. 2001; Coster et al. 2010).

Recently, several empirical evaluations have been published comparing predictive models and, like ours, reporting similar results across models (Moser et al. 2009; Heffner et al. 2011; Riedelsheimer et al. 2012; Resende et al. 2012b; Daetwyler et al. 2013; Wang et al. 2015). Some aspects of this similarity might be associated to statistical and biological properties. Statistically, the high discrepancy between number of observation and parameters can restrict the process of statistical learning resulting in similar predictive performances among methods (de Los Campos et al. 2013; Gianola 2013). Biologically, this similarity can be associated with the complex nature of traits. For real data, distribution of QTLs effects for most traits is perhaps less extreme than has been hypothesized in simulation studies (Hayes et al. 2009; de Los Campos et al. 2013; Daetwyler et al. 2013).



**Fig. 5** Total runtimes, in minutes, for fitting the 13 genomic prediction models in all cross-validation runs. Runtimes are shown separately for the three coffee traits—production of coffee beans, incidence of coffee leaf rust and yield of green beans—in the two *Coffea canephora*

populations (Intermediate and Premature). Cross-validation were performed using a Replicated Training-Testing. In each replication, 80% of the data were assigned to train the model and 20% remaining to validate the model. The procedure was replicated 30 times

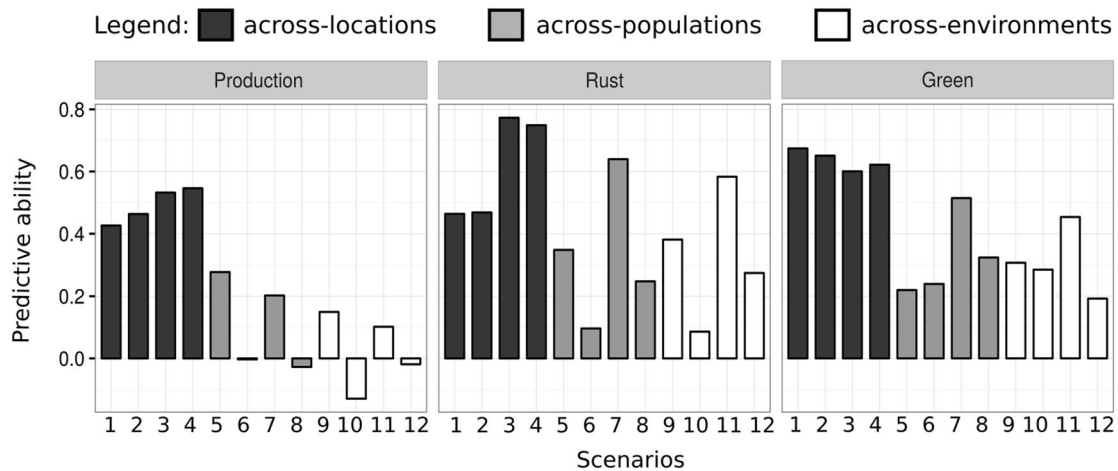
One method that consistently performed worse than the others was the fixed regression method (denoted by “**fixedMLR**” in the results). Fixed regression has been useful to detect genome-wide associations. However, these associations typically explain only a small fraction of the genetic variance of quantitative traits (Manolio et al. 2009). By contrast, methods that simultaneously fit effects for all markers are able to account for a much greater proportion of the genetic variation and, consequently, these approaches are more appropriate to predictive purposes (Meuwissen et al. 2001; Moser et al. 2009).

Contrasting to the predictive results, computational requirements significantly differ across the models. Consistent with previous studies, we found that Bayesian methods typically involved greater computational demand (Moser et al. 2009; Heslot et al. 2012; Neves et al. 2012). Particularly, computational cost is an important consideration since frequent re-estimation of marker effects is necessary in breeding programs (Moser et al. 2009). Judged by the overall performance, we found that RR-BLUP method presented important attributes for GS implementation, including straightforward implementation using existing mixed models software, relative simplicity, good performance, and limited computing time.

