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

Practical considerations for genotype imputation and multi-trait multi-environment genomic prediction in a tropical maize breeding program

Amanda Avelar de Oliveira

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

Piracicaba
2019

Amanda Avelar de Oliveira
Agronomist

Practical considerations for genotype imputation and multi-trait multi-environment genomic prediction in a tropical maize breeding program

Advisor:
Prof. Dr. **GABRIEL RODRIGUES ALVES MARGARIDO**

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

Piracicaba
2019

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

Oliveira, Amanda Avelar de

Practical considerations for genotype imputation and multi-trait multi-environment genomic prediction in a tropical maize breeding program / Amanda Avelar de Oliveira. - - Piracicaba, 2019.

72p.

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

1. Seleção genômica 2. Dados perdidos 3. Modelos multivariados 4. Genotipagem por sequenciamento I. Título

I dedicate with love to

my parents Carlos Alberto and Maria Lilian,

my sister Ana Luiza

and my husband Ricardo.

ACKNOWLEDGEMENTS

To the “Luiz de Queiroz” College of Agriculture - University of São Paulo (ESALQ-USP) and the Graduate Program in Genetics and Plant Breeding for all the support throughout my academic time.

The financial support from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), are greatfully acknowlegded. Additionally, to Embrapa Milho e Sorgo, who provided the phenotypic and genotypic data.

Appreciation goes to my advisor Dr. Gabriel Rodrigues Alves Margarido for his generosity, guidance and support, which were essential during my graduate studies.

Special acknowledgement goes to Dr. Maria Marta Pastina, who contributed enormously to this work and was always available to assist me, as well for being an example of woman in science.

I am also sincerely grateful to Dr. Marcio Resende Jr. for the opportunity to join his amazing team during six months at Sweet Corn Genomics and Breeding, University of Florida, Gainesville, FL, USA. All his support and guidance during this time were essential for the conducted of this work.

Additionally, I would like to thank all the professors and staff of the Department of Genetics at ESALQ/USP and of the Horticultural Science Department at University of Florida, for their support and extensive knowledge provided.

Thanks to all my colleagues from the Laboratory of Bioinformatic Applied to Bioenergy for all their support, talks, friendship and scientific discussions. I am also grateful to the colleagues from Sweet Corn and Genomic Breeding, it was a pleasure worked with you guys.

My sincerely gratitude goes to Luis Felipe Ferrão e Juliana Benevenuto, who hosted me at Gainesville and were a family to me. I am also thankfull to Rodrigo Amandeu and Maria Fernanda Trientini, I really miss the "Rep UF-ESALQ".

I extend my acknowledgements to all my family. Specially, to my parents, Carlos Alberto and Maria Lilian and my sister Ana Luiza. This work would not be possible without their continuous love, help and support.

To my husband Ricardo for his unconditional support, patience, encouragement and love.

Finally, to God, my guide throughout all the success and frustration.

RESUMO

Considerações práticas para a imputação de genótipos e predição genômica aplicada a múltiplos caracteres e ambientes em um programa de melhoramento de milho tropical

