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 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.