**University of São Paulo**
**"Luiz de Queiroz" College of Agriculture**

# Development and application of statistical genetic methods to genomic prediction in *Coffea canephora*

**Luís Felipe Ventorim Ferrão**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

**Piracicaba**
**2017**

**Luís Felipe Ventorim Ferrão**
**Bachelor in Biological Sciences**

# Development and application of statistical genetic methods to genomic prediction in *Coffea canephora*

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

**Piracicaba**
**2017**

**Dados Internacionais de Catalogação na Publicação**
**DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Ferrão, Luís Felipe Ventorim
Development and application of statistical genetic methods to genomic prediction in *Coffea canephora* / Luís Felipe Ventorim Ferrão. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2017 .
60 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Seleção genômica 2. Marcadores moleculares 3. Modelos lineares 4. Café . I. Título.

# DEDICATORY

WITH LOVE
to my parents Romario and Liliâm,
my brothers Guilherme and Arthur,
my awesome future wife, Juliana Benevenuto,
and in loving memory
to my grandmother Maria Teresa Gava Ferrão,
and my grandfather Anizio Ventorim.

## ACKNOWLEDGEMENTS

To the University of São Paulo, specially the Superior School of Agriculture "Luiz de Queiroz"-ESALQ for the provision of physical and intellectual infrastructure to develop this project.

Financial support from the FAPESP/CAPES (São Paulo Research Foundation), grants 2014/20389-2 and 2016/05127-7, is gratefully acknowledged. Phenotypic and genotypic evaluations were supported by Fapes (Espírito Santo Research Foundation), grants 55207464/11 and 65192036/14. Additional support was provided by the Instituto Capixaba de Pesquisa, Assitência Técnica e Extensão Rural (Incaper) and Embrapa Café.

Appreciation goes to my advisor, Prof. Antonio Augusto Franco Garcia, for his support, patience, and encouragement throughout my graduate studies. His technical and friendly advice was essential to the completion of this dissertation and has taught me innumerable lessons and insights on the workings of academic research in general.

Special acknowledgement goes to Prof. Matthew Stephens (University of Chicago, USA), who contributed a significant proportion of their valuable knowledge and time, and provided encouragement and assistance when it was needed.

Additionally, I would like to thank all the Incaper team (Dr. Romario G. Ferrão, Dra. Maria A. G. Ferrão, Dr. Aymbiré Fonseca and Paulo Volpi) for their contributions, encouragement and assistance. My sincere thanks also goes to Profa. Anete Pereira de Souza and Dra. Livia Souza, who provided me an opportunity to join their team in the CBMEG Lab (Unicamp, SP) as intern, and who gave access to the laboratory and research facilities. Without they precious support it would not be possible to conduct this research.

I also thanks all the staff and professors of ESALQ/USP for their help throughout my academic time. Special thanks go to Prof. Gabriel Margarido, Profa. Cláudia B. Monteiro Vitorello, Prof. Roberto Fritsche Neto, Profa. Clarice Demetrio and Prof. Roland Vencovsky (*in memorian*) for their valuable guidance during classes and discussions about genetic, breeding, statistical and programming computation. I could not forget to mention a few names in the team staff, including Berdan, Seu Antonio, Valdir, Fernandinho and Léia, who made my stay at the Department of Genetics much more enjoyable. I also would like to mention others who participated in my academic training at the Federal University of Viçosa (UFV) including Prof. Paulo Roberto Cecon, Prof. Cosme Damião Cruz, Prof. Fabyano Fonseca and Dra. Eveline Caixeta.

Others that have given their time, advice and support to help me complete this dissertation include all the students in the Statistical Genetic Lab (ESALQ/USP). Certainly, we have a cooperative and friendly working group! In particular, I would like to mention the "old group"of the Laboratory, which four years ago received and helped me, including Marcelo Mollinari, Guilherme Pereira, Rodrigo Amadeu and João Ricardo. I also thank other colleagues who assisted me during this time, including Amanda Avelar, the students of Claudia's Laboratory (ESALQ/USP) and all old friends and roommates of my last city (Viçosa / MG).

I extend my acknowledgments to my parents in law (Benevenuto family) and to all my family, (Ventorim and Ferrão families) including my cousins, uncles and aunts in both sides .

Last, but not least, special thanks go to my parents, Romario Gava Ferrão and Liliâm Maria Ventorim Ferrão, and my brothers, Guilherme Ventorim Ferrão and Arthur Ventorim Ferrão; whose advice, wisdom and support have been priceless. Said clearly: "Pais e irmãos, minha eterna gratidão pelo amor e apoio incondicional. Amo vocês!".

Finally, I would like to express my gratitude to my future wife, Juliana Benevenuto, for your support through all the frustration and success, thank you.

# SUMMARY

# RESUMO

## Desenvolvimento e aplicação de métodos genético-estatísticos para predição genômica em *Coffea canephora*

Seleção Genômica pode ser definida como a seleção simultânea de centenas ou milhares de marcadores moleculares, os quais cobrem o genoma de forma densa, de modo que locos de caracteres quantitativos (QTL) estejam em desequilíbrio de ligação com uma parte desses marcadores. Assim, marcadores associados a QTLs, independentemente da significância dos seus efeitos, são utilizados na predição do mérito genético de um indivíduo para um determinado caráter. Simulações e estudos empíricos mostram que essa abordagem apresenta acurácia suficiente para garantir o sucesso em programas de melhoramento genético, quando comparado com os métodos tradicionais de seleção fenotípica. Para tanto, uma das etapas requeridas é o uso de modelos genético-estatísticos que contemplem a predição fidedigna da performance fenotípica da população sob estudo. Apesar da relevância, o número de estudos no gênero *Coffea* ainda são reduzidos, não havendo relatos sobre o desempenho desses modelos em diferentes populações e ambientes, ou mesmo, a sua performance para diferentes caracteres agronômicos do cafeeiro. Dessa forma, este estudo tem como finalidade investigar aspectos relacionados a modelagem estatística, a fim de compreender quais são os fatores que tornam os modelos preditivos mais acurados e utiliza-los em programas aplicados de melhoramento genético. Dados reais de duas populações de seleção recorrente de *Coffea canephora*, avaliados em dois ambientes e genotipados pela tecnologia de genotipagem por sequenciamento (GBS, do inglês *Genotyping-by-Sequencing*) foram considerados para o estudo da relação entre genótipo-fenótipo. Em termos de modelagem estatística, duas classes de modelos foram considerados: i) Modelos mistos, baseados no cálculo da matriz de parentesco realizado como medida de (co)variância genética entre indivíduos (modelo GBLUP); e ii) Modelos de associação multilocos, no qual milhares de marcadores moleculares são modelados simultaneamente e os efeitos estimados dos marcadores são somados, a fim de computar o mérito genético dos indivíduos. Ambas estratégias foram descritas em capítulos separados no formato de artigo científico. O capítulo intitulado "A mixed model to multiplicative harvest-location trial applied to genomic prediction in *Coffea canephora*" abordou uma expansão do modelo GBLUP de modo a contemplar efeitos de interações entre Genótipo×Colheita e Genótipo×Local. Para tanto, apropriadas estruturas de variância e covariância para modelagem da heterogeneidade e correlação dos efeitos genéticos e residuais foram testadas. O modelo proposto, denominado de MET.GBLUP, apresentou melhor qualidade de ajuste e capacidade preditiva, quando comparado com outros métodos. O capítulo em sequência, intitulado de "Comparison of statistical methods and reliability of genomic prediction in *Coffea canephora* population" investigou a capacidade preditiva de diferentes modelos de associação multilocos. A suposição usual de efeitos dos marcadores amostrados de uma distribuição normal foi relaxada, a fim de testar métodos alternativos que pudessem melhor descrever o fenômeno biológico e, consequentemente, resultar em maior capacidade preditiva. Embora os modelos testados sejam conceitualmente distintos, diferenças mínimas nos valores de acurácia de predição foram observadas nos cenários testados. Em termos de demanda computacional, modelos Bayesianos apresentaram maior tempo de análise. Os resultados descritos em ambos os capítulos apoiam o potencial do uso da seleção genômica em programas de melhoramento assistido de café. Em termos práticos, comparado com métodos tradicionais de avaliação fenotípica, é esperado que a implementação desses conceitos em programas de seleção recorrente possam acelerar o ciclo de melhoramento, manter a diversidade genética e, sobretudo, aumentar o ganho genético por unidade de tempo.

**Palavras-chave:** Seleção genômica, Marcadores moleculares, Modelos lineares, Café

# ABSTRACT

## Development and application of statistical genetic methods to genomic prediction in *Coffea canephora*

Genomic selection (GS) works by simultaneously selecting hundreds or thousands of markers covering the genome so that the majority of quantitative trait loci are in linkage disequilibrium (LD) with such markers. Thus, markers associated with QTLs, regardless of the significance of their effects, are used to explain the genetic variation of a trait. Simulation and empirical results have shown that genomic prediction presents sufficient accuracy to help success in breeding programs, in contrast to traditional phenotypic analysis. For this end, an important step addresses the use of statistical genetic models able to predict the phenotypic performance for important traits. Although some crops have benefited from this approach, studies in the genus *Coffea* are still in their infancy. Until now, there have been no studies of how predictive models work across populations and environments or, even, their performance for different complex traits. Therefore, the main objective of this research is investigating important aspects related to statistical modeling in order to enable a more comprehensive understanding of what makes a robust prediction model and, as consequence, apply it in practical breeding programs. Real data from two experimental populations of *Coffea canephora*, evaluated in two brazilian locations and SNPs identified by Genotyping-by-Sequencing (GBS) were considered to investigate the genotype-phenotype relationship. In terms of statistical modelling, two classes of models were considered: i) Mixed models, based on genomic relationship matrix to define the (co)variance between relatives (called GBLUP model); and ii) Multilocus association models, which thousands of markers are modeled simultaneously and the marker effects are summed, in order to compute the genetic merit of individuals. Both approaches were considered in separated chapters. Chapter entitled "A mixed model to multiplicative harvest-location trial applied to genomic prediction in *Coffea canephora*" addressed an expansion of the traditional GBLUP to accommodate interaction effects (Genotype×Local and Genotype×Harvest). For this end, we have tested appropriate (co)variance structures for modeling heterogeneity and correlation of genetic effects and residual effects. The proposed model, called MET.GBLUP, showed the best goodness of fit and higher predictive ability, when compared to other methods. Chapter in the sequence was entitled "Comparison of statistical methods and reliability of genomic prediction in *Coffea canephora* population" and addressed the use of different modelling assumptions considering multilocos association models. The usual assumption of marker effects drawn from a normal distribution was relaxed, in order to seek for a possible dependency between predictive performance and trait, conditional on the genetic architecture. Although the competitor models are conceptually different, a minimal difference in predictive accuracy was observed in the comparative analysis. In terms of computational demand, Bayesian models showed higher time of analysis. Results discussed in both chapters have supported the potential of genomic selection to reshape traditional breeding programs. In practice, compared to traditional phenotypic evaluation, it is expected to accelerate the breeding cycle in recurrent selection programs, maintain genetic diversity and increase the genetic gain per unit of time.

**Keywords:** Genomic selection, Molecular markers, Linear models, Coffee

# 1 PREFACE

Coffee is the world's most widely traded tropical agricultural commodity (Tran *et al.*, 2016). It is estimated that more than 125 million people have been benefited, directly or indirectly, by the coffee agribusiness (IOC, 2016). As result, the crop is part of the economy of more than 70 countries and it is one of the most popular beverages in western countries (Moncada *et al.*, 2015). In this scenario, Brazil has a prominent position, given it is responsible for about a third of all world production making it the world's largest producer, a position that has held for the last 150 years (IOC, 2016). For this reason, among the activities related to agricultural business in the country, coffee crop has been one of the most important in economic and social aspects.

Coffee belongs to the Rubiaceae family and the genus *Coffea*, which comprises hundreds of tropical species. Among them, two species present commercial production: *Coffea arabica*, more aromatic with more perceptible acidity; and *Coffea canephora*, which beverage have a bitter, full bodied taste and higher caffeine level (Tran *et al.*, 2016). In the 50's, with the raise of soluble coffee consume, *C. canephora* species, known as a coffee of lower quality, began to be commercially exploited in the so-called blends (coffee drink composed by grain mixture of both species). In addition to counteract the acidity and add full bodied taste, blends conferred good industrial efficiency which resulted in low cost and, hence, more competitive being therefore of more interest. This fact boosted world production of *C. canephora*, particularly, in tropical countries. Popularly known as "Robusta coffee", currently, Brazil stands out as the second largest producer in the world. In this context, Espírito Santo (ES) State is responsible for 78% of all grains produced in the country. This total represents 20% of the *C. canephora* worldwide production representing the importance of the crop in a global scenario (Ferrão *et al.*, 2007).

Much of this success is due to the breeding program that has been conducted by the Incaper Institution (Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural). Since the first variety developed by the Incaper was released (1993), it is estimated that the average productivity increased in the State in the order of 310%, with an increase of only 7.5% in the area. Nine *C. canephora* cultivars were released by Incaper. Despite this evident success, traditionally, breeding schemes in coffee are still based entirely on phenotypic evaluations collected in field trials. It is undeniable that important advances were obtained in the last decades. However, it is also important to take into account the time required to achieve these gains (Ferrão *et al.*, 2007). Breeding programs supported only by phenotypic metrics are coupled with long testing phases resulting in low gains per unit of time.

The possibility to predict accurately the genetic merit based on molecular information, a process known as genomic selection (GS), is revolutionizing breeding schemes (Jannink *et al.*, 2010). The importance and interest in this methodology is driven by the desire to increase the rate of genetic gain per unit of time. This is caused by a higher of selection, when compared with traditional selection schemes. Additionally, genomic selection allows for selection of juvenile plants without phenotypes. Although some crops have benefited from this contemporary approach (Grattapaglia and Resende, 2010; Poland *et al.*, 2012a; Crossa *et al.*, 2013; Spindel *et al.*, 2015), studies in the genus *Coffea* are still in their infancy. Until now, there is not evidence supporting how predictive models works across populations and environments or, even, their performance for different complex traits. In this scenario, studies in *Coffea canephora* can be considered as a good starting point. Despite its economic importance there are genetic motivations, including the ploidy (2n = 2x) and wide genetic variability (Ferrão *et al.*, 2015). Both features make the genotyping and statistical modelling more feasible than in *C. arabica*, which is allotetraploid and has a narrow genetic base. Furthermore, the first high-quality genome sequence of Robusta coffee was recently completed and reported, which supports the use of *C. canephora* species as an important model in coffee investigations (Denoeud *et al.*, 2014).

In order to investigate the GS performance in coffee breeding, the main objective of this research

is to discuss aspects of statistical modelling for genomic prediction. Until now, multiple methods and models have been proposed. As a general rule, these approaches combine concepts of quantitative genetic (Falconer and Mackay, 1996), linear regression (Rencher and Schaalje, 2008), mixed models (Henderson, 1949), genetic relationships (VanRaden, 2008) and Bayesian analysis (Gelman *et al.*, 2014). Aiming to introduce these topics, a critical overview about GS implementation was considered in the chapter "*Genomic Selection-State of the Art*" , that is part of the book "*Genetic Improvement of Tropical Species*", under responsibility of the *Springer* editor. Due to copyright issues, this chapter was omitted in this dissertation.

Subsequent chapters (1 and 2) focus on aspects and development of genomic prediction models and their performance considering coffee data set. Multiple methods and models have been proposed for implementing genomic selection (VanRaden, 2008; de Los Campos *et al.*, 2013). In statistical terms, prediction begins with the specification of a model involving effects and other parameters that describe the factors that determine observed values (Garrick *et al.*, 2014). In GS context, a statistical model is proposed to associate phenotypic observations with variations at DNA level. The major challenge of genomic prediction researches is to accurately model the true QTL effects. This challenge is caused by the disparity between the large number of markers (p) and the number of records (n) that are available to predict marker effects. This is the well documented "curse of dimensionality"or "p>n statistical problem"(Gianola and van Kaam, 2008).

Any model to be used for genomic prediction must be able to accommodate more predictors than observations, which prevents the use of the classical theory of linear models (e.g., ordinary least square or maximum likelihood) (Gianola, 2013). In the literature, two different approaches have been widely used for this end: i) Mixed models based on genomic relationship matrix; and ii) multilocus association models (also called "polygenic modeling" or "marker effects models") (Kärkkäinen and Sillanpää, 2012; Zhou and Stephens, 2012; Garrick *et al.*, 2014). Both methods were addressed in Chapter 1 and 2, respectively.

Mixed model approach is a method that utilizes genomic relationships to estimate the genetic merit of an individual. For this purpose, a genomic relationship matrix is estimated from DNA marker information. The matrix defines the covariance between individuals based on observed similarity at the genomic level, rather than on expected similarity based on pedigree. The similarity with the traditional BLUP (Henderson, 1949) motivated the classification of this method as "Genomic BLUP" or "GBLUP". Chapter 1, entitled "*A Mixed model to multiple harvest-location trial applied to genomic prediction in Coffea canephora*" considered this approach for genomic predictions in coffee. Some key points were discussed, as the following: i) Evaluate the GS performance, in contrast to phenotypic methods; ii) Handle the interaction effects (Genotype×Harvest and Genotype×Location) in GS context; iii) Given the modest genomic resources and the absence of a standard genotyping platform, investigate the potential of the Genotyping-by-Sequencing (GBS). In order to consider the raised points, a predictive model was proposed addressing the coffee breeding scenario, which involves measures in a series of replicated field trials grown across multiple years and location. Experiments of this nature are typically referred to as multi-environment trials (MET) (Smith *et al.*, 2005). The central point discussed in this chapter was an expansion of the GBLUP model in order to accommodate interaction effects.