A disponibilidade de marcadores moleculares cobrindo todo o genoma, como os polimorfismos de nucleotídeos individuais (*single nucleotide polymorphism* - SNP), aliada aos recursos computacionais para o processamento de grande volume de dados, tornou possível o desenvolvimento de uma abordagem de melhoramento assistido para caracteres de herança quantitativa, conhecida como seleção genômica. Na última década a seleção genômica tem sido implementada com sucesso em uma enorme variedade de espécies animais e vegetais, comprovando suas vantagens sobre a seleção assistida por marcadores tradicional e a seleção baseada apenas em informações de parentesco. No entanto, alguns desafios práticos ainda podem limitar a implementação deste método em um programa de melhoramento de plantas. Como exemplos, citam-se o custo da genotipagem de alta densidade de um grande número de indivíduos e a aplicação de modelos mais complexos, que consideram múltiplos caracteres e ambientes. Dessa forma, este estudo teve como objetivos: *i*) investigar estratégias de identificação de SNPs e imputação que possibilitem uma genotipagem de alta densidade economicamente viável; e *ii*) avaliar a aplicação de modelos multivariados de seleção genômica para múltiplos caracteres e ambientes. Este trabalho foi divido em dois capítulos. No primeiro capítulo, comparou-se a acurácia de quatro métodos de imputação: NPUTE, Beagle, KNNI e FILLIN, usando dados de genotipagem por sequenciamento (*genotyping-by-sequencing* – GBS) de 1.060 linhagens de milho, que foram genotipadas usando diferentes profundidades de cobertura. Além disso, duas estratégias de identificação de SNPs e imputação foram avaliadas. Os resultados indicaram que a combinação de estratégias de detecção de polimorfismos e imputação pode possibilitar uma genotipagem economicamente viável, resultando em maiores acurácias de imputação. No segundo capítulo, modelos multivariados de seleção genômica, para múltiplos caracteres e ambientes, foram comparados com suas versões univariadas. Dados de 415 híbridos avaliados na segunda safra em quatro anos (2006-2009) para os caracteres produtividade de grãos, número de espigas e umidade foram utilizados. Os genótipos dos híbridos foram inferidos *in silico* com base nos genótipos das linhagens parentais usando marcadores SNPs obtidos via GBS. No entanto, informações genotípicas estavam disponíveis para apenas 257 híbridos, de modo que foi necessário fazer uso da matriz **H**, a qual combina informações de parentesco genético baseadas em pedigree e marcadores. Os resultados obtidos demonstraram que o uso de modelos de seleção genômica para múltiplos caracteres e ambientes pode aumentar a capacidade preditiva, especialmente para predizer a performance de híbridos nunca avaliados em qualquer ambiente.

Palavras-chave: Seleção genômica; Dados perdidos; Modelos multivariados; Genotipagem por sequenciamento

ABSTRACT

Practical considerations for genotype imputation and multi-trait multi-environment genomic prediction in a tropical maize breeding program

The availability of molecular markers covering the entire genome, such as single nucleotide polymorphism (SNP) markers, allied to the computational resources for processing large amounts of data, enabled the development of an approach for marker assisted selection for quantitative traits, known as genomic selection. In the last decade, genomic selection has been successfully implemented in a wide variety of animal and plant species, showing its benefits over traditional marker assisted selection and selection based only on pedigree information. However, some practical challenges may still limit the wide implementation of this method in a plant breeding program. For example, we cite the cost of high-density genotyping of a large number of individuals and the application of more complex models that take into account multiple traits and environments. Thus, this study aimed to *i*) investigate SNP calling and imputation strategies that allow cost-effective high-density genotyping, as well as *ii*) evaluating the application of multivariate genomic selection models to data from multiple traits and environments. This work was divided into two chapters. In the first chapter, we compared the accuracy of four imputation methods: NPUTE, Beagle, KNNI and FILLIN, using genotyping-by-sequencing (GBS) data from 1060 maize inbred lines, which were genotyped using different depths of coverage. In addition, two SNP calling and imputation strategies were evaluated. Our results indicated that combining SNP-calling and imputation strategies can enhance cost-effective genotyping, resulting in higher imputation accuracies. In the second chapter, multivariate genomic selection models, for multiple traits and environments, were compared with their univariate versions. We used data from 415 hybrids evaluated in the second season in four years (2006-2009) for grain yield, number of ears and grain moisture. Hybrid genotypes were inferred *in silico* based on their parental inbred lines using SNP markers obtained via GBS. However, genotypic information was available only for 257 hybrids, motivating the use of the **H** matrix, which combines genetic information based on pedigree and molecular markers. Our results demonstrated that the use of multi-trait multi-environment models can improve predictive abilities, especially to predict the performance of hybrids that have not yet been evaluated in any environment.

Keywords: Genomic selection; Missing data; Multivariate models; Genotyping-by-sequencing

1 GENERAL INTRODUCTION

The leveraging of heterosis has been extremely successful in affording continuous improvement in commercial maize grain yield. Since the beginning of the hybrid era, maize breeders achieved increases in grain yield that are unmatched among other cereals or oil seeds (Lee and Tracy 2009; Hallauer and Miranda Filho 2010). Classical maize breeding consists of crossing lines from different heterotic groups and measuring phenotypic performance of hybrids in multiple environment trials. However, phenotyping has become one of the most costly and laborious stages in a breeding program. Thus, genomic prediction stands out in virtue of its ability to reduce the time required to complete a breeding cycle, to enable an earlier and more efficient selection of superior genotypes, and to reduce phenotyping costs, representing a promising tool for use in maize breeding programs (Crossa et al. 2017; Wang et al. 2018).