In a GS context, the possibility to predict phenotypic performance within and across environments is an outstanding question that has not been fully explored in coffee. As expected, within-environment predictions (Scenarios

A–D) yielded higher correlation values than cross-predictions (Scenarios 1–12). It has long been recognized that expression of genotypes are affected by environmental conditions and, as a consequence, across-location predictions (Scenarios 1–4) exhibited lower predictive performance than within-environment predictions (Scenarios A–D). In particular, this suggests that genotype-by-location (G×L) interactions are important, even considering that both locations are within the same breeding zone. In theory, G×L interactions occur because the capture and conversion abilities of a plant are determined by its particular ensemble of genes, which are expressed conditionally to the amount and quality of inputs received in each environment (Malosetti et al. 2013). This differential expression is captured by the estimate of marker effects and ultimately influences the predictions. Decaying accuracy across locations has been observed in GS studies in trees (Resende et al. 2012a; Beaulieu et al. 2014; Gamal El-Dien et al. 2015), cassava (Ly et al. 2013), and maize (Windhausen et al. 2012a, b).

For predictions in different populations (Scenarios 5–8), lower accuracy values can be explained by quantitative genetic concepts, which supports allele substitution effects varying between populations due to differences in allele frequency and LD pattern between SNPs and QTLs (Asoro et al. 2011; Windhausen et al. 2012a, b; Lehermeier et al. 2015). Similar results are presented in Neves et al. (2012) in mice populations and by Hayes et al. (2009) in dairy cattle. Scenarios 9–12 represented the most challenging condition



**Fig. 6** Predictive ability of cross-environment genomic predictions using the RR-BLUP method. “Environment” refers to a combination of location (FEM, FES) and population (Intermediate, Premature). As explained in Fig. 1, Scenarios 1–4 are used to evaluate GS performance across locations; Scenarios 5–8 assess GS performance across

populations; and Scenarios 9–12 evaluate prediction accuracy when both the location and population differ between the training and tests sets. Predictive ability was recorded for three coffee traits: production of coffee beans, incidence of coffee leaf rust, and yield of green beans

for GS as these scenarios combine predictions across locations and populations simultaneously. In most cases, models yielded poor prediction accuracy, with predictive abilities near to zero for the production trait.

Our results suggest the feasibility of incorporating GS into recurrent selection programs so long as predictive models are used to make predictions within the same environment, or within the same breeding zone. Traditionally, one cycle of phenotypic recurrent selection in coffee is divided into three phases: (i) progenies are obtained from a base population; (ii) field trials are conducted in multiple environments and harvests; and (iii) a new base population is generated via selection and recombination of the best individuals. In coffee, due to the long juvenile period associated with multiple evaluations across harvests, 5–6 years are required on average to complete a breeding cycle. Another challenge is that evaluating and maintaining multiple field trials is expensive and laborious. Therefore, incorporating GS prediction models can potentially reduce the time and expense of recurrent selection. We suggest applying GS methods in the second and third stages of recurrent selection programs by coupling prediction and selection during the seedling phase inside of greenhouses. Rapid-cycle recurrent selection supported by GS has potential to accelerate the increase of favorable alleles in the population and reduce both monetary and time costs associated with phenotyping (Windhausen et al. 2012a, b; Grenier et al. 2015). In a modern breeding scheme, phenotypic trials in multiple environments might be considered in advanced phases (e.g., third recurrent cycle), in order to re-estimate marker effects.

Outside the main topic of genomic prediction models, several other aspects of our study may be of interest to the

development of GS in coffee. For some traits, we found that prediction accuracy did not greatly improve as we included more SNPs in the models. In particular, we noted high predictive values for models trained with ~10,000 SNPs. Other studies have reported similar results (Vazquez et al. 2010; Spindel et al. 2015). According to de Los Campos et al. (2013), the point at which adding SNPs does not yield any improvement depends, mainly, on the span of LD in the genome and sample size. In both coffee populations, we expected large LD blocks since the populations were originated from only one cycle of recombination between a finite number of selected candidates. Regarding that resources need to be allocated to genotyping, 10,000 markers might be considered as a reference to design a SNP array as an alternative to reduce genotyping cost and popularize the use of genomic resources in coffee. Additional recommendations are given in Ferrão et al. (2017), including the possibility of exploiting the genomic information generated in GS investigations to guide parental selection.