In order to address the conjugate use of genomic information and MET modeling, appropriate (co)variance structures for modeling heterogeneity and correlation of genetic effects and residual effects were considered. Among the advantages, the flexibility to consider correlated information for the genetic and residual terms is an important factor, since they are not easy to handle considering traditional analysis (e.g., ANOVA models) (Smith *et al.*, 2005; Malosetti *et al.*, 2014). This approach has been used in recent years in our research group for data modeling in perennial crops, in special for sugarcane crop. Much of these ideas were described by Pastina *et al.* (2012) in QTL mapping studies and by

Balsalobre *et al.* (2016) in phenotypic analysis. Therefore, given our expertise in this topic, it was a natural way to consider the mixed model theory as a starting point for genomic prediction studies in coffee. This study was supported by FAPESP (Sao Paulo Research Foundation), grant 2014/20389-2. The approach and the results was orally presented at Coffee Workshop during the PAG XXIV (Plant and Animal Genome Conference), San Diego, USA; and submitted for publication in *Tree Genetics & Genomes* journal.

The research covered in the Chapter 2 addressed the use of multilocus association models to predictive analysis. Thousands of markers are modeled simultaneously and the genetic value of an individual is obtained by the sum of these estimated effects (Garrick *et al.*, 2014). To this end, all markers have been included as explanatory variables under a Bayesian framework or considering Machine Learning algorithms (Kärkkäinen and Sillanpää, 2012; Zhou *et al.*, 2013; James *et al.*, 2013). This categorization in Bayesian or Machine Learning group occurred in accordance to how they tackle the underlying statistical question about "p>n problem", a data dimensionality dilemma where the number of markers (p) significantly exceeds the number of phenotypic records (n). As a general rule, in this scenario some modelling assumptions are required either by discarding the unimportant predictors or by shrinking their effects toward zero (Kärkkäinen and Sillanpää, 2012). The procedure adopted will differentiate the methods in terms of predictive ability, computational efficiency and genetic assumptions.

Although comparisons between methods have been carried out in different species and traits, to our knowledge, investigations in coffee are still modest. Therefore, Chapter 2, entitled as "*Comparison of statistical methods and reliability of genomic prediction in Coffea canephora populations*", was addressed to compare the performance of a range of genomic prediction models across *C. canephora* traits. Likewise, these analysis were extended for predictions across locations and populations of coffee, in order to check the reliability of GS studies to predict genetic merit in multiple conditions of the plant breeding. This research was developed during the FAPESP/BEPE (Research Internship Abroad, grant 2016/05127-7) at the University of Chicago, USA; under the supervision of Prof. Matthew Stephens. Results was orally presented at Coffee Workshop during the PAG XXV (Plant and Animal Genome Conference), San Diego, USA. This chapter was wrote in a manuscript format expressing our intention to submit it in the *Genetics – G3 Genes/Genomes/Genetics* journal. To the best of our knowledge, this is the first research addressing the use of predictive models in multiple traits and populations in the genus *Coffea*.

In the final chapter (Chapter 3), a summary of the key findings are presented and the implications of the research outcomes are discussed. In addition, the potential impacts of this research in the coffee breeding community are discussed as possible future directions, indicating the increasing potential of genomic selection.

# 2 A MIXED MODEL TO MULTIPLE HARVEST-LOCATION TRIAL APPLIED TO GENOMIC PREDICTION IN *COFFEA CANEPHORA*

**Keywords:** Genomic Selection; Genotyping-by-Sequencing(GBS); GBLUP; Multi-Environments Trials (MET); Perennial crops.

## 2.1 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 multiples locations and harvests (MET) 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 proposed and genomic analysis was performed using a genotyping-by-sequencing (GBS) approach, which showed good potential to be used in coffee breeding programs. Interactions were modeled using the multiplicative mixed model theory, that is commonly used in MET analysis in perennial crops. The effectiveness of the proposed method was compared with other genetic models in terms of goodness of fit and predictive accuracy. Different scenarios that mimic coffee breeding were used in the cross-validation process. The proposed approach had the lowest AIC and BIC values and, consequently, the best fit. In terms of predictive capacity, 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.

## 2.2 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. It 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 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. The potential to increase the rates of genetic gain and reduce the breeding cycle is a widely accepted concept in animal and plant breeding. Popularly called

genomic selection (GS), the methodology has potential to redirect resources and activities in breeding programs (de los Campos *et al.*, 2009).

Although GS is a promising method to help 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 in which the species is inserted. Typically, coffee trials consist of evaluations in multiple locations and harvests to understand interactions and predict the performance of untested genotypes. Experiments of this nature are collectively referred to as Multi-Environments Trials (MET) and are not restricted to coffee but are also used in perennial crops (Kelly *et al.*, 2009).

A large number of statistical models have been developed to address interactions in MET studies. In a modern framework, the genotypic performance across environments has been modeled as a correlated trait. 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 natural 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 the best linear unbiased prediction (BLUP) (Kelly *et al.*, 2009).

BLUP methodology relies on pedigree information to define the covariance between known relatives. However, this covariance can also be defined at the genomic level using DNA information rather than an expected value based in pedigree record. This matrix 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, traditional GBLUP was extended to accommodate the 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.*, 2014; 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 about issues in relation 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 as sugarcane (Pastina *et al.*, 2012; Margarido *et al.*, 2015), forage grass (Smith and Casler, 2004) and cereal (Kelly *et al.*, 2009).

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 a representative approach 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 the one-step approach, GBS can discover new markers and genotype entire populations. It is rapid, flexible and perfectly suited for GS versus traditional molecular markers. Investigations using GBS are common in many crops (Poland *et al.*, 2012b; Crossa *et al.*, 2013), but there is a still an important gap in the coffee literature.

The central purpose of this research was to consider a genomic selection model suitable for use

in *C. canephora* and other perennial crops. The breeding scenario in which the species is inserted was considered and the importance of the interaction investigated. To the best of our knowledge, studies applying high-throughput genotyping to genomic predictions are still relatively novel in *Coffea*. We present aspects related to the applicability of Genotyping-by-Sequencing (GBS) as well as future perspectives.

## 2.3 Material and Methods

### 2.3.1 Phenotypic data

The experimental population was developed and evaluated by the Instituto Capixaba de Pesquisa, Assistência Tecnica e Extensão Rural Incaper; ES State, Brazil. Phenotypic data 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 clones were selected as progenitor due the high production and the same grain maturity date. The latter is an important trait for new coffee varieties because it allows for harvest standardization.

After one cycle of recombination and evaluation, the top 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 $19^0 24'$ south, longitude $40^0 31'$ west, 70 m altitude; and Sooretama Experimental Farm (FES) - latitude $15^0 47'$ south, longitude $43^0 18'$ west, 40 m altitude. he complete experiment used 3570 coffee trees, and the average of each plot was evaluated for grain production (kilograms of mature coffee fruit in the cherries stages) over four consecutive harvest-production years (2008, 2009, 2010 and 2011).

### 2.3.2 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 treatment were collected and lyophilized. DNA extraction used 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 is implemented in the TASSEL-GBS (v.4.3.7) (Glaubitz *et al.*, 2014). Sequenced tags were aligned against the *C. canephora* genome (Denoeud *et al.*, 2014). The raw 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 10x (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). The graphical analyzes were performed using the OmicCircos (Hu *et al.*, 2014).

### 2.3.3 Phenotypic Models

The following model uses a notation presented by Pastina *et al.* (2012). The statistical model in which the underlined terms indicate a random variable is:

$$\underline{y}_{ijkr} = \mu + L_j + B|L_{rj} + H_k + LH_{jk} + \underline{G}_{ijk} + \underline{e}_{ijkr} \tag{2.1}$$

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 $\mu$ is the overall mean; $L_j$ is the effect of location;

**Table** 2.1: Variance and covariance structures examined for the random effects in model 2.1

| **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.

$B|L_{rj}$ 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 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 $k^{th}$ harvester and the genetic variance within the $j^{th}$ location, respectively. The VCOV structures for these matrices are represented in Table 3.1. 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: 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 $\Sigma_g$ 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 ($\Sigma_g$). 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 lake of covariance between individuals). The identity assumption ensures that 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 via to genetic effects. It assumed 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 to harvest, location and block, respectively. The $I_g$ is an identity residual (co)variance matrix to genotypes. In principle, all the structures mentioned in Table 3.1 were tested for the residual term. In addition, spatial adjustments were tested, in order to correct 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.

### 2.3.4 GBLUP version to multiples harvest-location trial (MET-GBLUP)

The aforementioned Model 2.1 was used to test the importance of the interaction modeling (Genotype × Location- G×L and Genotype × Harvest- G×H) as well as the inclusion of the molecular

**Table** 2.2: Summary of the tested models and the assumption on the variance and covariance structure related to the random effects specified in the Model 2.1 description. MET prefix on the methods indicates models where the interaction is explicitly modeled, testing covariance structures for location and harvest.

| **Method** | $\mathbf{G}^a$ | $\mathbf{R}^a$ |
|---|---|---|
| **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 2.1. The $I_G$, $A_p$ and $A_m$ represent a Identify matrix, additive relationship matrix and realized kinship, respectively. [1] First class of methods, that ignored the multiple-environment (MET) modeling; [2] Second class of methods, that considered the multiple-environment (MET) modeling.

information in predictive models. Thus, different assumptions about the random effects distribution were tested. Two classes of models were defined in accordance with interaction inclusion (MET modeling) (Table 2.2).

The first class ignored the MET modeling; simple structures for genetic and residual 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 $\theta_{xy}$) 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 was computed using molecular marker information considering $A_m = MM'$, where $M$ is the matrix containing the SNPs information centered on the average and standardized by the variance (VanRaden, 2008)

The second class of models considered the MET modeling. Here, the genetic and residual matrices were modeled considering the structures cited in Table 3.1 as well as variations on the genetic covariances ($\Sigma_g$ 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 multiples harvest-location trial" and refers to the idea of accommodating the G ×L and G ×H interactions using MET theory and a genomic selection model (GBLUP).

### 2.3.5 Comparison of Models

Two criteria were used to compare the models (Table 2.2): i) Goodness of fit, via AIC (Akaike, 1974) and BIC (Schwarz, 1978) ; and ii) Predictive capacity measured by cross-validation. The cross-validation considered three hypothetical scenarios in coffee breeding. Scenario 1 (CV1) represents the full genotypic prediction, i.e, simulation of genotypes that were not evaluated in any block, location and harvest. Scenario 2 (CV2) represents genotypic 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 made in CV1 followed by CV2 and CV3.

The predictive abilities were assessed using a Replicated Training-Testing evaluation. In each replication, 90% of the individuals 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

**Table** 2.3: Goodness of fit for the genetic matrix, factored by location ($G_L$) and harvest ($G_H$) considering the AIC and BIC criteria. 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.

random assignments into TRN and TST. A similar scheme was used by Crossa *et al.* (2013). The predictive capacity was measured using the accuracy mean and the mean squared prediction error (MSPE) across the 10 repetitions. The predictive accuracy 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. All analysis were performed using the Genstat software (VSN International, 2011).

## 2.4 Results

### 2.4.1 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 2.3). The values of ID combinations highlight the poor quality of the goodness of fit data when traditional ANOVA assumptions are considered — even when homogeneous variances across locations and harvesters are applied.

All the structures mentioned in Table 3.1 were also tested for the residual. Convergence problems and negative variance components were 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, we highlight that 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, spacial adjustment was tested to corrects possible trends in the field trial data. No improvements on the AIC and BIC criterion were observed when the correlation between immediately neighboring plots was specified (results not shown).

Figure 2.1 presents the phenotypic dispersion across the harvests and the variance component magnitude. The dispersion of the phenotypic observations showed 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 differential behavior 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 2010) and low production years

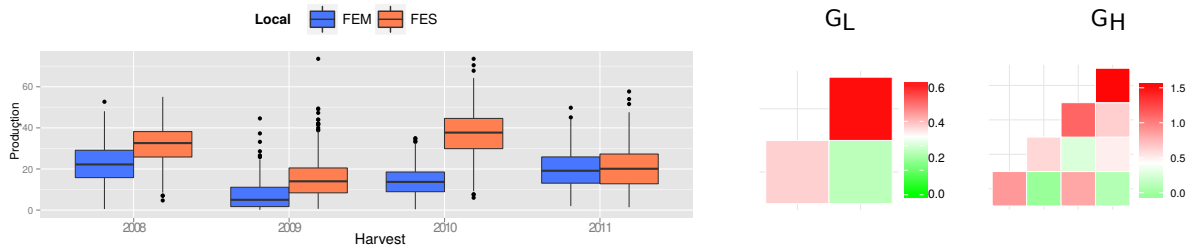


**Figure** 2.1: Boxplot of grain production (kilograms of mature coffee fruit in the cherries stages) across the locations (FEM and FES) and harvests (2008,2009,2010 and 2011), and a heatmap representing the unstructured form estimated for locations ($G_L$) and harvests ($G_H$)

**Table** 2.4: SNP density summary in the *Coffea canephora* GBS libraries for each chromosome (Chr), considering the sequential filtering: Trial: removing all triallelic SNPs; mAF: removing SNPs with mAF $< 0.01$, plus Trial filtering; MD: removing SNPs that are present in less than $< 50\%$ of the samples, plus Trial and mAF filtering; Depth Coverage: removing SNPs with mean number of sequence reads per locus averaged across all individuals less than 1x, 5x, 10x and 15x, plus Trial, mAF and MD filtering.

| **Chr** | **Raw**[a] | **Trial** | **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 SNPs in a unfiltered VCF format.
[b] The percentage of SNPs remaining after filtering.

(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.

### 2.4.2 Genotypic Data

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 2.4). 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 SNP total, but 15x 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 (Figure 2.2). The first (from outer to inner

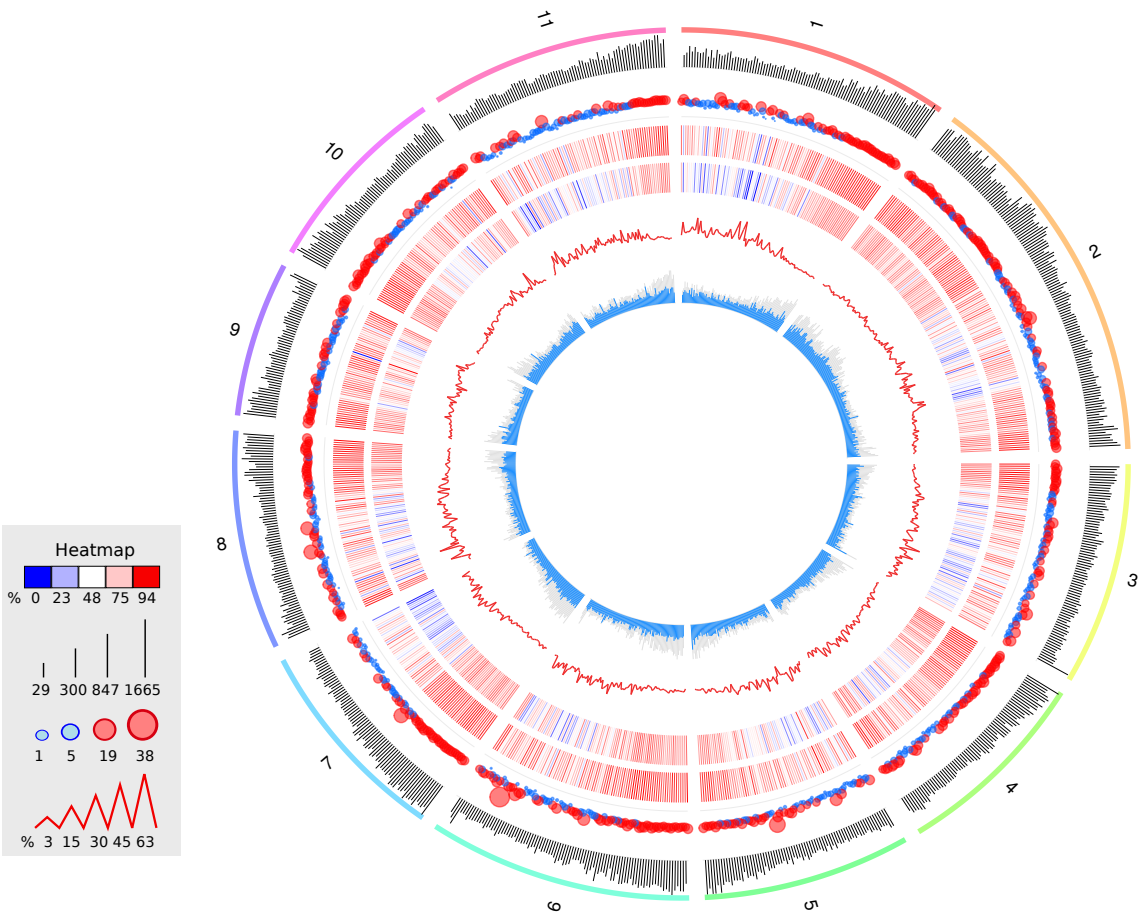


**Figure** 2.2: Circular visualization about GBS information across the *Coffea canephora* chromosomes. From from outer to inner layers, the graphic is separated by seven layers: i) Chromosomes; ii) number of raw SNPs; iii) depth coverage; iv) percentage of SNPs eliminated considering the Minor Allele 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.

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 int 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 is the percentage of SNPs per window with Minor Allele Frequency (mAF) lower than or 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.

### 2.4.3 MET and GS models

Models that ignored the MET modeling (**Id**, **BLUP** and **GBLUP**) showed higher AIC and BIC values and hence poor fit (Table 2.5). The inclusion of molecular pedigree consistently improved the results based on the criteria of minimum AIC and BIC. The **MET.GBLUP** proposed here had the

**Table** 2.5: AIC and BIC values for models with different variance and covariance structures for the genetic and residual random effects. MET prefix on the methods 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** |

Bold 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 2.1. The $I_G$, $A_p$ and $A_m$ represent a Identify matrix, additive relationship matrix and realized kinship, respectively.