Finally, we view this work as an initial investigation of genomic prediction in the coffee breeding, and there are additional questions that remain unanswered. In particular, we primarily focused on additive models for GS. Considering that *C. canephora* is a clonally propagated species, predictive models designed to explore the total genetic per se value of an individual is also a relevant question. Our option to predict additive effects was motivated by the breeding context, which includes the accumulation of favorable alleles through early and short cycles of recurrent selection. We believe that our results are sufficiently promising to justify further research, including the extension to modelling non-additive effects and incorporation of

genotype-by-environment (G×E) interactions by considering benefit from genetic correlations between locations (Lopez-Cruz et al. 2015; Cuevas et al. 2016, 2017). We also emphasize that one of the main difficulties in conducting coffee studies is the lack of information about genetic architecture of complex traits (Tran et al. 2016). Therefore, advances in our basic understanding of the genetic architecture, which include studies at genomic and phenotypic level, will lead to further improvements in GS.

## Conclusions

In this research, we explored the potential for genomic prediction in *C. canephora*, a perennial species for which GS concepts are appealing, but the basis for its implementation is still in its infancy. To this end, we investigated GS in multiple recurrent selection populations that were evaluated for three agronomic traits. In addition, we explored different models for accurate prediction of these traits within and across environments.

Some of our key findings include: (i) similar predictive abilities were obtained in most of the models that were compared, consistent with previous studies using real data; (ii) the predictive models exhibited very different computational costs; (iii) the RR-BLUP method achieved a good balance of predictive ability and computational cost; (iv) diversity and genetic relationship between training and testing data sets are important requirements; and (v) positive predictive ability supports the idea of implementing GS in conventional schemes of recurrent selection in coffee. Compared to traditional phenotypic methods, we expect that GS implementation can accelerate breeding cycles in coffee.

## Data archiving statement

The genotypes in the study belong to the germplasm collection and breeding program of the Incaper institution (ES, Brazil). Data available from Dryad <https://doi.org/10.5061/dryad.1139fm7>.

**Acknowledgements** This work was partially supported by FAPESP/CAPES (São Paulo Research Foundation), grants 2014/20389-2 and 2016/05127-7 for LFVF and AAFG. Phenotypic evaluations and GBS data is supported by Fapes (Espírito Santo Research Foundation), grants 55207464/11 and 65192036/14. Additional support was provided by the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper) and Embrapa Cafe. We thank Livia Souza and Anete P. de Souza (CBMEG, Unicamp/Brazil) for their assistance in DNA extraction; Paulo Volpi (Incaper/Brazil) for his support with phenotype evaluations; and members of the Stephens lab (University of Chicago, USA) and the Statistical Genetics lab (ESALQ, Brazil) for their feedback on the results. Finally, we thank three anonymous referees and the editor for helpful comments on the initial submission.

**Author contribution** L.F.V.F, A.A.F.G, R.G.F, M.A.G.F, M.S, and A.F conceived the study and designed the experiments. R.G.F, M.A.G.F, and A.F developed the experimental design and collected the phenotype data. L.F.V.F performed the DNA extraction. L.F.V.F, M.S, P.C, and A.A.F.G performed the genomic prediction analysis. L.F.V.F wrote the paper, with input from all authors.

## Compliance with ethical standards

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

## References

- Akaike H (1974) A new look at statistical model identification. *IEE Trans Autom Control* 19:716–723
- Asoro FG, Newell Ma, Beavis WD, Scott MP, Jannink J-L (2011) Accuracy and training population design for genomic selection on quantitative traits in Elite North American Oats. *Plant Genome J* 4:132
- Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J (2014) Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC Genom* 15:1048
- Carbonetto P, Zhou X, Stephens M (2017) varbvs: Fast variable selection for large-scale regression. *arXivpreprint arXiv:170906597*
- Carbonetto P, Stephens M (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal* 7:73–108
- Chan AW, Hamblin MT, Jannink J-L (2016) Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data. *PLoS ONE* 11:e0160733
- Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet Sel Evol* 42:1–11
- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C et al. (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda)* 3:1903–1926
- Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P, de los Campos G (2017) Bayesian genomic prediction with genotype x environment interaction kernel models. *G3 Genes, Genomes, Genet* 7:41–53
- Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, de los Campos G et al (2016) Genomic prediction of genotype x environment interaction kernel regression models. *Plant Genome* 3:1–20
- Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. *J Agr Biol Envir St* 11:381
- Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193:347–365
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. (2011) The variant call format and VCFtools. *Bioinforma* 27:2156–2158
- Denoeud F, Carretero-paulet L, Dereeper A, Guyot R, Pietrella M, Zheng C et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. 345.
- Elshire RJ, Glaubitz JC, Sun Q, Poland Ja, Kawamoto K, Buckler ES et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379



- Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome J* 4:250
- Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci* 95:4114–4129
- Ferrão L, Caixeta E, Pena G, Zambolim E, Cruz C, Zambolim L et al. (2015) New EST–SSR markers of *Coffea arabica*: transferability and application to studies of molecular characterization and genetic mapping. *Mol Breed* 35:1–5
- Ferrão LFV, Ferrão RG, Ferrão MAG, Francisco A, Garcia AAF (2017) A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*. *Tree Genet Genomes* 13:95
- Fisher RA (1919) XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 52:399–433
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Software* 33:1–22
- Gamal El-Dien O, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby Ya (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genom* 16:370
- Garrick D, Dekkers J, Fernando R (2014) The evolution of methodologies for genomic prediction. *Livest Sci* 166:10–18
- Gianola D (2013) Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194:573–596
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. 9
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330
- Grenier C, Cao T-V, Ospina Y, Quintero C, Châtel MH, Tohme J et al. (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PLoS ONE* 10: e0136594
- Habier D, Fernando R, Kizilkaya K, Garrick D (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186
- Hayes BJ, Bowman PJ, Chamberlain aJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443
- Heffner EL, Jannink J, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome* 4:65–75
- Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci* 52:146–160
- James G, Witten D, Hastie T, Tibshirani R (2013). *An introduction to statistical learning*. (vol. 112) New York: Springer.
- Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. *Genetics* 191:969–987
- Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
- Lehermeier C, Schon C-C, de los Campos G (2015) Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* 201: 323–337
- Liaw A, Wiener M (2002) Classification and regression by random-Forest. *R News* 2:18–22
- Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L et al. (2015) Increased prediction accuracy in wheat breeding trials using a marker x environment interaction genomic selection model. *G3 Genes/Genomes/Genet Genes/Genomes/Genet* 5:569–582
- 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* 193:327–345
- Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG et al. (2013) Relatedness and genotype x environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci* 53:1312–1325
- Malosetti M, Ribaut J-M, van Eeuwijk FA (2013) The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front Physiol* 4:44
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Del Moncada PM, Tovar E, Montoya JC, González A, Spindel J, McCouch S (2015) A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. *Tree Genet Genomes* 12:1–17
- Moser G, Tier B, Crump R, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet Sel Evol* 41:56
- Neves HHR, Carvalheiro R, Queiroz SA (2012) A comparison of statistical methods for genomic selection in a mice population. *BMC Genet* 13:100
- Oakey H, Cullis B, Thompson R, Comadran J, Halpin C, Waugh R (2016) Genomic selection in multi-environment crop trials. *G3 Genes, Genomes, Genet* 6:1313–1326
- Park T, Casella G (2008) The Bayesian Lasso. *Journal of the American Statistical Association* 103:681–686
- Pastina MM, Malosetti M, Gazaffi R, Mollinari M, Margarido GRA, Oliveira KM et al. (2012) A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theor Appl Genet* 124:835–849
- Pérez PR, de los Campos G (2013). BGLR: Bayesian generalized linear regression. *R Packag version*
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2013). nlme: Linear and Nonlinear Mixed Effects Models
- R Core Team (2013). R: A Language and Environment for Statistical Computing
- Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D et al. (2012a) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol* 193:617–624
- Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM et al. (2012b) Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.). *Genet* 190:1503–1510
- Riedelsheimer C, Technow F, Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genom* 13:452
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Silva FF, Varona L, de Resende MDV, Filho JSSB, Rosa GJM, Viana JMS (2011) A note on accuracy of Bayesian LASSO regression in GWS. *Livest Sci* 142:310–314

- 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* 143:449
- Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E et al. (2015) Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. *PLoS Genet* 11:1–25
- 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* 96:3300–3312
- Vazquez AI, Rosa GJM, Weigel KA, De los Campos G, Gianola D, Allison DB (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J Dairy Sci* 93:5942–5949
- Wang X, Yang Z, Xu C (2015) A comparison of genomic selection methods for breeding value prediction. *Sci Bull* 60:925–935
- Wehrens R, Mevik BH (2007). *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*
- Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. *Genet Res.* 75:249–252
- Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28:2086–2087
- Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J.-L, Sorrells ME et al. (2012a) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 (Bethesda)* 2:1427–1436
- Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J.-L, Sorrells ME et al. (2012b) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3 (Bethesda)* 2:1427–1436
- Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* 9:e1003264