**Table** 2.6: Predictive capacity measured by the accuracy (r) and mean squared prediction error (MSPE) considering six modeling methods and three breeding scenarios: Scenario 1 (CV1) represents the full genotypes prediction, *i.e*, genotypes that were not evaluated in any block, location and year (harvesting); Scenario 2 (CV2) represents the prediction of missing genotypes in one of the environments; Scenario 3, represents the genotypic prediction 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 |

Bold numbers represent the greatest r values and the smallest MSPE.

lowest AIC and BIC values and was the best model.

The second class of models had better predictive ability (Table 2.6). In the most complex scenario (CV1) the difference in predictive accuracy 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 close accuracy implying that inclusion of kinship could not improve the predictive ability.

Another comparative criterion used during the cross-validation was the MSPE, which was held in the perception of the distance among true 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. 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.

## 2.5 Discussion

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. In coffee breeding programs, genotype performances are typically measured in a series of replicated field trials grown across

multiple years and locations. Experiments of this nature are typically referred to as multi-environment trials (MET), and they measure the performance of genotypes across a range of environmental conditions that cultivars might be exposed to. The challenge in this case 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.

Here, we proposed an expansion of the traditional GBLUP to address the conjugate use of genomic information and MET modeling. A similar approach was described by Burgueño *et al.* (2012), 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 species with similar experimental design.

Multiplicative mixed models have been commonly used for MET analysis (Smith *et al.*, 2001, 2005). 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 genotypics, 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-harvester 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) reported that the most general form is the fully unstructured (UNS) matrix, although it often leads to troubles during estimation. 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 description of biology. Pastina *et al.* (2012) and Margarido *et al.* (2015) 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 was observed when spatial correlation was fitted (results not shown). This might be because of the experimental design, which was not a typical square 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). This has evidence of the necessity to perform selection in each specific location. 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 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). Here, we reinforce the relevance of using more than one criterion to draw conclusions. The goodness of fit, commonly used in QTL studies (Pastina *et al.*, 2012), 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 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 in 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 siblings, for example, could show a variation from 0.4 to 0.6 in the genomic relationship matrix. A fixed value of 0.5 is calculated using only the pedigree record. The exploitation of this level variation usually results in better goodness of fit 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 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 recovery the information via the covariance matrix (Malosetti *et al.*, 2016). It also offers the use of 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. Account the interaction has

been showing as an important source of variation in many phenotypic studies (Burgueño *et al.*, 2011) as well as in GS studies (Burgueño *et al.*, 2012; Malosetti *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 by the **MET.GBLUP** and were compared with the top genotypes selected via competing methods. Changes in the genotype ranking were observed and support the results of Kelly *et al.* (2009). The main goal of a breeding program is identify the best genotypic performance for commercial release and to use this genotype as a parent in future crosses. In essence, the changes in the genotypic ranking means erroneous selection and thus a loss of gain selection in subsequent generations.

Finally, the low number of studies using high throughput genotyping in coffee motivates a brief discussion of this subject. 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 in traditional molecular approaches. 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 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.

## 2.6 Conclusion

We highlighted the GS approach as an important approach in marker assisted selection. For coffee, the reduction of repeated cycles of selection, breeding and testing are the main motivation. Developing new cultivars can take decades, but this can be accelerated with GS implementation. Good prospects have been reported in maize (Crossa *et al.*, 2013), wheat (Poland *et al.*, 2012b), rice (Spindel *et al.*, 2016) and forest (Grattapaglia and Resende, 2010).

In our context, prediction based on GS models will be considered in the selection of progenies for the program of recurrent selection in *C. canpehora*. We believe that some factors are essentials for the GS implementation: 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. Imputation methods and improved on the bioinformatic steps, specially in the SNP and genotype calling, are important future trends in coffee studies. In terms of statistical modeling, studies focusing on the importance of non-additives and epistatic effects are necessary. Finally, the use of alternative approaches, such as hierarchical Bayesian regressions, are an important perspective for future studies (Ferrão *et al.*, 2016).

## Acknowledgments and grant support

This work is partially supported by FAPESP/CAPES (Sao 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. 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 Contribution

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.

### Conflict of Interest

The authors declare that they have no conflict of interest

# 3 COMPARISON OF STATISTICAL METHODS AND RELIABILITY OF GENOMIC PREDICTION IN *COFFEA CANEPHORA* POPULATIONS

**Keywords:** Bayesian model; Genomic Selection; G×E interaction; Genotyping by Sequencing (GBS); Perennial crops;

## 3.1 Abstract

Genomic selection is defined as the prediction of genetic merit of individuals based on dense genotypic marker information. Simulation and empirical results have been shown that predictions based on molecular data present sufficient accuracy to help success in breeding programs. Although some crops have benefited from this contemporary approach, studies in the genus *Coffea* are still in their infancy. Until now, there have been no studies of how predictive models work across populations and environments or, even, their performance for different complex traits in coffee. Considering that predictive models are based on biological and statistical assumptions, it is expected that their performance varies depending on the true underlying genetic architecture of the phenotype. We used real data from two experimental populations of *Coffea canephora*, evaluated in two locations, and SNPs identified by Genotyping-by-Sequencing (GBS) to investigate the genotype-phenotype relationship. For this end, we considered thirteen prediction models commonly used in genomic selection analysis, including penalized and Bayesian estimation procedures, as well as nonparametric regressions and dimension reduction procedure. Analysis were extended for predictions within-environment and predictions across locations and populations, in order to check the reliability of GS results to predict the genetic merit in multiple scenarios of the plant breeding. Considering the three traits under analysis (grain production, leaf rust incidence and yield of green grains), we observed minimal differences in terms of predictive accuracy among the competitor models. Bayesian methods showed a slight superiority, although more computation was required. Predictive accuracies for within-environment analysis, on average, were higher than predictions across locations and populations. Results discussed in this research have supported the potential of GS to reshape traditional coffee breeding schemes. In practice, compared to traditional phenotypic evaluation, GS is expected to accelerate the breeding cycle, maintain genetic diversity and increase the genetic gain per unit of time.

## 3.2 Introduction

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

The advent of molecular markers opened an important perspective to achieve fast and longstanding genetic gains (Lande and Thompson, 1990). For this purpose, Meuwissen *et al.* (2001) suggested use all available molecular markers to predict quantitative traits in animal and plant breeding. Known as Genomic Selection (GS), the methodology became widely accepted by its potential to maximize the genetic gain and reduce the breeding cycle. Theoretical bases of this process is that whenever marker density is high enough, most QTL 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 trait (Goddard and Hayes, 2007).

The major challenge of genomic prediction is to accurately model the true QTL effects, since all markers are included as potential explanatory variables considering sparse regression models (Kärkkäinen and Sillanpää, 2012; Zhou *et al.*, 2013; Garrick *et al.*, 2014). This is caused by disparity between the large number of SNP markers ($p$) and the number of records ($n$) that are available to estimate the SNP effects. As a result, parameters of interest (e.g., marker effects) cannot be estimated accurately considering classical theory of linear models (i.e, ordinary least square or maximum likelihood) (Gianola, 2013). This leads to a situation where some kind of selection of the predictors is required, either by discarding the unimportant predictors or by shrinking their effects toward zero (Kärkkäinen and Sillanpää, 2012). Several analytical approaches have been proposed for genome-based prediction of genetic values, including penalized and Bayesian estimation procedures, as well as nonparametric regressions and dimension reduction procedure (Gianola *et al.*, 2009; de Los Campos *et al.*, 2013).

Comparisons between predictive models have been carried out in different scenarios for different species and traits (Riedelsheimer *et al.*, 2012; Resende *et al.*, 2012; Daetwyler *et al.*, 2013; Heslot *et al.*, 2014; Crossa *et al.*, 2016; Thavamanikumar *et al.*, 2015; Wang *et al.*, 2015). However, empirical and simulations studies have shown the absence of a benchmark algorithm, since biological and technical factors can affect the predictive accuracy, such as, the population size, genetic architecture and relation among the training and validation data set (de Los Campos *et al.*, 2013; Daetwyler *et al.*, 2013). All these factors have supported the search for a convenient predictive model to be considered in the implementation of genomic selection. Although some crops have benefited from this investigation, studies in the genus *Coffea* are still in their infancy.

Another important aspect related to practice implementation is define the breeding scenario in which genomic prediction will be applied (Windhausen *et al.*, 2012). So far, GS prediction accuracy has mostly been evaluated within single environments (Windhausen *et al.*, 2012; Beaulieu *et al.*, 2014; Gamal El-Dien *et al.*, 2015). In coffee, breeding schemes are commonly delineated in multiple environments, in order to measure the performance of genotypes across a range of conditions. However, it is unknown whether marker effects estimated in a set of environments are useful to predict the genotype performance in a different set of environments. If feasible, this means that predictive models could be calibrated in a specific condition and used to predict phenotypic performance in other environments, resulting in time and cost economy.

Until now, there is no evidence in coffee research supporting how predictive models work across populations and environments or, even, their performance for different complex traits. In this sense, using *C. canephora* species as a starting point in studies applied to the genus *Coffea* is considered a promising approach. Economic and genetic reasons can be argued in favor of this idea. The genetic motivation is based on the species ploidy (2n = 2x) and the wide genetic variability; both make the process of genotyping and statistical modeling more feasible than in *C. arabica* - allotetraploid and with a narrow genetic base (Ferrão *et al.*, 2015; Tran *et al.*, 2016). Economic motivations are based on the grain production and crop cultivation. It is estimated that *C. canephora* is responsible for 40% of the world coffee production and its grain is the main source of raw material for soluble coffee (Tran *et al.*, 2016). Further, the species has lower costs of production, mainly due to less stringency for control of biotic and abiotic factor, crop management and primary processing (van der Vossen *et al.*, 2015).

Given the potential of GS to reshape breeding programs and achieve fast selection gains, this research addressed the comparison of performance of a range of genomic prediction models across *C. canephora* traits. This includes machine learning algorithms and the Bayesian framework. These analyses were extended for predictions across-locations and across-populations, in order to check the reliability of GS studies to predict genetic merit in multiple conditions of the plant breeding. To the best of our knowledge, this is the first studies that discuss the use of predictive models in a broad scope for multiple traits and populations in the genus *Coffea*.

## 3.3 Material and Methods

### 3.3.1 Experimental data

Experimental population was designed by the Instituto Capixaba de Pesquisa, Assistência Tecnica e Extensão Rural Incaper; ES State, Brazil. Phenotypic data consisted of two recurrent selection populations formed from the recombination of superior clones of *C. canephora*. To summarize in brief, of the thousands of genotypes maintained in Incaper these clones were selected as progenitor (or founders) due the high production and same grain maturity group. The latter is an important requirement for new coffee varieties because it allows for harvest standardization. Based on the maturity group, coffee populations were designated as Intermediate and Premature. Intermediate population, on average, started the grain maturity on March/April, which harvests in June. Premature genotypes show an anticipation of grain maturity and the harvest, on average, one month before.

Intermediate and Premature population were designed from the recombination of 16 and 9 progenitors, respectively. After one cycle of recombination and evaluation, the founder and 103 progenies in the Intermediate population and 87 progenies in the Premature population were cloned and evaluated in randomized complete blocks with three replications and five plants per plot. These populations were installed in two representative environments (locations) for the Brazilian production of *C. canephora*: Marilândia Experimental Farm (FEM) - latitude $19^0 24'$ south, longitude $40^0 31'$ west, 70 m altitude; and Sooretama Experimental Farm (FES) - latitude $15^0 47'$ south, longitude $43^0 18'$ west, 40 m altitude.

The complete experiment used 3570 coffee trees in the Intermediate population and 2880 coffee trees in the Premature population. Over four consecutive harvest-production years (2008, 2009, 2010 and 2011), both populations were evaluated for grain production (kilograms of mature coffee fruit in the cherries stage); natural infection of coffee leaf rust, caused by the *Hemileia vastatrix* fungus (scale of scores ranging from 1 to 9, according to visual sporulation intensity evaluated in field); and the yield of green grains (post-harvest trait, grams of mature grains after processed by dry methods for removing the entire dried husk in samples of 2 kilograms of coffee fruit in the cherries stage).

### 3.3.2 Experimental design

In this research the GS potential was investigated under two aspects in coffee breeding: i) Comparing models in terms of predictive ability across different traits and conditions; and ii) checking the GS performance for within and across-environments predictions. For this end, specific scenarios were designed, as follow. Scenarios A, B, C and D represent GS implementation for within-environment prediction (Figure 3.1). Here, by environment we mean a specific combination between location and population. In these scenarios, predictive abilities were assessed using a Replicated Training-Testing evaluation (Crossa *et al.*, 2016). In each replication, 80% of the individuals were assigned randomly for training data set (TRN), while the remaining 20% were assigned for testing data set (TST). This division was replicated 30 times with independent random assignments into TRN and TST. Models were fitted to the TRN data set and prediction accuracy was evaluated in the TST data set.

Scenarios 1, 2, 3 and 4 represented GS performance across-locations (Figure 3.1). For this purpose, the training (TRN) and testing (TST) data set ranged in accordance to the scenario under analysis. For example, in Scenario 1 the predictive ability was conditioned for the same population, but considering different locations. A predictive model was calibrated in the Intermediate population and FEM location (TRN dataset) and, then, the estimated marker effects were used to predict the genetic merit of individuals in the FES location (TST dataset). Following the same reasoning, scenario 5, 6, 7 and 8 represented the predictive ability conditioned for the same location, but considering different populations (called across-populations predictions). Finally, scenarios 9, 10, 11 and 12 represented across-

environments predictions.

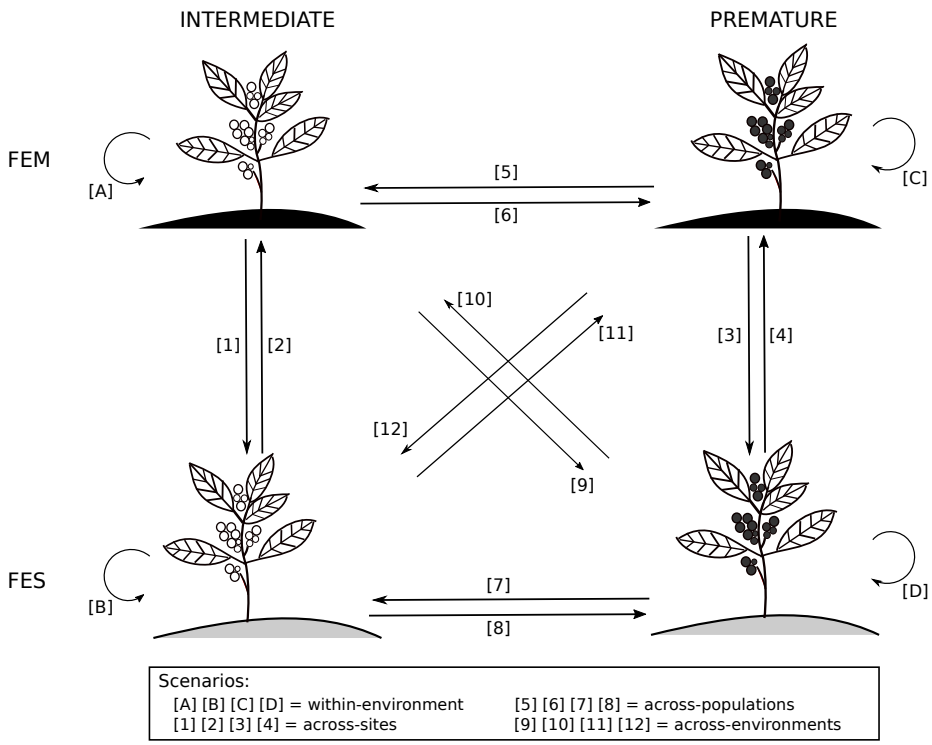


**Figure** 3.1: Scenarios were genomic selection was investigated. Here, by environment we mean a specific combination between location (FEM or FES) and population (Intermediate or Premature). Scenarios A, B, C and D represented GS performance for within-environment and where genomic selection models were compared. Scenarios 1, 2, 3 and 4 represented GS performance across-locations. Scenarios 5, 6, 7 and 8 represented GS performance across-populations. Scenarios 9, 10, 11 and 12 represented across-environments predictions. Direction of the arrows represented changes on training and testing data sets.

### 3.3.3 Phenotypic Models

A phenotypic model was adjusted for each combination of environment and trait. Coffee experimental design are represented, in the majority, for longitudinal data across multiple harvests (years). Different VCOV structures were tested, in order to better explain this temporal variation.

Using a similar notation to presented by Pastina *et al.* (2012), the following statistical model was used (underlined terms indicate a random variable):

$$\underline{y}_{ijk} = \mu + B_j + H_k + \underline{G}_{ik} + \underline{\epsilon}_{ijk} \tag{3.1}$$

where $\underline{y}_{ijk}$ is the phenotype of the $j^{th}$ block (j=1,2,3) of the $i^{th}$ individual (i = 1,2...,n) and $k^{th}$ harvest (k = 1,2,3,4); $\mu$ is the overall mean; $H_k$ is the harvest effect; $B_j$ is the block effect; $\underline{G}_{ik}$ is a random genetic effect of individual $i$ at harvests $k$; $\underline{\epsilon}_{ijk}$ and is a random non-genetic residual error term.

For the genetic effects, we assumed a multivariate normal distribution with a zero mean vector and a VCOV matrix indexed by two factors (harvest and genotype) written as the Kronecker product ($\otimes$) of matrices, as follow: $G = \sum_H^{k\times k} \otimes I_g^{n\times n}$; in which $\sum_H^{k\times k}$ is VCOV relate to harvest. The diagonal element represents the genetic variance within the $k^{th}$ harvest. Several VCOV structures for this matrix were investigated (Table 1). Two important points were considered in MET analysis: i) Each structure has different assumptions about the heterogeneity of variance and may be used to quantify the interactions;

**Table** 3.1: Variance and covariance (VCOV) structures examined for the random effects in Model 3.1 for *Coffea canephora* phenotypic analysis.

| **Model** | **Num.Par**[a] | **Description** |
|---|---|---|
| Ident | 1 | Identical variation |
| Diag | K | Heterogeneous variations |
| CompSym | K+1 | Compound symmetry with heterogeneous variance |
| Uns | K(K+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. K is the number of harvests.

and ii) the number of estimated parameters represents the variation in the degree of complexity. The term $I_g$ is an Identity matrix.

The residual term was factored in a similar way. It assumed a multivariate normal distribution implying a zero mean and VCOV matrix indexed by three factors: $R = R_H^{k\times k} \otimes I_B^{r\times r} \otimes I_g^{n\times n}$. The terms $I_g$ and $I_B$ are an Identity residual matrix to genotypes and blocks, respectively. For the term $R_H^{k\times k}$, structures "Ident" and "Diag" were tested (Table 3.1).

In order to choose an appropriate phenotypic model, VCOV structures were examined and compared via AIC (Akaike Information Criterion) (Akaike, 1974) and BIC (Bayesian Information Criterion) (Schwarz, 1978) criteria. The calculation of the heritability in complex linear mixed models is not straightforward (Cullis *et al.*, 2006; Oakey *et al.*, 2016). Here, broad-sense heritability ($h^2$) was computed from the simplest phenotypic model (Identity structure for the genetic and residual matrix) as: $h^2 = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2/bh}$; where $\sigma_g^2$ is the genotype variance component, $\sigma_e^2$ is the residual variance component, $b$ and $h$ are the number of blocks and harvests, respectively. All analysis were performed using the "nlme" R-package (Pinheiro *et al.*, 2016), an open-source statistical software.

### 3.3.4 Genotypic Data

Both Intermediate and Premature populations were genotyped using the GBS approach Elshire *et al.* (2011). The GBS protocol followed that from the Genomic Diversity Facility, Cornell University (http:// www.biotech.cornell.edu/brc/genomics-facility). Leaves of each treatment were collected and lyophilized. DNA extraction used 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 is implemented in the TASSEL-GBS (v.4.3.7) (Glaubitz *et al.*, 2014). Sequenced tags were aligned against the *C. canephora* genome (Denoeud *et al.*, 2014). SNPs were extracted from the raw VCF file and filtered manually considering the following cutoff: i) Triallelic SNPs were removed; ii) Minimum minor allele frequency (0.01 mAF) and iii) SNPs that are present in less than $< 70\%$ of the samples were eliminated.

To ensure that genotype were defined in a consistent manner, we considered a probabilistic model to perform the genotype calling in SNPs with low coverage ($\leq 5$ reads sequenced at genotyped loci). For this end, a Bayesian framework was proposed in order to consider the genetic background during the inferential process. A similar model is described by Chan *et al.* (2016). Here, a two-step approach was proposed, which first the parental genotypes are computed using the likelihoods defined by Chan *et al.* (2016) and assuming a uniform genotype prior; and, given their genotypes, these probabilities are considered as prior information in the inference of the genotype progenies. Two important requisites should be considered: i) The proposed model is valid for populations whose progenitors are known; ii)

It is essential that parental information is accurate, in order that a priori distribution is well defined to call the progenies genotypes. In our experiments, accurate information were achieved increasing the sequencing coverage depth (progenitors were sequenced 3x more than progenies).

All SNP manipulation and genotype calling were carried out using VCFtools package (Danecek *et al.*, 2011), customized scripts in R (R Core Team, 2013) and bash (GNU, 2007). The graphical analyzes were performed using the "symbreed" R package (Wimmer *et al.*, 2012).

### 3.3.5 Genomic predictions

The following genetic model was fitted to estimate the genetic marker effects:

$$\boldsymbol{y} = 1_n\mu + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim MVN_n(0, \boldsymbol{I}_n\tau^{-1})$$

Here $\mathbf{y}$ is an $n$-vector of genetic values measured on $n$ individuals, adjusted for the environmental effects as described on the "Phenotypic Models" section; $\mathbf{X}$ is a $n{\times}p$ matrix of genotypes measured on the same individuals at $p$ genetic markers; $\boldsymbol{\beta}$ is a $p$-vector of (unknown) SNP effects; $1_n$ is an $n$-vector of 1's; $\mu$ is a scalar representing the mean, and $\boldsymbol{\varepsilon}$ is an $n$-vector of error terms that have variance $\tau^{-1}$. $MVN_n$ denotes the n-dimensional multivariate normal distribution.

Model choice may be important insofar as it must align with the genetic architecture of the trait (Beaulieu *et al.*, 2014). In order to compare different models applied to genomic predictions, we defined three class of methods commonly used to GS purposes: fixed multiple regression, machine learning algorithm and Bayesian framework.

#### 3.3.5.1 Fixed Multiple Regression

This class of method mimic traditional GWAS algorithms ("single-SNP" analysis), which test each SNP, one at a time, for association analysis with the phenotype. Fixed regression using a subset of markers derived from "single-SNP" analysis served as our non-GS marker-based prediction control, as applied by Spindel *et al.* (2015). For each replication in the cross-validation scheme (Replicated Training-Testing evaluation), single marker regression was run for all markers and p-values determined for each marker by F-test. Linear models were then tested using the first 100 most significant markers.

#### 3.3.5.2 Machine Learning algorithm

Machine learning algorithm refers to a vast set of tools dedicated to building and studying methods that are capable of learning from data, endeavoring to find an optimal solution to minimizes a given loss (Xavier *et al.*, 2016). Supervised machine learning methods involves, in general, statistic models addressed to predict an output based on one or more inputs (James *et al.*, 2013). Three classes of supervised machine learning methods have been considered to predict genetic values: regularized regression, dimension reduction methods and random forest.

Regularized regression fit a model containing all $p$ predictor, regularizing the coefficients estimates, or equivalently, shrinking the SNP estimates toward to zero. Estimates are derivate as the solution to an optimization problem that balances goodness of fit and model complexity, represented by: $\hat{\beta} = \sum_{i=1}^{n}(y_i - \mu - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda J(\beta)$; where $\lambda \geq 0$ is a regularization parameter that controls the trade-off between lack of fit and model complexity. Several regularized approaches have been proposed and and they differ on the choice of penalty function (James *et al.*, 2013). Ridge-regression (RR) and LASSO are the two most popular ones. The RR estimator solves the regression problem using the L2

norm, while LASSO estimator consider the L1 norm (Hoerl and Kennard, 1970; Tibshirani, 1996; de Los Campos *et al.*, 2013). As practical result, RR shrink all of coefficients toward zero, but it not set any of them exactly to zero. On the other hand, the L1 norm has the effect of forcing some of the coefficients estimates to null values performing the variable selection (James *et al.*, 2013). RR-BLUP uses the same estimator as RR, but estimates the penalty parameter by REML as $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$, where $\sigma_e^2$ is the residual variance and $\sigma_\beta^2$ is the variance of the regression coefficients. Penalty function in the LASSO analysis was estimated by cross-validation considering fixed values of varying between 0.1 and 199.1 (0.1, 1.1, ..., 199.1), as described by Silva *et al.* (2011). RR-BLUP and LASSO models were implemented using, respectively, the "rrBLUP" (Endelman, 2011) and "glmnet" (Friedman *et al.*, 2010) R-packages.

Partial Least Squares Regression (PLSR) is a dimension reduction procedure that transform the predictor and then fit a least square model using the transformed variables. PLSR method will try to find the multidimensional direction in the predictor space that explains the maximum multidimensional variance in the phenotypic space. For this end, orthogonal components are built from the original predictor matrix $\mathbf{X}$ and reduces the *p+1* coefficients to the simpler problem of estimation *m+1* coefficients, where $m < p$. Let $Z_1, Z_2, ..., Z_m$ represent $M < p$ linear combinations of our original $p$ SNP effects, for some constants $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}$, where $m = 1, ..., M$. The following linear regression model can be fit considering the transformed predictors: $y = \Phi_0 + Z\Phi + \varepsilon$. PLSR is similar to the well-known principal component regression (PCR), since both methods construct a matrix of latent components as a linear transformation (James *et al.*, 2013). PLSR approach was implemented using the "pls" (Mevik *et al.*, 2007) R-package.

Tree-based methods for involves segmentation the predictor space into a number of simple regions (James *et al.*, 2013). Random Forest (RF) is a collection of classification or regression trees grown on bootstrap samples of observations using a random subset of predictors to define the best split at each node. Different variables are used at each split in different trees. The RF prediction for an observation is computed by averaging the predictions over trees for which the given observation was not used to build the tree. This model was implemented using the "RandomForest" R package (Liaw and Wiener, 2002). We used the default setting of the function, where $p/3$ predictors are considerer when building a random forest of regression trees. RF approach is a non-parametric model and make no assumptions about the distribution or any other properties of the data, which is an advantage.

#### 3.3.5.3 Bayesian Framework

Bayesian models are specified as hierarchical linear regression and, as general rule, differ in the priors adopted for the regression coefficients, while sharing the same sampling model: a Gaussian distribution with mean vector represented by a regression on $p$ markers (SNP) and a residual variance (Gianola, 2013).

For mathematical description, a notation similar to presented by de Los Campos *et al.* (2013) was used:

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

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

where $p(\mu, \beta, \sigma^2|y, \omega)$ is the posteriori density of model to unknowns $\mu, \beta, \sigma^2$ given the data ($y$) and hyper-parameters ($\omega$); $p(y|\mu, \beta, \sigma^2)$ is the likelihood of the data given the unknowns, which for continuous traits are commonly independent normal densities, with mean $X\beta$ and variance $\sigma^2$; and $p(\mu, \beta, \sigma^2|\omega)$, is the joint prior density of model unknowns, including the intercept $\mu$, which is commonly assigned a flat prior; markers effects $\beta$, which are commonly assigned IID informative priors; and the residual variance

**Table** 3.2: Summary of effect size distributions proposed to GS studies, adapted from Zhou *et al.* (2013).

| **Name** | $p(\beta\|\omega)^{**}$ | **keyword** | **Software** |
|---|---|---|---|
| t | $\beta_j \sim t(0, \upsilon, \sigma_a^2)$ | bayesA[1] | BGLR[7] |
| point-t | $\beta_j \sim \pi t(0, \upsilon, \sigma_a^2) + (1-\pi)\delta_0$ | bayesB[1] | BGLR[7] |
| point-normal | $\beta_j \sim \pi N(0, \sigma_a^2) + (1-\pi)\delta_0$ | bayesC[2], bayesVS[3] | BGLR[7], varbvs[3] |
| point-normal-mixture | $\beta_j \sim \pi_1 N(0, \sigma_a^2) + \pi_2 N(0, 0.1\sigma_a^2) + \pi_3 N(0, 0.01\sigma_a^2) + (1-\pi_1-\pi_2-\pi_3)\delta_0$ | bayesR[4] | BayesR[4] |
| normal | $\beta_j \sim N(0, \sigma_a^2)$ | bayesRR[5] | BGLR[7] |
| normal -mixture | $\beta_j \sim \pi N(0, \sigma_a^2 + \sigma_b^2) + (1-\pi)N(0, \sigma_b^2)$ | gemma[6] | GEMMA[6] |
| double exponential | $\beta_j \sim DE(0, \omega)$ | bayesLASSO[8] | BGLR[7] |

References: [1] Meuwissen *et al.* (2001); [2]Habier *et al.* (2011); [3] Carbonetto and Stephens (2012); [4] Erbe *et al.* (2012);[5] Whittaker *et al.* (2000) ; [6] Zhou *et al.* (2013); [7]Pérez and de los Campos (2013); [8]Park and Casella (2008).

[**]Notation presented by Zhou *et al.* (2013). Abbreviations: DE denotes double exponential distribution, NEG denotes normal exponential gamma distribution. In the scaled t-distribution $\upsilon$ and $\sigma_a^2$ are the degree of freedom parameter and scale parameter, respectively. In the DE distribution, $\theta$ is the scale parameter. In the NEG distribution, $\kappa$ and $\theta$ are the shape and scale parameters, respectively. Notes: 1. Some applications of these methods combine a particular effect size distribution with a random effects term, with covariance matrix K, to capture sample structure or relatedness. If $K \propto XX^T$ then this is equivalent to adding a normal distribution to the effect size distribution. The listed effect size distributions in this table do not include this additional normal component. Note 2: In some papers, keywords listed here have been used to refer to fitting techniques rather than effect size distributions

$\sigma^2$,which is commonly assigned a scaled-inverse chi-square prior with degree of freedom *d.f* and scale parameter *S*. Table 3.2 summarizes the Bayesian models considered in this research

For all models the same condition for Markov chain Monte Carlo (MCMC) convergence was considered during the computational implementation (30000 iterations and a burn-in period of 2000). "bayesVS" model is the only Bayesian approach that is not based on MCMC, but builds on an approximate Bayesian inference with variational inference (Carbonetto and Stephens, 2012). Hyper-parameters were defined following the default definitions implemented in each software. All comparison of statistical methods was performed in R language (R Core Team, 2013).

### 3.3.6 Fitting and comparing models

The SNP effects were estimated on the basis of thirteen different statistical methods and,then, they were used to compute the predicted genetic merits of an individual considering the formula: $\hat{g}_i = \sum_j^n X_{ij}\hat{\beta}_j$, where $X_{ij}$ is the specific allele of the $j^{th}$ marker on the $i^{th}$ individual; and $n$ is the total number of markers. Predictive ability was measured using the predictive accuracy and the mean squared prediction error (MSPE) following the scenarios specified in the Figure 3.1.

Predictive accuracy ($r_{gp}$) was computed via Pearson correlation between predicted ($\hat{y}_i$) and observed genetic values ($y_i$). The mean squared prediction error (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. The MSPE was also used as a measure of the predictive ability, once combines quality assessment in terms of variance and bias of predictions. The linear regression coefficient of the observed on predicted genetic value was considered to express the magnitude of inflation/deflation of the predictions relative to the response variable. To ensure impartial statistical comparisons the same independent random assignments into TRN and TST were used for each model in all the within-environment analysis (Figure 3.1). The average computational time was also considered to compare the methods. Computations were performed on a single core of an Intel Core i7-3770 3.40 GHz CPU and 8 GB of RAM memory.

Since the degree of relationships between training and validation datasets influences prediction ability, the relationship between Intermediate and Premature populations were investigated using PCA

analysis and the $F_{st}$ metric. Likewise, the effect of SNP density on the predictive accuracy was also investigated. For this end, random and guided SNPs subset were sampled. Guided SNPs subsets were sampled across windows in each *C. canephora* chromosome. The windows size in the genome ranged from 50,000 to 900,000 bp (by a increment of 100,000 bp) and ten selections of SNPs were made inside of each genomic window, considering the markers with highest MAF and call rate, as described by Spindel *et al.* (2015). In contrast, random subset of SNPs were sampled across the genome. Same number of SNPs was considered in both subset. Hence, for the Intermediate population the resulting SNP densities were: 35427, 20450, 13690, 10189, 7989, 6577, 5559, 4780 and 4240. For the Premature population the resulting SNP densities were: 40767, 21433, 13969, 8019, 6587, 5560, 4780 and 4240. Differences in SNP density across the populations is due the difference in the final number of SNPs that were mapped in the populations.

## 3.4 Results

### 3.4.1 Phenotypic Models

Figure 3.2 summarizes the phenotypic variation in both populations and locations. The phenotypic dispersion evidence that both populations have similar performances. On average, FES location was more productive than FEM and showed higher incidence of rust. Considering grain production, an important pattern was observed across the harvests in both populations: a lack of annual production stability. Evidences of differential pattern across the years, observed in the boxplot results, was confirmed by the mixed model analysis with better goodness of fit (lower AIC and BIC) when heterogeneity of residual and genetic variance were accounted (Supplementary material Table 3.4). Boxplot highlighted this cyclical production, intercalating years of high (2008 and 2010) and low production (2009 and 2011). This behavior was more evident for the production than the post-harvest trait. Lack of stability is an evidence of Genotype-by-Harvest (G×H) interaction.

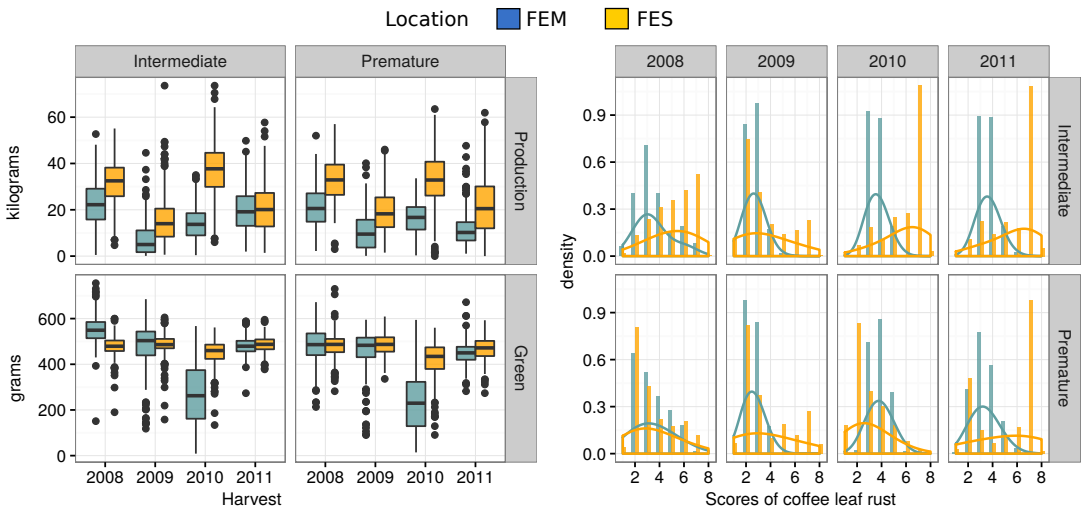


**Figure** 3.2: Boxplot of production (kilograms of mature coffee fruit in the cherries stage) and yield of green grains (grams of mature grains after processed by dry methods for removing the entire dried husk in samples of 2 kilograms of coffee fruit in the cherries stage) across locations (FEM and FES), harvests (2008, 2009, 2010 and 2011) and coffee populations (Intermediate and Premature); scale of scores of coffee leaf rust (*Hemileia vastatix*) ranging from 1 to 9, according to sporulation observation.

The heritability across traits and environments ranged from 0.56 to 0.92 (Table 3.3). Incidence of rust and yield of green grains showed the highest values of heritability (0.89 and 0.92, respectively).

**Table** 3.3: Broad-sense heritability of production (kilograms of mature coffee fruit in the cherries stage), incidence coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) across locations (FEM and FES) and coffee populations (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 |

On average, traits evaluated in FES location and in Premature population showed higher heritability than FEM location and Intermediate population.

### 3.4.2 Genotypic Data

It is our interest evaluate how genomic prediction works across locations and populations. Both genetic distance and number of SNPs shared between populations are important results. Among the factors that are under control of the breeder, the strength of genetic relationships between training and validation populations is an important component affecting prediction accuracy (de Los Campos *et al.*, 2013; Daetwyler *et al.*, 2013). Figure 3.3a shows the first two principal components from PCA analysis on the full genotypic data set. A high degree of genetic relationships between both populations was observed and there is not a clear distinction between clusters. Low value of $F_{st}$ (0.0158) is supporting this evidence of similarity between both populations.

After the SNP filtering process, a total of 45,748 and 59,332 SNPs were maintained in the Intermediate and Premature populations, respectively. Of this total, 38,106 SNPs were identified in both population(Figure 3.3b). The SNP density across the 11 *C. canephora* chromosomes showed that GBS was effective to sample markers across the *C. canephora* genome (Figure 3.3c).

### 3.4.3 Genomic Prediction

For within-environment analysis, thirteen well-established genomic prediction methods were compared for three coffee traits. Although the methods differ in assumptions of the marker effects, minimal differences were observed across the competitor models (Figure 3.4). One exception is the "fixedMLR" approach that, consistently, showed poor results.

Ignoring the "fixedMLR" results, mean values of predictive accuracy ranged from 0.17 for production trait to 0.51 for rust incidence (Supplementary material, Table 3.5). Overall, Bayesian methods presented a slight superiority compared to machine learning algorithms (0.42 versus 0.41, predictive accuracy). Predictive accuracy ranged across traits, locations and populations. Conditionalizing the results for traits and considering all the predictive models, on average, the marginal values for incidence of leaf rust and yield of green grains were higher (0.52 and 0.49, respectively) than production (0.38). These results are in accordance with the trait hereditability (Table 3.3). Likewise, but conditionalizing the predictive accuracies for locations, slight superiority was observed for the FEM in contrast to FES location (0.44 versus 0.39). For populations, on average, superior performance was observed for the Premature in contrast to the Intermediate population (0.44 versus 0.40).

Although higher predictive accuracy and lower MSPE were observed in the Bayesian approach, more computational demand was required (Figure 3.5). Among the competitor methods,"rrblup" and "gemma" showed a good balance in terms of predictive accuracy and computational demand. Considering the computation time among methods, some differences may reflect implementation issues, including language environment rather than fundamental differences between algorithms.

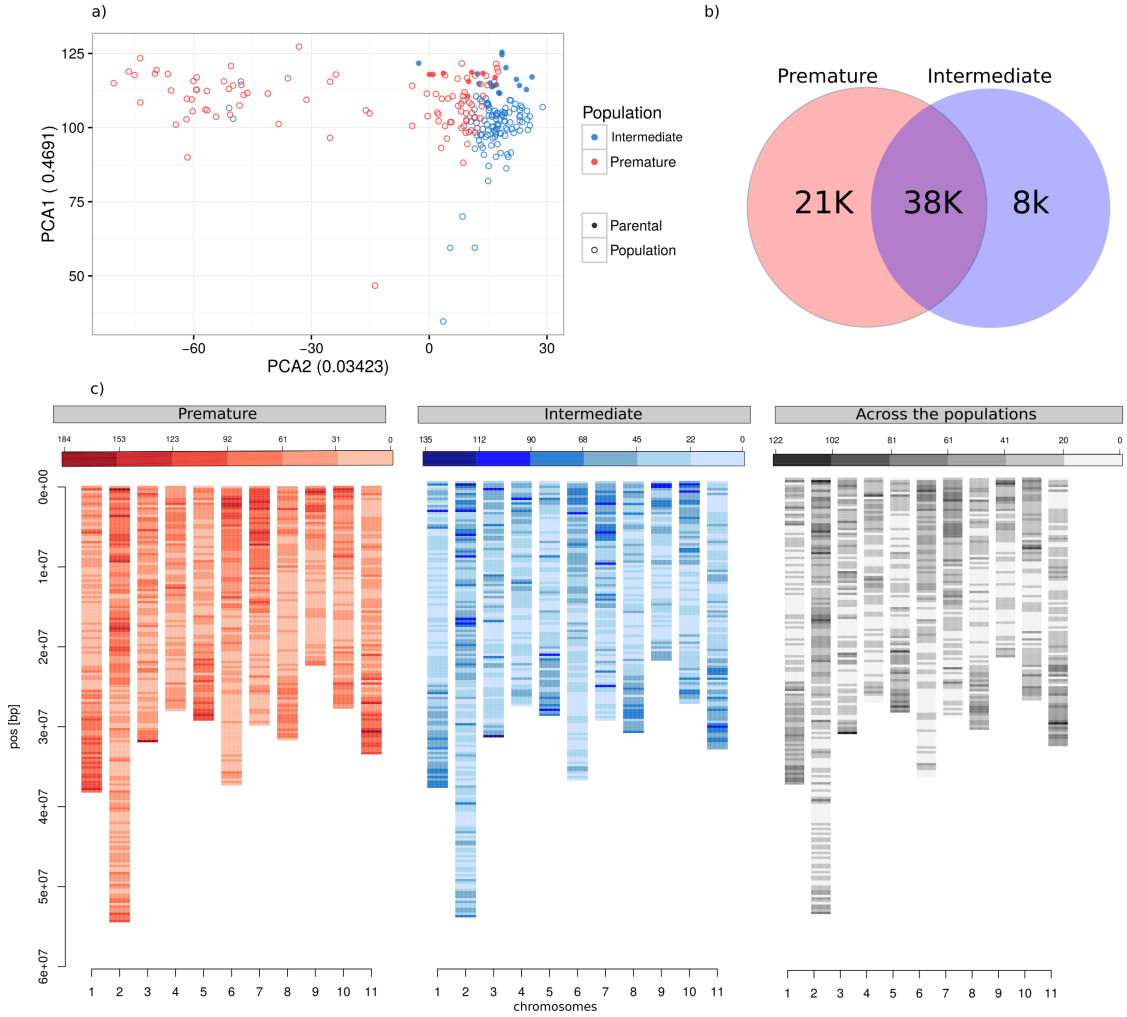


**Figure** 3.3: a) Principal component analysis (PCA) of two *Coffea canephora* breeding populations (Intermediate and Premature), where the filled points represent the parental (founders) genotypes and empty points the progenies; b) Venn diagram is showing common and differential SNPs between both population ($K = 10^3$ SNPs); c) SNP density across the 11 *C. canephora* chromosomes in 400,000 sliding windows for the Premature, Intermediate and considering common SNP across both population.

Similar results across the competitor models were also observed for the slope and MSPE metrics (Supplementary material, table 3.6). The coefficient of regression (slope) of genetic values on predicted values was calculated as a measurement of the bias of each method (Moser *et al.*, 2009). Ideally, a value of slope equal to one indicates no bias in the prediction. For all traits, similar slope values were observed across the models (Supplementary material, table 3.7); exceptions were the "fixedMLR" and "BayesVS" approaches. The MSPE was also used as a measure of prediction ability, which combines quality assessment in terms of variance and bias of predictions - inferior values are desire and indicate lower error of predictions. Once again, similar results were observed across the competitor models; "pls" "rrblup" and "fixedMLR" showed, on average, the highest error of predictions.

Given the satisfactory performance in terms of mean computation time and predictive ability, the "rrblup" approach was considered for all subsequent analysis. In order to check the impact of the SNP density on the predictive ability, SNPs were sampled: i) Randomly and ii) guided by the mAF and call rate. Regardless the approach considered to sampling the markers, it was observed a stable predictive ability across different SNP densities (Figure 3.6). A slight superior performance was observed for SNPs

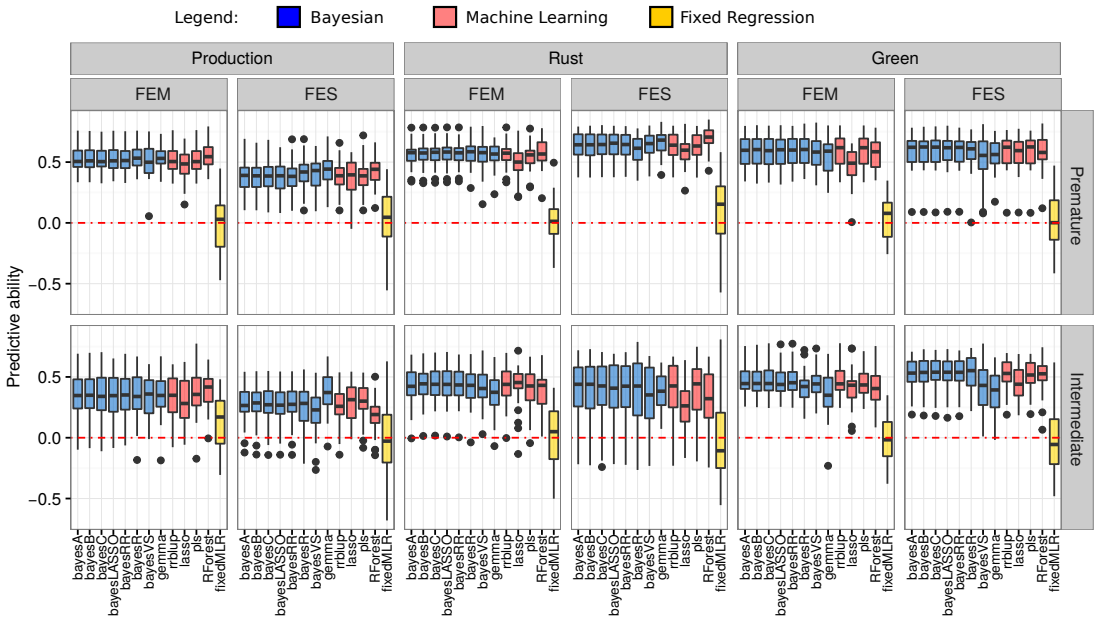


**Figure** 3.4: Comparison of the predictive ability of thirteen statistical methods applied to prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two locations (FEM and FES).

randomly sampled.

In addition to comparing statistical models, an important step in coffee breeding is to check GS performance to guide the selection over environments. As a general rule, positive values were observed for all three traits for the different scenarios (Figure 3.7). Conditionalizing the results for the same population, for all the traits, positive values of predictive accuracy were observed. Consistently, the incidence of rust and yield of green grains showed better predictive ability than production trait. These results are evidencing that GS approach has good perspectives of success for across-locations predictions. Indirectly, these positive predictions shed light on the magnitude of the genotype-by-location (G×L) interaction.

Across-populations analysis showed positive, but a lower predictive ability than across-locations. For the production trait, negative values were observed in the scenarios 5 and 7, respectively. Rust incidence and the yield of green grains traits showed promising results. Positive predictive ability is supported by the high genetic similarity between both populations, evidenced in the PCA analysis and $F_{st}$ value. An important difference was observed between models calibrated in the Intermediate and Premature populations. As a general rule, models trained at the Premature population (scenarios 6, 8, 10 and 12) showed lower predictive performance when compared to models calibrated at the Intermediate population. This difference is due to the population size, given that Intermediate population is larger than Premature. In general, accuracy of estimates of marker effects increases with sample size (de Los Campos *et al.*, 2013).

The last scenario considered across locations and populations predictions, simultaneously. Lowest predictions were observed in this condition, evidencing that across-environments analysis is a complex scenario. Interaction effects between populations and locations are important confounding sources that affect predictive performance. Negative predictions were observed for production trait.

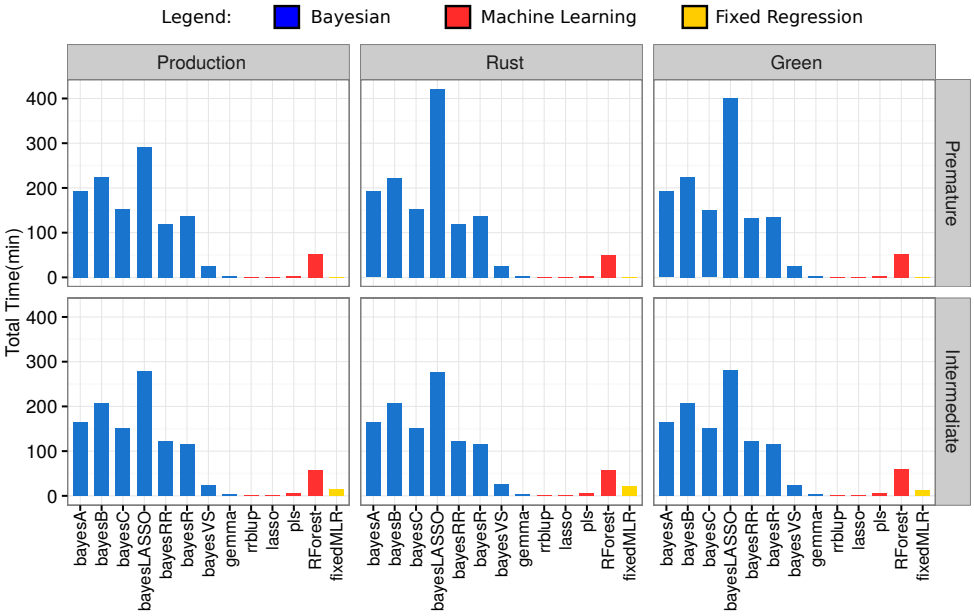


**Figure** 3.5: Computation time, in minutes, of different methods applied to prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two locations (FEM and FES).

## 3.5 Discussion

The GS potential compared with traditional phenotypic evaluations are well documented, and increasingly widely appreciated (de Los Campos *et al.*, 2013). But we believe this potential nonetheless remains under-exploited in coffee research. Some reasons can be pointed out: i) the modest number of genomic resources available; ii) difficulty in maintaining field experimentation; iii) target traits expressed later and iv) long life cycle. This breeding scenario is not restricted to coffee but represents the majority of perennial crops. Our aim here is to provide a broad discussion related to GS implementation. In particular, we designed this research aiming to check the GS impact in conventional coffee breeding schemes. At its core, it is important to consider that any GS investigation begins with the obtaining of good phenotypic metrics, followed by high throughput genotyping and use of a statistical model to combine these information to perform predictions.

Coffee experimental design are represented, in their majority, for longitudinal data evaluated in multiple environments and harvests. Experiments of this nature are collectively referred to as Multi-Environments Trials (MET) (Smith *et al.*, 2005). A large number of statistical models have been developed to address MET analysis. The use of conventional fixed models (ANOVA models) suggests the violation of basic assumptions, *e.g.*, homogeneity and independence of variances (Smith *et al.*, 2005). Ignore these source of variation can introduce bias on the estimation of genetic values and, eventually, affects predictive performance. Given its importance, in the mixed model, we have included appropriate (co)variance structures for modeling heterogeneity and correlation of genetic effects and non-genetic residual effects(Smith *et al.*, 2005; Pastina *et al.*, 2012; Balsalobre *et al.*, 2016). This approach has several advantages that are not easy to handle in traditional analysis (Malosetti *et al.*, 2014). The flexibility to fit residual and genetic variances conditional to harvest showed a better goodness of fit than traditional ANOVA.

High-throughput genotyping was boosted by the rapid progress of next-generation DNA sequencing (NGS). Genotyping-by-Sequencing (GBS) is a product of this advance (Elshire *et al.*, 2011).

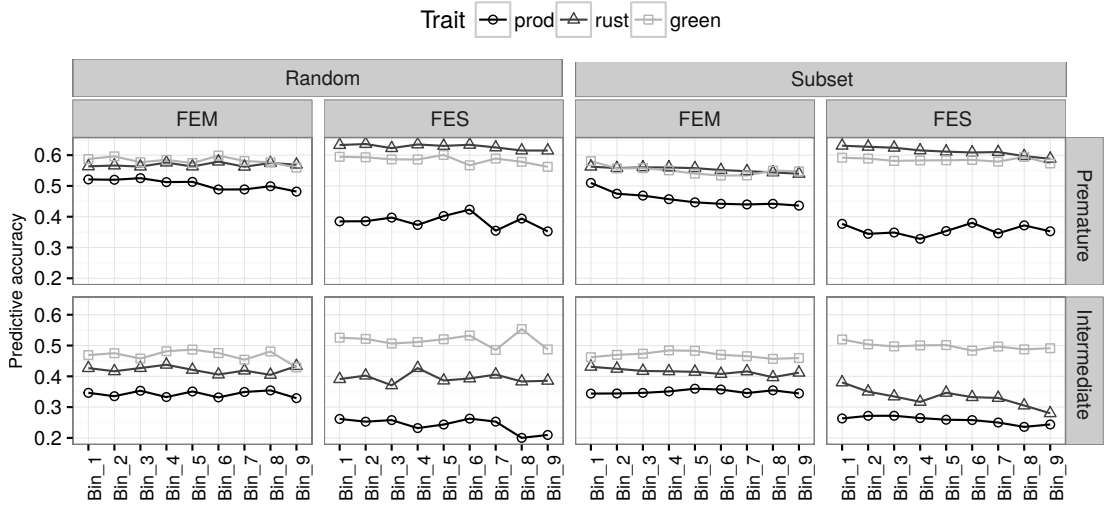


**Figure** 3.6: Mean predictive ability of cross-validation for prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two locations (FEM and FES), using 10 selections of SNP subsets either distributed evenly throughout the genome (right column) or chosen at random (left column). From Bin1 to Bin9, in the Intermediate population, the SNP density are: 35427, 20450, 13690, 10189, 7989, 6577, 5559, 4780 and 4240. For the Premature population, the SNP density are: 40767, 21433, 13969, 8019, 6587, 5560, 4780 and 4240

A total of 45748 and 59332 filtered SNPs were identified in the Intermediate and Premature coffee populations, respectively. This SNP density is superior than a previous studies reported in *Coffea arabica* using similar approach (DaRT methodology) (Moncada *et al.*, 2015). Difference in polymorphism levels between both species are consistent with the evolutionary history, since *C. canephora* possesses a high genetic diversity due its origin, reproduction method and dissemination (Ferrão *et al.*, 2015). In a further investigation, we considered GBS results to investigate the genetic differentiation of both populations. An important point to the success of GS implementation is the genetic closeness between the reference population (TRN) and the breeding population. It is expects a reduction in predictive ability when the degree of relationship decreases (De Roos *et al.*, 2009; de Los Campos *et al.*, 2013; Daetwyler *et al.*, 2013). PCA analysis and $F_{st}$ metric showed a little genetic differentiation between both populations. An $F_{st}$ value of 0.0158 was observed. As rule of thumb, some authors have been considering that $F_{st} < 0.05$ indicates little genetic differentiation (Hartl *et al.*, 1997). For our populations, lower genetic variation is expected since the population's founders were genotypes selected during the breeding program (accessions within a coffee germplasm collection), which share common agronomic features. This similarity was supported by the PCA analysis.

For predictive analysis, we compared thirteen predictive models on within-environment predictions. It seems to be consensual that predictive performance is dependent on biological and technical factors, including population size, genetic architecture and relation among the training and validation data set (de Los Campos *et al.*, 2009; Daetwyler *et al.*, 2013). Considering that models were evaluated in all environments under the same technical conditions, we were expecting a possible dependence between predictive performance and trait, conditional on the genetic architecture. For example, "bayesRR" approach considers the marker effects as sampled from a normal distribution with fixed variance (Meuwissen *et al.*, 2001); hence, as a practical consequence, effects are shrinking to the same degree assuming our believes that the trait is controlled by many loci with small effects, in reference to the infinitesimal model (Fisher, 1919). In contrast, "bayesB" makes assumptions that most loci have no effect on the trait and therefore more markers are left out of the prediction model; so, the underlying biological phenomenon in

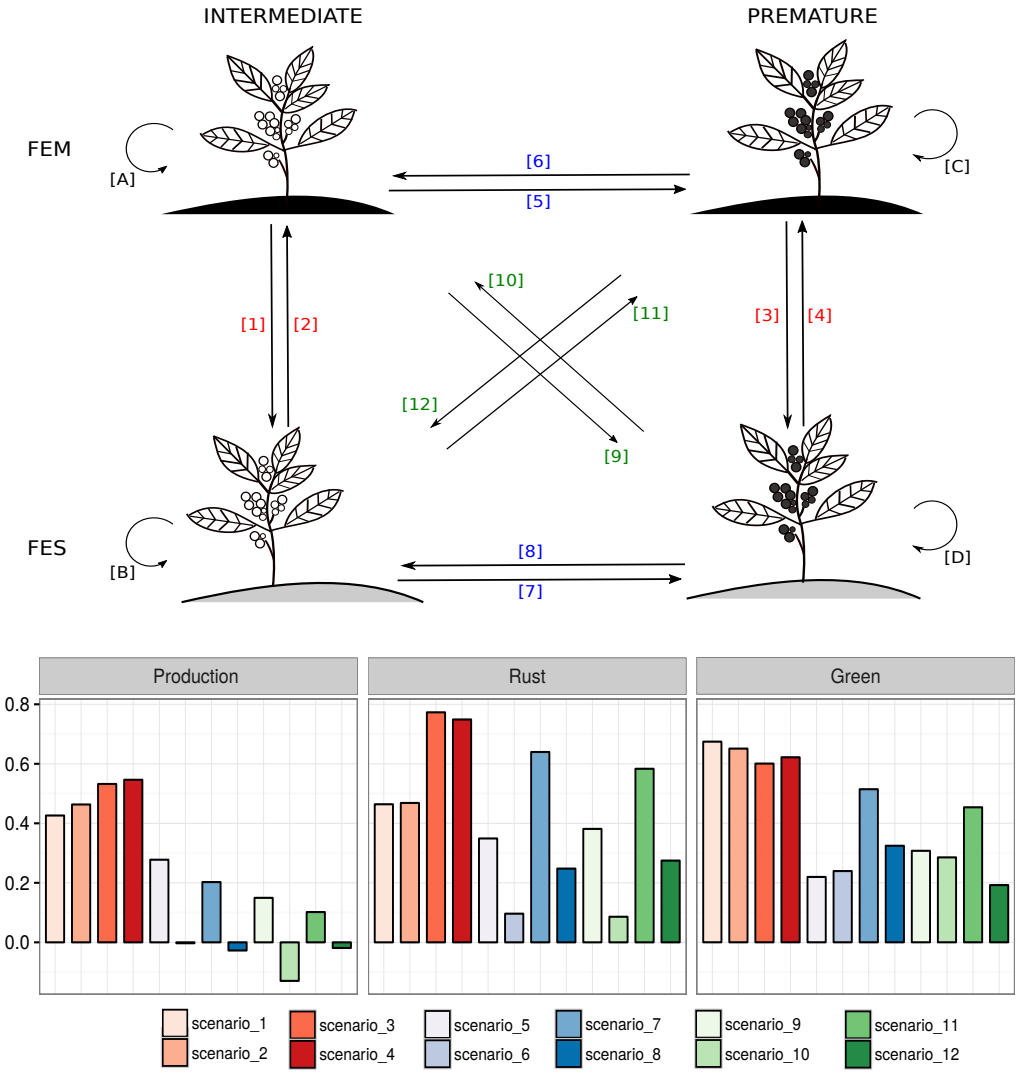


**Figure** 3.7: Across-environments predictions. Here, by environment we mean a specific combination between location (FEM and FES) and population (Intermediate and Premature). Scenarios 1, 2, 3 and 4 represented GS performance across-locations. Scenarios 5, 6, 7 and 8 represented GS performance across-populations. Scenarios 9, 10, 11 and 12 represented across-environments predictions. Direction of the arrows represented changes on training and testing data sets. Predictive ability were measured across production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage).

this case is a trait controlled by relatively few loci, whose effects that vary in size (Meuwissen *et al.*, 2001). Although conceptually different, we observed a minimal difference in terms of predictive accuracy across the models, evidencing an apparent divergence of our empirical results with simulation studies (Meuwissen *et al.*, 2001; Coster *et al.*, 2010).

In recent years, a number of empirical evaluations methods have been published comparing predictive models (Moser *et al.*, 2009; Heffner *et al.*, 2011; Riedelsheimer *et al.*, 2012; Resende *et al.*, 2012; Daetwyler *et al.*, 2013; Wang *et al.*, 2015; Júnior *et al.*, 2016). As a general rule, these studies have supported our finds, indicating similar predictive performance across competitor models. At this point, biological and statistical hypotheses have been proposed. The high discrepancy between number of observation and parameters can restrict the process of statistical learning resulting in similar predictive performances among methods (Gianola, 2013; de Los Campos *et al.*, 2013). Considering

the observed similarity between Bayesian models, Tempelman (2015) pointed out other general issues concerning on the hyperparameters specification, number and diagnostic of MCMC chains and, also, problems related to data dimensionality. Following a biological perspective, similarity across methods can be associated with the complex nature of the traits. For real data, the distribution of QTLs effects for most traits is perhaps less extremes than suggested (de Los Campos *et al.*, 2013; Daetwyler *et al.*, 2013). In maize, Riedelsheimer *et al.* (2012) associated minor differences among the models due to the high level of linkage disequilibrium (LD). In this scenario, predictive accuracies are quite similar irrespective whether the effect of large QTL are precisely captured (as in the case of "lasso" and "bayesB" algorithms, for example) or spread over a larger region (as in the case of "rrblup" method).

The only exception in terms of predictive accuracy was observed for the fixed regression method ("fixedMLR" algorithm), an approach commonly used for genome-wide association analyses (GWAS). Fixed regression has been useful to detect associations, but explaining only a small fraction of the genetic variance of quantitative traits (Manolio *et al.*, 2009). In contrast, methods that simultaneously fit all markers as random effects are able to account for most of the genetic variance and, for this reason, they are more appropriate to predictive purposes (Meuwissen *et al.*, 2001; Moser *et al.*, 2009).

Although we observed a slight difference in terms of predictive ability, computational demand significantly differ across the methods. Machine learning algorithms showed less computational time than Bayesian analysis, in agreement with previous studies comparing predictive models (Moser *et al.*, 2009; Heslot *et al.*, 2012; Neves *et al.*, 2012). Computing time is important, particularly for implementation in practice which requires frequent re-estimation of genetic merits (Moser *et al.*, 2009). At this point, it seems useful to highlight the "rrblup" and "gemma" approaches. Judged by their overall performance across traits and computational requirements, both seem to be particularly appealing for practice application in plant breeding.

The "rr-blup" approach was one of the first methods proposed for genomic selection and is equivalent to best linear unbiased prediction (BLUP), in the context of mixed models (Whittaker *et al.*, 2000; Endelman, 2011). Among the additive models, RR-BLUP method has been widely used in animal and plant breeding due its straightforward implementation using existing mixed models software, relative simplicity, good performance and limited computing time (de Los Campos *et al.*, 2013). On the other hand, the "gemma" approach is a method that belong to the Bayesian class. Originally proposed by Zhou *et al.* (2013), it is a hybrid between assumptions assumed by linear mixed models (all variants have at least a small effect) and sparce linear regression methods (some portion of variants have an additional effect). This almost diametrically opposed assumptions are addressed assuming that effects come from a mixture of two normal distributions. Commonly used in GWAS analysis, we are reinforcing their potential to be applied in predictive analysis. Implemented in GEMMA software, the methodology addresses other two important practical issues: i) emphasizes the benefits of estimating the hyperparameters from the data, rather than setting the pre-specified values; and ii) provide a computational algorithm faster than traditional Bayesian methods (Zhou and Stephens, 2012; Zhou *et al.*, 2013; Zhou and Stephens, 2014).

Regarding that resources need to be allocated to genotyping, a further inspection considered the number of markers necessary to performing GS. In our scenario, for all three traits and both validation schemes (Random and Subset), thousand of markers were sufficient to maintain the predictive performance. This information is important for future practical implementations. Empirical studies have evaluated the effects of marker density on prediction accuracy and reported similar results (Vazquez *et al.*, 2010; Spindel *et al.*, 2015). As a general rule, prediction accuracy reaches a plateau and does not increase beyond certain marker density. According to de Los Campos *et al.* (2013) the level at which this plateau takes place depends, mainly, of the span of LD in the genome and sample size. In both populations, we expected large LD blocks since they were originated from only one cycle of recombination.

Experiments in perennial crops consider evaluations in multiple environmental conditions. The central purpose is measure the performance of breeding stocks across a range of conditions that cultivars might be exposed and provide information about the adaptability of genotypes to specific environments or to sets of environments (Malosetti *et al.*, 2014, 2016). In GS context, the possibility to predict phenotypic performance across environments is certainly a relevant question. Further, the similarity between both coffee populations - supported by the PCA analysis- shed light about the possibility to perform predictions across-populations (Intermediate and Premature).

Combining genotypes from different breeding populations into one training set tends to be advantageous, since markers effects could be estimated from a larger number of observations. Methods and models have been discussed in this direction by De Roos *et al.* (2009), Schulz-Streeck *et al.* (2012) and Hayes *et al.* (2009). However, consider such models in plant breeding implies multiple populations or environments delineated in field, phenotyped for different traits and genotyped. In practical terms, this means increases in costs and time. A simpler scenario addressed in this research considers whether a unique training population could be used to calibrate a predictive model and, as consequence, the estimated markers effects used to predict phenotypic performances in other conditions (locations or populations). However, for this purpose, molecular markers just can be used if their estimated effects remain accurate over the environments. In order to check the reliability of genomic prediction in this scenario, we have tested their potential across different combinations of location and population.

In contrast to within-environments predictions, analysis across locations, populations and environments resulted, on average, in lower predictive accuracies in coffee. For across-locations analysis, this results are expected due the Genotype-by-location (G×L) interaction, even when both sites are located within one breeding zone. The G×L occurs because the capture and conversion abilities of a plant are determined by its particular ensemble of genes, which are express conditionally to the amount and quality of inputs received in each environmental condition (Malosetti *et al.*, 2014). This differential expression is captured by the marker effects computed in each location and, as consequence, influences in the across-location predictions. Decaying in accuracy for across-locations predictions are supported by studies in forest (Beaulieu *et al.*, 2014; Gamal El-Dien *et al.*, 2015), cassava (Ly *et al.*, 2013) and maize (Windhausen *et al.*, 2012). In terms of across-populations predictions, lower accuracy values are based in quantitative genetic theory, which supports that allele substitution effects may vary between populations due to, for example, differences in allele frequency. In this scenario, it is reasonable expect that marker effects vary between populations due to differences in marker–QTL linkage disequilibrium (LD) patterns (Asoro *et al.*, 2011; Windhausen *et al.*, 2012; Lehermeier *et al.*, 2015). Similar results are discussed by Neves *et al.* (2012), in mice populations.

The magnitude of correlations between the predicted and observed performance in validation across locations and populations open an important perspective to implement marker-assisted selection in conventional schemes of recurrent selection. Traditionally, one cycle of phenotypic recurrent selection in *C. canephora* consists of: i) Development of progenies from a base population; ii) phenotypic evaluation of the progenies in multiple environments and harvests; and iii) selection and recombination of the best individuals to form a new base population. Given the long juvenile period, on average, 5-6 years are necessary for each cycle in coffee. Additionally, evaluate and maintain multiple trials on field is expensive and laborious. The expected increase in genetic gain is assumed to come from the acceleration of the breeding cycles (more cycles per unit time), and from higher selection intensity, through genotypic evaluation of a larger number of candidates (Grenier *et al.*, 2015). To this end, GS would be implemented in the second and third stage of a conventional recurrent selection program, considering prediction and selection during the seedling phase inside of greenhouses. Rapid-cycle of recurrent selection has potential to accelerate the increase of the frequency of favorable alleles in the population and substantially reducing both monetary and time costs associated with phenotyping (Windhausen *et al.*, 2012). In a modern

breeding scheme, phenotypic trials in multiple environments could be considered in advanced phases (e.g., third recurrent cycle), in order to re-estimate the marker effects or recommend promising materials to compose a new clonal variety of coffee.

Finally, one of the main difficulties in coffee studies is the lack of information about genetic architecture of complex traits (Tran *et al.*, 2016). It seems clear that understanding the underlying genetic basis of a phenotypic trait and its variational properties can aid the selection in plant breeding. Intuitively, the seek for genetic variants that present common and specific effects across environments and/or populations can guide the GS implementation. The effort to identify relevant genetic factors and, as consequence, regard this information for predictive purposes, in such way, approximate GS and genome-wide association studies (GWAS). Recent studies have shown that GS model combined with GWAS analysis can increase predictive results (Spindel *et al.*, 2015, 2016). As pointed out by Spindel *et al.* (2015), "genetic architecture must also be taken into account when considering the implementation of genomic selection".

## 3.6 Conclusion

Genomic selection depends on the breeding scenario that the breeder is attempting to address. In this research we investigated a hypothetical situation where GS was considered to predict genetic merits in different locations and populations for different traits in coffee. In addition, we relaxed the usual assumption of marker effects drawn from a normal distribution seeking for a possible association between model and trait, conditional to the genetic architecture. Although each model is based on particular genetic and statistical assumptions, our main finding diverges of previous simulation studies reported in the literature. Minimal differences in terms of predictive ability were observed across the competitor models for different traits. However, we observed significant differences in terms of computational demand, supporting our intention to consider simpler and fast algorithms for future studies with similar purposes. Finally, positive predictive accuracies across locations and populations open an important perspective to implement marker-assisted selection in conventional schemes of recurrent selection 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. As a future perspective, we reinforce the relevance to advance in studies about genetic architecture of complex traits in coffee.

## Acknowledgments and grant support

This work is partially supported by FAPESP/CAPES (Sao Paulo Research Foundation), grants 2014/20389-2 and 2016/05127-7 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. 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 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 installed the experimental design and collected the phenotypic data. L.F.V.F performed the DNA extraction. L.F.V.F, M.S and A.A.F.G performed the genomic prediction analysis. L.F.V.F wrote the paper.

### 3.7 Supplementary material

**Table** 3.4: Goodness of fit for the genetic and residual matrix, factored by harvest. Bold numbers represent the smallest AIC and BIC values, indicating the best fitted phenotypic model.

| Site | $G_h$ | $R_h$ | Production | | Rust | | Green | |
|---|---|---|---|---|---|---|---|---|
| | | | AIC | BIC | AIC | BIC | AIC | BIC |
| **Intermediate Population** | | | | | | | | |
| **FEM** | **Ident** | **Ident** | 9709.4 | 9751.4 | 9709.4 | 9751.4 | 10816.0 | 10849.7 |
| | **Diag** | **Ident** | 9679.8 | 9737.5 | 9679.8 | 9737.5 | 10746.8 | 10795.0 |
| | **CompSym** | **Ident** | 9695.6 | 9742.7 | 9695.6 | 9742.7 | 10805.6 | 10844.1 |
| | **Uns** | **Ident** | 9639.7 | 9728.8 | 9639.7 | 9728.8 | 10750.6 | 10827.7 |
| | **Ident** | **Diag** | 9648.9 | 9706.6 | 9648.9 | 9706.6 | 10511.5 | 10559.7 |
| | **Diag** | **Diag** | 9627.0 | 9700.4 | 9627.0 | 9700.4 | 10516.1 | 10578.7 |
| | **CompSym** | **Diag** | 9632.6 | 9695.5 | 9632.6 | 9695.5 | **10484.3** | **10537.2** |
| | **Uns** | **Diag** | **9587.0** | **9691.9** | **9587.0** | **9691.9** | 10492.8 | 10584.2 |
| **FES** | **Ident** | **Ident** | 10296.6 | 10338.5 | 5218.2 | 5260.2 | 9661.9 | 9695.7 |
| | **Diag** | **Ident** | 10281.3 | 10339.0 | 5220.7 | 5278.5 | 9646.1 | 9694.4 |
| | **CompSym** | **Ident** | 10224.0 | 10271.2 | 5094.0 | 5141.3 | 9570.2 | 9608.8 |
| | **Uns** | **Ident** | 10141.2 | 10230.4 | 5077.4 | 5166.6 | 9536.6 | 9613.8 |
| | **Ident** | **Diag** | 10275.7 | 10333.5 | 5190.8 | 5248.5 | 9595.7 | 9643.9 |
| | **Diag** | **Diag** | 10256.3 | 10329.7 | 5192.3 | 5265.8 | 9595.9 | 9658.6 |
| | **CompSym** | **Diag** | 10203.6 | 10266.5 | 5064.0 | **5127.1** | 9498.2 | **9551.3** |
| | **Uns** | **Diag** | **10114.3** | **10219.3** | **5048.4** | 5153.4 | **9483.9** | 9575.6 |
| **Premature Population** | | | | | | | | |
| **FEM** | **Ident** | **Ident** | 8279.9 | 8320.7 | 3318.7 | 3359.6 | 9470.3 | 9503.1 |
| | **Diag** | **Ident** | 8284.3 | 8340.5 | 3195.2 | 3251.5 | 9413.4 | 9460.2 |
| | **CompSym** | **Ident** | 8257.3 | 8303.3 | 3319.2 | 3365.2 | 9447.9 | 9485.4 |
| | **Uns** | **Ident** | 8259.6 | 8346.5 | 3185.6 | 3272.6 | 9389.3 | 9464.3 |
| | **Ident** | **Diag** | 8239.4 | 8295.6 | 3236.0 | 3292.3 | 9312.6 | 9359.5 |
| | **Diag** | **Diag** | 8243.8 | 8315.3 | 3134.5 | **3206.1** | 9309.2 | 9370.2 |
| | **CompSym** | **Diag** | **8215.5** | **8276.8** | 3236.2 | 3297.6 | **9267.4** | **9318.9** |
| | **Uns** | **Diag** | 8219.0 | 8321.2 | **3124.0** | 3226.3 | 9268.1 | 9357.1 |
| **FES** | **Ident** | **Ident** | 8661.8 | 8702.5 | 4509.7 | 4550.3 | 8338.5 | 8371.0 |
| | **Diag** | **Ident** | 8641.0 | 8696.9 | 4484.0 | 4539.9 | 8297.5 | 8344.0 |
| | **CompSym** | **Ident** | 8586.9 | 8632.6 | 4360.4 | 4406.1 | 8186.1 | 8223.2 |
| | **Uns** | **Ident** | 8525.2 | **8611.5** | 4305.1 | 4391.4 | 8110.6 | **8185.0** |
| | **Ident** | **Diag** | 8644.8 | 8700.7 | 4437.3 | 4493.1 | 8324.4 | 8370.9 |
| | **Diag** | **Diag** | 8629.0 | 8700.1 | 4414.8 | 4485.9 | 8291.3 | 8351.7 |
| | **CompSym** | **Diag** | 8569.6 | 8630.5 | 4286.3 | 4347.3 | 8154.3 | 8205.4 |
| | **Uns** | **Diag** | **8513.2** | 8614.8 | **4233.4** | **4334.9** | **8103.0** | 8191.4 |

Ident:Identical variation; Diag: Heterogeneous variations; CompSym: compound symmetry with heterogeneous variance; Uns: unstructured model.
Bold numbers represent the smallest AIC and BIC values, indicating the best fitted phenotypic model.

**Table** 3.5: Mean predictive accuracy of thirteen methods applied to prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two sites (FEM and FES). Predictive abilities were assessed using a Replicated Training-Testing evaluation. In each replication, 80% of the individuals were assigned randomly for training data set, while the remaining 20% were assigned for testing data set (TST).

| **Intermediate Population** | | | | | | |
|---|---|---|---|---|---|---|
| | **FEM** | | | **FES** | | |
| **Model** | **Production** | **Rust** | **Green** | **Production** | **Rust** | **Green** |
| **bayesA** | 0.3591 | 0.4241 | 0.4634 | 0.2686 | 0.3938 | 0.5252 |
| **bayesB** | 0.3539 | 0.4299 | 0.4656 | 0.2687 | 0.3936 | 0.5250 |
| **bayesC** | 0.3531 | 0.4332 | 0.4685 | 0.2673 | 0.3936 | 0.5290 |
| **bayesLASSO** | 0.3509 | 0.4350 | 0.4663 | 0.2623 | 0.3903 | 0.5280 |
| **bayesRR** | 0.3568 | 0.4325 | 0.4674 | 0.2655 | 0.3962 | 0.5283 |
| **bayesR** | 0.3465 | 0.4244 | 0.4119 | 0.2493 | 0.3856 | 0.5356 |
| **bayesVS** | 0.3470 | 0.4029 | 0.4493 | 0.2165 | 0.3499 | 0.4151 |
| **gemma** | 0.3533 | 0.3565 | 0.3601 | 0.3712 | 0.3814 | 0.3754 |
| **rrblup** | 0.3442 | 0.4280 | 0.4653 | 0.2562 | 0.3911 | 0.5222 |
| **lasso** | 0.3013 | 0.4308 | 0.4160 | 0.2937 | 0.2445 | 0.4454 |
| **pls** | 0.3701 | 0.4047 | 0.4522 | 0.2891 | 0.3926 | 0.5151 |
| **RForest** | 0.3844 | 0.3970 | 0.4073 | 0.1763 | 0.3399 | 0.5104 |
| **fixedMLR** | 0.1219 | 0.0298 | -0.0193 | 0.0056 | -0.0175 | -0.0188 |
| **Premature Population** | | | | | | |
| **bayesA** | 0.5253 | 0.5637 | 0.5792 | 0.3796 | 0.6348 | 0.5980 |
| **bayesB** | 0.5248 | 0.5661 | 0.5807 | 0.3787 | 0.6341 | 0.5980 |
| **bayesC** | 0.5245 | 0.5663 | 0.5802 | 0.3793 | 0.6357 | 0.6003 |
| **bayesLASSO** | 0.5229 | 0.5706 | 0.5779 | 0.3723 | 0.6383 | 0.6012 |
| **bayesRR** | 0.5252 | 0.5687 | 0.5804 | 0.3792 | 0.6352 | 0.5993 |
| **bayesR** | 0.5371 | 0.5777 | 0.5895 | 0.4050 | 0.5980 | 0.5863 |
| **bayesVS** | 0.5067 | 0.5625 | 0.5754 | 0.3973 | 0.6371 | 0.5264 |
| **gemma** | 0.5290 | 0.5562 | 0.5563 | 0.4317 | 0.6556 | 0.5568 |
| **rrblup** | 0.5181 | 0.5649 | 0.5872 | 0.3822 | 0.6313 | 0.5923 |
| **lasso** | 0.4693 | 0.4953 | 0.4679 | 0.3679 | 0.5831 | 0.5717 |
| **pls** | 0.5184 | 0.5496 | 0.5870 | 0.3867 | 0.6305 | 0.5908 |
| **RForest** | 0.5537 | 0.5770 | 0.5646 | 0.4309 | 0.6907 | 0.5849 |
| **fixedMLR** | 0.0050 | 0.0025 | 0.0428 | 0.0321 | 0.0861 | 0.0193 |

**Table** 3.6: Mean value of the mean squared prediction error (MSPE) of thirteen statistical methods applied to prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two sites (FEM and FES). MSPE were assessed using a Replicated Training-Testing evaluation. In each replication, 80% of the individuals were assigned randomly for training data set, while the remaining 20% were assigned for testing data set (TST).

| **Intermediate Population** | | | | | | |
|---|---|---|---|---|---|---|
| | **FEM** | | | **FES** | | |
| **Model** | **Production** | **Rust** | **Green** | **Production** | **Rust** | **Green** |
| **bayesA** | 0.870 | 0.900 | 0.755 | 0.877 | 0.821 | 0.713 |
| **bayesB** | 0.871 | 0.894 | 0.755 | 0.873 | 0.827 | 0.715 |
| **bayesC** | 0.874 | 0.894 | 0.753 | 0.872 | 0.833 | 0.712 |
| **bayesLASSO** | 0.876 | 0.897 | 0.757 | 0.867 | 0.839 | 0.723 |
| **bayesRR** | 0.869 | 0.894 | 0.754 | 0.868 | 0.832 | 0.712 |
| **bayesR** | 0.890 | 0.920 | 0.819 | 0.888 | 0.833 | 0.734 |
| **bayesVS** | 0.928 | 0.964 | 0.848 | 0.925 | 0.925 | 0.868 |
| **gemma** | 0.878 | 0.886 | 0.891 | 0.857 | 0.850 | 0.879 |
| **rrblup** | 1.045 | 1.045 | 1.045 | 0.910 | 0.910 | 0.910 |
| **lasso** | 0.951 | 0.884 | 0.835 | 0.847 | 0.997 | 0.769 |
| **pls** | 3.088 | 3.758 | 2.881 | 1.014 | 1.280 | 1.051 |
| **RForest** | 0.851 | 0.933 | 0.819 | 0.913 | 0.874 | 0.777 |
| **fixedMLR** | 762.467 | 108.657 | 50.575 | $10^4$ | $10^4$ | $10^3$ |
| **Premature Population** | | | | | | |
| **bayesA** | 0.745 | 0.666 | 0.656 | 0.875 | 0.593 | 0.653 |
| **bayesB** | 0.746 | 0.665 | 0.657 | 0.877 | 0.593 | 0.654 |
| **bayesC** | 0.748 | 0.664 | 0.658 | 0.877 | 0.592 | 0.654 |
| **bayesLASSO** | 0.761 | 0.665 | 0.672 | 0.886 | 0.594 | 0.669 |
| **bayesRR** | 0.748 | 0.661 | 0.661 | 0.877 | 0.592 | 0.656 |
| **bayesR** | 0.765 | 0.684 | 0.703 | 0.878 | 0.695 | 0.695 |
| **bayesVS** | 0.824 | 0.711 | 0.741 | 0.933 | 0.633 | 0.787 |
| **gemma** | 0.738 | 0.677 | 0.681 | 0.846 | 0.568 | 0.698 |
| **rrblup** | 0.767 | 1.585 | 0.932 | 1.496 | 1.952 | 1.516 |
| **lasso** | 0.800 | 0.752 | 0.760 | 0.928 | 0.679 | 0.826 |
| **pls** | 0.765 | 1.676 | 0.933 | 1.640 | 1.976 | 1.563 |
| **RForest** | 0.739 | 0.669 | 0.706 | 0.852 | 0.536 | 0.679 |
| **fixedMLR** | $10^4$ | $10^4$ | $10^4$ | $10^4$ | 4.733 | $10^4$ |

**Table** 3.7: Mean value of the linear regression coefficient of the observed on predicted genetic value (slope) of thirteen statistical methods applied to prediction of production (kilograms of mature coffee fruit in the cherries stage), incidence of coffee leaf rust (scaling score) and yield of green grains (grams in samples of 2 kilograms of coffee fruit in the cherries stage) in two *Coffea canephora* populations (Intermediate and Premature) evaluated in two sites (FEM and FES). Slopes were assessed using a Replicated Training-Testing evaluation. In each replication, 80% of the individuals were assigned randomly for training data set, while the remaining 20% were assigned for testing data set (TST).

| **Intermediate Population** | | | | | | |
|---|---|---|---|---|---|---|
| | **FEM** | | | **FES** | | |
| **Model** | **Production** | **Rust** | **Green** | **Production** | **Rust** | **Green** |
| **bayesA** | 1.003 | 1.002 | 1.001 | 0.740 | 1.056 | 1.183 |
| **bayesB** | 1.044 | 1.054 | 1.070 | 0.809 | 1.121 | 1.214 |
| **bayesC** | 1.114 | 1.102 | 1.114 | 0.865 | 1.171 | 1.248 |
| **bayesLASSO** | 1.197 | 1.240 | 1.230 | 0.903 | 1.275 | 1.396 |
| **bayesRR** | 1.118 | 1.118 | 1.127 | 0.857 | 1.220 | 1.253 |
| **bayesR** | 1.658 | 1.497 | 1.425 | 1.409 | 1.426 | 1.542 |
| **bayesVS** | 3.514 | 2.132 | 2.918 | 3.302 | 2.697 | 1.869 |
| **gemma** | 1.109 | 1.093 | 1.176 | 1.184 | 1.164 | 1.184 |
| **rrblup** | 1.370 | 1.370 | 1.370 | 1.239 | 1.239 | 1.239 |
| **lasso** | 0.758 | 1.091 | 1.021 | 0.724 | 0.667 | 1.029 |
| **pls** | 0.779 | 0.780 | 0.801 | 0.613 | 0.835 | 0.963 |
| **RForest** | 1.334 | 1.397 | 1.525 | 0.719 | 1.264 | 1.859 |
| **fixedMLR** | 0.070 | 0.027 | -0.002 | 0.002 | 0.006 | -0.017 |
| **Premature Population** | | | | | | |
| **bayesA** | 1.170 | 0.987 | 0.851 | 1.093 | 1.033 | 1.168 |
| **bayesB** | 1.189 | 1.014 | 0.911 | 1.119 | 1.048 | 1.183 |
| **bayesC** | 1.210 | 1.039 | 0.938 | 1.154 | 1.071 | 1.215 |
| **bayesLASSO** | 1.366 | 1.143 | 1.202 | 1.316 | 1.168 | 1.364 |
| **bayesRR** | 1.217 | 1.044 | 0.954 | 1.173 | 1.070 | 1.221 |
| **bayesR** | 1.584 | 1.456 | 1.881 | 1.813 | 1.315 | 1.423 |
| **bayesVS** | 2.027 | 1.495 | 2.375 | 2.510 | 1.454 | 1.580 |
| **gemma** | 1.110 | 0.983 | 0.893 | 1.105 | 1.042 | 1.002 |
| **rrblup** | 1.042 | 1.007 | 0.847 | 1.093 | 0.993 | 1.079 |
| **lasso** | 1.095 | 0.979 | 0.455 | 0.916 | 1.038 | 1.135 |
| **pls** | 1.026 | 0.865 | 0.694 | 0.941 | 0.930 | 1.022 |
| **RForest** | 1.493 | 1.211 | 1.186 | 1.376 | 1.264 | 1.306 |
| **fixedMLR** | -0.002 | -0.002 | 0.000 | 0.000 | 0.014 | 0.001 |

# 4 CONCLUSION

Simulation and empirical results have shown that genomic prediction presents sufficient accuracy to help success in breeding programs. Although some crops have benefited from this methodology, studies in the genus *Coffea* are still modest. The main objective in this research was discuss aspects related to statistical modeling in order to enable a more comprehensive understanding of what makes a robust and accurate prediction model. Additionally, it was explored new possibilities introduced through genomic selection to accelerate coffee breeding programs. Aspects of statistical modelling were discussed in Chapters 1 and 2, considering two different approaches: Mixed model and multilocus association models.

In addition to statistical modeling, Chapter 1 and 2 addressed questions that underlie a coffee breeding program. In Chapter 1, for a given population, both locals were jointly modeled in order to answer questions related to the importance of interaction modelling, compare phenotypic and genomic models and investigate the potential of the Genotyping-by-Sequencing (GBS) in coffee studies. On the other hand, Chapter 2 addressed a hypothetical situation where GS was considered to predict genetic merits in different environments and populations.

In terms of practical implementation, the use of mixed model theory (Chapter 1) presents software and concepts well established in the breeder routine (Mrode, 2014), which means that predictive models and derivations of them (e.g., inclusion of interaction and/or non-additive effects) can be straightforwardly implemented. About modelling statistical, another advantage is the possibility to consider one-stage approach. Most GS studies use a two-stage analysis, where in a first stage the phenotypic data are pre-adjusted with estimates of non-genetic effects and, in a second stage, these adjusted metrics are considered in penalized regressions methods (RR-BLUP, in most cases) (Oakey *et al.*, 2016). Although represent lower computational demand, two-stages approach biases marker effects and induces heterogeneous residual variances and residual correlations, that are not completely eliminated by a weighted analysis (de Los Campos *et al.*, 2013). For this reason, when feasible, one-stage approach should be preferred. Chapter 2 investigated whole-genome regressions, including penalized and Bayesian estimation procedures, as well as non-parametric regressions and dimension reduction procedure. A central idea was relaxing the usual assumption of marker effects drawn from a normal distribution, which means seek for a possible association between model and trait, conditional to the genetic architecture. Although based on particular genetic and statistical assumptions, minimal differences were observed in terms of predictive ability. Therefore, models that showed less computational demand ("rrblup" and "gemma") can be considered for future investigations.

Considering some questions addressed to practical implementation in coffee breeding program, in Chapter 1 the MET.GBLUP model showed the best goodness of fit and predictive ability. Traditionally, one cycle of phenotypic recurrent selection in *C. canephora* consists of: i) Development of progenies from a base population; ii) phenotypic evaluation of the progenies in multiple environments and harvests; and iii) selection and recombination of the best selected individuals to form a new base population. Intuitively, the objectives is to generate an improved population by increasing the frequency of favorable alleles while maintaining sufficient genetic variation for subsequent cycles of selection (Windhausen *et al.*, 2012). A short term, a potential application is select individuals in both population (Intermediate and Premature) considering genomic prediction. Hence, after on recombination cycle, progenies can be genotyped and MET.GBLUP model would be used to predict the genetic merit of individuals unphenotyped in both locals. Our prospect is the reducing of the breeding cycle (avoiding long testing phases) and increases the selection intensity, through genotypic evaluation of a larger number of candidates. In contrast to the conventional recurrent selection program, including marker-assisted in coffee breeding schemes, it is expected a reduction of two-thirds (5-6 years) to the total time required to advance one generation.

In Chapter 2 it was discussed a hypothetical situation which a unique training population would be considered to calibrate a predictive model and the estimated markers effects used to predict phenotypic performances in other conditions (locals or populations). It is noteworthy that positive accuracy values were observed, in special, for across-locals predictions. As perspective, these results have potential to be included in new breeding schemes.

An open question addressed in Chapter 2 is the lack of information about genetic architecture of complex traits. Certainly, towards in this direction is a challenge in coffee research (Tran *et al.*, 2016). A recent approach that has been investigated in GS research is not focus only on predictions, but also aggregate two important features: identify SNP associated with the trait and understand its genetic architecture (Spindel *et al.*, 2015; MacLeod *et al.*, 2016). It seems clear that investigate which genetic variants have common and specific effects on environments or populations can help the selection of generalist genotypes (good performance in all conditions; i.e., broad adaptation) or specialist (performance directed for a specific condition; i.e., narrow adaptation). Broadly speaking, the problem of identifying relevant SNPs considering multilocus association models, in such way, approximate GS methods with contemporaneous GWAS algorithm (O'Hara and Sillanpää, 2009). The primary rationale of GWAS investigations is the idea that, by examining SNPs in details, important insights about the underlying biologic phenomenon can be discovery (Guan and Stephens, 2011). Therefore, it is reasonable to consider that modern GS analysis can borrow particularity from GWAS method - identify important covariates and learn about underlying biologic process – and uses them for prediction tasks.

A further conclusion addressed the use of GBS approach. The biallelic nature of SNP markers makes them less informative than microsatellites, molecular marker commonly used in coffee studies (Ferrão *et al.*, 2015; Moncada *et al.*, 2015). However, this disadvantage is easily overcome by their high abundance, ease and high throughput of their discovery and the robustness and automation of SNP genotyping assays. Promising results in terms of number and density of SNPs across the genome suggesting that GBS can be used as an efficient genotyping method in coffee research. Considering that coffee species suffer with the absence of a standard genotyping platform, GBS approach presents the advantage to simultaneous marker discovery and genotyping across the whole population of interest, making it rapid, flexible and suitable for species with limited genomic resources.

As a final message, GS approach is recommended as a promising and innovative approach to be applied in coffee breeding programs. 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 GS implementation.

## REFERENCES


Akaike, H., 1974 A new look at the statistical model identification. IEEE transactions on automatic control **19**: 716–723.

Asoro, F. G., M. a. Newell, W. D. Beavis, M. P. Scott, and J.-L. Jannink, 2011 Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. The Plant Genome Journal **4**: 132.

Balsalobre, T. W. A., M. C. Mancini, G. d. S. Pereira, C. O. Anoni, F. Z. Barreto, H. P. Hoffmann, A. P. de Souza, A. A. F. Garcia, and M. S. Carneiro, 2016 Mixed Modeling of Yield Components and Brown Rust Resistance in Sugarcane Families. Agronomy Journal **108**: 1–14.

Beaulieu, J., T. K. Doerksen, J. MacKay, A. Rainville, and J. Bousquet, 2014 Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC genomics **15**: 1048.

Burgueño, J., J. Crossa, J. M. Cotes, F. S. Vicente, and B. Das, 2011 Prediction assessment of linear mixed models for multienvironment trials. Crop Science **51**: 944–954.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa, 2012 Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Science **52**: 707.

Carbonetto, P. and M. Stephens, 2012 Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis **7**: 73–108.

Chan, A. W., M. T. Hamblin, and J.-L. Jannink, 2016 Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data. PLoS ONE **11**: e0160733.

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

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis, 2010 Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genetics Selection Evolution **42**: 1–11.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J. M. Hickey, C. Chen, G. de Los Campos, J. Burgueño, V. S. Windhausen, E. Buckler, J.-L. Jannink, M. a. Lopez Cruz, and R. Babu, 2013 Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. G3:Genes|Genomes|Genetics **3**: 1903–1926.

Crossa, J., G. de los Campos, M. Maccaferri, R. Tuberosa, J. Burgueño, and P. Pérez-Rodríguez, 2016 Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Science **56**: 2193–2209.

Cullis, B. R., A. B. Smith, and N. E. Coombes, 2006 On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics **11**: 381.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics **193**: 347–65.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and . G. P. A. Group, 2011 The variant call format and VCFtools. Bioinformatics **27**: 2156–2158.

de Los Campos, G., D. Gianola, and G. J. Rosa, 2009 Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci **87**.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics **182**: 375–85.

De Roos, A., B. Hayes, and M. Goddard, 2009 Reliability of genomic predictions across multiple populations. Genetics **183**: 1545–1553.

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

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one **6**: e19379.

Endelman, J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal **4**: 250.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman, C. Reich, B. Mason, and M. Goddard, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science **95**: 4114–4129.

Falconer, D. S. and T. F. C. Mackay, 1996 *Quantitative Genetics*. Pearson Education Limited, England.

Ferrão, L., E. Caixeta, G. Pena, E. Zambolim, C. Cruz, L. Zambolim, M. Ferrão, and N. Sakiyama, 2015 New EST–SSR markers of Coffea arabica: transferability and application to studies of molecular characterization and genetic mapping. Molecular Breeding **35**: 1–5.

Ferrão, L. F. V., R. G. Ferrão, M. A. G. Ferrão, A. Fonseca, M. Stephens, and A. A. F. Garcia, 2016 Genomic prediction in Coffea canephora using Bayesian polygenic modeling. In *5th International Conference on Quantitative Genetics*, p. 203, Madison, WI.

Ferrão, R. G., M. A. G. Ferrão, A. Fonseca, and B. Pacova, 2007 Melhoramento Genético de Coffea canephora. In *Cafe Conilon*, edited by R. Ferrão, A. Fonseca, S. Bragança, M. Ferrão, and L. D. Muner, pp. 123–173, Vitória-ES, incaper edition.

Fisher, R. A., 1919 XV.—The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the royal society of Edinburgh **52**: 399–433.

Friedman, J. H., T. Hastie, and R. Tibshirani, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software; Vol 1, Issue 1 (2010) .

Gamal El-Dien, O., B. Ratcliffe, J. Klápště, C. Chen, I. Porth, and Y. a. El-Kassaby, 2015 Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC genomics **16**: 370.

Garrick, D., J. Dekkers, and R. Fernando, 2014 The evolution of methodologies for genomic prediction. Livestock Science pp. 1–9.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2014 *Bayesian data analysis*, volume 2. Taylor & Francis.

Gianola, D., 2013 Priors in whole-genome regression: the bayesian alphabet returns. Genetics **194**: 573–596.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the bayesian alphabet. Genetics **183**: 347–363.

Gianola, D. and J. B. C. H. M. van Kaam, 2008 Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178**: 2289–2303.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, 2014 Tassel-gbs: a high capacity genotyping by sequencing analysis pipeline. PloS one **9**: e90346.

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.

Goddard, M. E. and B. J. Hayes, 2007 Genomic selection. Journal of animal breeding and genetics = Zeitschrift für Tierzüchtung und Züchtungsbiologie **124**: 323–30.

Grattapaglia, D. and M. D. V. Resende, 2010 Genomic selection in forest tree breeding. Tree Genetics & Genomes **7**: 241–255.

Grenier, C., T.-V. Cao, Y. Ospina, C. Quintero, M. H. Châtel, J. Tohme, B. Courtois, and N. Ahmadi, 2015 Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLOS ONE **10**: 1–25.

Guan, Y. and M. Stephens, 2011 Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics pp. 1780–1815.

Habier, D., R. Fernando, K. Kizilkaya, and D. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**: 186.

Hartl, D. L., A. G. Clark, and A. G. Clark, 1997 *Principles of population genetics*, volume 116. Sinauer associates Sunderland.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution **41**: 51.

Heffner, E. L., J.-L. Jannink, and M. E. Sorrells, 2011 Genomic selection accuracy using multi-family prediction models in a wheat breeding program. The Plant Genome **4**: 65–75.

Henderson, C. R., 1949 Estimation of changes in herd environment. J Dairy Sci **32**: 706.

Heslot, N., D. Akdemir, M. E. Sorrells, and J.-L. Jannink, 2014 Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics **127**: 463–480.

Heslot, N., H.-P. Yang, M. E. Sorrells, and J.-L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science **52**: 146–160.

Hoerl, A. E. and R. W. Kennard, 1970 Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**: 55–67.

Hu, Y., C. Yan, C.-H. Hsu, Q.-R. Chen, K. Niu, G. A. Komatsoulis, and D. Meerzaman, 2014 OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data. Cancer Informatics **13**: 13–20.

IOC, 2016 International Coffee Organization - Trade Statistics Tables.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013 *An introduction to statistical learning*, volume 6. Springer.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics **9**: 166–77.

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

Júnior, G. A. F., G. J. Rosa, B. D. Valente, R. Carvalheiro, F. Baldi, D. A. Garcia, D. G. Gordo, R. Espigolan, L. Takada, R. L. Tonussi, *et al.*, 2016 Genomic prediction of breeding values for carcass traits in nellore cattle. Genetics Selection Evolution **48**: 7.

Kärkkäinen, H. P. and M. J. Sillanpää, 2012 Back to Basics for Bayesian Model Building in Genomic Selection. Genetics **191**: 969–987.

Kelly, A. M., B. R. Cullis, A. R. Gilmour, J. A. Eccleston, and R. Thompson, 2009 Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genetics Selection Evolution **41**: 1.

Lande, R. and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**: 743–756.

Lehermeier, C., C.-C. Schon, and G. de los Campos, 2015 Assessment of Genetic Heterogeneity in Structured Plant Populations Using Multivariate Whole-Genome Regression Models. Genetics **201**: 323–337.

Liaw, A. and M. Wiener, 2002 Classification and regression by randomforest. R news **2**: 18–22.

Lopez-Cruz, M., J. Crossa, D. Bonnett, S. Dreisigacker, J. Poland, J.-L. Jannink, R. P. Singh, E. Autrique, and G. de los Campos, 2015 Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. G3; Genes|Genomes|Genetics **5**: 569–82.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H. G. Gauch Jr, R. Okechukwu, A. G. Dixon, P. Kulakow, and J.-L. Jannink, 2013 Relatedness and genotype× environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Science **53**: 1312.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, a. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics **17**: 144.

Malosetti, M., D. Bustos-Korts, M. Boer, and F. van Eeuwijk, 2016 Predicting Responses in Multiple Environments: Issues in Relation to Genotype x Environment Interactions. Crop Science **13**: (accepted).

Malosetti, M., J.-M. Ribaut, and F. A. van Eeuwijk, 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**: 53.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, 2009 Finding the missing heritability of complex diseases. Nature **461**: 747–753.

Margarido, G. R. A., M. M. Pastina, A. P. Souza, and A. A. F. Garcia, 2015 Multi-trait multi-environment quantitative trait loci mapping for a sugarcane commercial cross provides insights on the inheritance of important traits. Molecular Breeding **35**: 175.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Mevik, B.-H., R. Wehrens, *et al.*, 2007 The pls package: principal component and partial least squares regression in r. Journal of Statistical software **18**: 1–24.

Moncada, P. M. D., E. Tovar, J. C. Montoya, A. González, J. Spindel, and S. McCouch, 2015 A genetic linkage map of coffee (Coffea arabica L.) and QTL for yield, plant height, and bean size. Tree Genetics & Genomes **12**: 1–17.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma, 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol **41**.

Mrode, R. A., 2014 *Linear models for the prediction of animal breeding values*. Cabi.

Neves, H. H., R. Carvalheiro, and S. A. Queiroz, 2012 A comparison of statistical methods for genomic selection in a mice population. BMC genetics **13**: 100.

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

O'Hara, R. B. and M. J. Sillanpää, 2009 A review of bayesian variable selection methods: What, how and which. Bayesian Analysis **4**: 85–118.

Park, T. and G. Casella, 2008 The Bayesian Lasso. Journal of the American Statistical Association **103**: 681–686.

Pastina, M. M., M. Malosetti, R. Gazaffi, M. Mollinari, G. R. A. Margarido, K. M. Oliveira, L. R. Pinto, A. P. Souza, F. A. van Eeuwijk, and A. A. F. Garcia, 2012 A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. Theoretical and Applied Genetics **124**: 835–849.

Pérez, P. R. and G. de los Campos, 2013 BGLR: Bayesian generalized linear regression. R package version .

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

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team, 2016 *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink, 2012a Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome Journal **5**: 103.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink, 2012b Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome Journal **5**: 103.

R Core Team, 2013 R: A Language and Environment for Statistical Computing.

Rencher, A. C. and G. B. Schaalje, 2008 *Linear Models in Statistics*. Hoboken, New Jersey, john wiley edition.

Resende, M. F. R., P. Muñoz, M. D. V. Resende, D. J. Garrick, R. L. Fernando, J. M. Davis, E. J. Jokela, T. A. Martin, G. F. Peter, and M. Kirst, 2012 Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics **190**: 1503–1510.

Riedelsheimer, C., F. Technow, and A. E. Melchinger, 2012 Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC genomics **13**: 452.

Schulz-Streeck, T., J. Ogutu, Z. Karaman, C. Knaak, and H. Piepho, 2012 Genomic selection using multiple populations. Crop Science **52**: 2453–2461.

Schulz-Streeck, T., J. O. Ogutu, and H. . P. Piepho, 2013 Comparisons of single-stage and two-stage approaches to genomic selection. Theor Appl Genet **126**.

Schwarz, G., 1978 Estimating the dimension of a model. The Annals of Statistics **6**: 461–464.

Silva, F. F., L. Varona, M. D. V. de Resende, J. S. S. B. Filho, G. J. M. Rosa, and J. M. S. Viana, 2011 A note on accuracy of Bayesian LASSO regression in GWS. Livestock Science **142**: 310–314.

Smith, a., B. Cullis, and R. Thompson, 2001 Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics **57**: 1138–1147.

Smith, a. B., B. R. Cullis, and R. Thompson, 2005 The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. The Journal of Agricultural Science **143**: 449.

Smith, K. F. and M. D. Casler, 2004 Spatial Analysis of Forage Grass Trials across Locations, Years, and Harvests. Crop Science **44**: 56–62.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redona, G. Atlin, J. L. Jannink, and S. R. McCouch, 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 Genetics **11**: 1–25.

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

Tempelman, R. J., 2015 Statistical and Computational Challenges in Whole Genome Prediction and Genome-Wide Association Analyses for Plant and Animal Breeding. Journal of Agricultural, Biological, and Environmental Statistics **20**: 442–466.

Thavamanikumar, S., R. Dolferus, and B. R. Thumma, 2015 Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3: Genes|Genomes|Genetics **5**: 1991–1998.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288.

Tran, H. T. M., L. S. Lee, A. Furtado, H. Smyth, and R. J. Henry, 2016 Advances in genomics for the improvement of quality in coffee. Journal of the Science of Food and Agriculture **96**: 3300–3312.

van der Vossen, H., B. Bertrand, and A. Charrier, 2015 Next generation variety development for sustainable production of arabica coffee (coffea arabica l.): a review. Euphytica **204**: 243–256.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science **91**: 4414–4423.

Vazquez, A., G. Rosa, K. Weigel, G. De los Campos, D. Gianola, and D. Allison, 2010 Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in us holsteins. Journal of dairy science **93**: 5942–5949.

VSN International, 2011 GenStat for Windows 14th Edition.

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

Wang, X., Z. Yang, and C. Xu, 2015 A comparison of genomic selection methods for breeding value prediction. Science Bulletin **60**: 925–935.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genetical research **75**: 249–52.

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schoen, 2012 synbreed: a framework for the analysis of genomic prediction data using r. Bioinformatics **28**: 2086–2087.

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink, M. E. Sorrells, B. Raman, J. E. Cairns, A. Tarekegne, K. Semagn, *et al.*, 2012 Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3: Genes| Genomes| Genetics **2**: 1427–1436.

Xavier, A., W. M. Muir, B. Craig, and K. M. Rainey, 2016 Walking through the statistical black boxes of plant breeding. Theoretical and Applied Genetics **129**: 1933–1949.

Zhou, X., P. Carbonetto, and M. Stephens, 2013 Polygenic modeling with bayesian sparse linear mixed models. PLoS genetics **9**: e1003264.

Zhou, X. and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nature genetics **44**: 821–4.

Zhou, X. and M. Stephens, 2014 Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods **11**: 407–409.