Genomic prediction was first proposed in 2001 by Meuwissen et al. Since then, it has been applied to a variety of crops and routinely practiced in breeding programs of major seed companies, especially for maize and soybean (Bernardo 2016). The key idea of this method is the simultaneous prediction of the effects of a large number of markers spread throughout the genome, in order to ensure that every quantitative trait locus (QTL) affecting a trait be in linkage disequilibrium (LD) with at least one marker. This method remained unexplored for a few years, because the molecular markers available at that time were limited and obtained at high costs. However, the emerging of next-generation sequencing technology presented the possibility of obtaining molecular markers densely distributed across the genome, using high-throughput techniques such as genotyping-by-sequencing (GBS) (Elshire et al. 2011).

Feature of GBS data are the high rates of missingness and heterozygote undercalling, prompting the use of approaches to impute these missing genotypes. In this scenario, several studies have assessed the efficiency of imputing missing data, using different methods and strategies (Howie et al. 2009; Cleveland et al. 2011; Hickey et al. 2012; Swarts et al. 2014; Bouwman et al. 2014; Nazzicari et al. 2016; Gonen et al. 2018). Besides that, the cost of genotyping many samples at high density is still high, representing a barrier to small or public plant breeding programs to routinely implement genomic prediction. Therefore, it is necessary to adopt low cost genotyping strategies to solve this limitation. For species for which genotyping chips are available, combining data from high and low density SNP arrays is a cost effective strategy (Jacobson et al. 2015; Hickey et al. 2015; Gorjanc et al. 2017). When genotyping chips are not available, the GBS technology allows breeders to adjust the amount of retrieved information and its cost by choosing different restriction enzymes, regulating sequencing depth and the level of multiplexing (Elshire et al. 2011; Deschamps et al. 2012; Poland and Rife 2012).

Currently, the majority of genomic prediction models applied are univariate ones. However, in breeding programs it is common to evaluate several traits simultaneously, because elite genotypes should concentrate favorable alleles for several traits of interest. The existence of genetic correlation between quantitative traits indicates that measures in one trait provide indirect information about other traits, a fact that can be used to improve the predictive ability of genomic selection (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014; Dos Santos et al. 2016; Marchal et al. 2016; Lyra et al. 2017; Covarrubias-Pazaran et al. 2018). Besides the correlation between traits, considering a model that also accommodate the genotype by environment interaction, is also an important issue to plant breeders, since genotypes are evaluated for multiple traits in multiple environments. The types of model that jointly take into account multiple traits and environments are referred to as multi-trait multi-environment (MTME) models. Nonetheless, few studies have simultaneously assessed multiple traits and multiple environments for genomic selection purposes (Montesinos-López et al. 2016; Gomes Torres et al. 2018; Ward et al. 2019).

The complexity of applying genomic prediction in plant breeding programs arises at different levels and is influenced by several factors. In order to investigate some of the challenges faced by breeders, when applying genomic prediction to a maize breeding program, this work is the result of a partnership among: Embrapa Milho e Sorgo (Sete Lagoas, MG, Brazil), the Laboratory of Bioinformatics Applied to Bioenergy at ESALQ/USP ("Luiz de Queiroz" College of Agriculture, University of São Paulo - Piracicaba, SP, Brazil) and the Sweet Corn Genomics and Breeding at University of Florida, Gainesville, FL, USA. In this context, we conducted two studies that are herein organized in two chapters. In the first chapter, we aimed to evaluate different SNP calling and imputation strategies using GBS data of maize lines from the Embrapa maize breeding program. Subsequently, chapter 2 focuses on applications of multi-trait multi-environment genomic prediction models to second season maize hybrids, which also originated from the Embrapa breeding program.

References

Bernardo R (2016) Bandwagons I, too, have known. Theor Appl Genet 129:2323–2332. https://doi: 10.1007/s00122-016-2772-5

Bouwman AC, Hickey JM, Calus MP, Veerkamp RF (2014) Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. Genet Sel Evol 46:6. https://doi: 10.1186/1297-9686-46-6

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43:26. https://doi: 10.1186/1297-9686-43-26

Cleveland MA, Hickey JM, Kinghorn BP (2011) Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. BMC Proc 5:S6. https://doi: 10.1186/1753-6561-5-S3-S6

Covarrubias-Pazaran G, Schlautman B, Diaz-Garcia L, et al (2018) Multivariate GBLUP improves accuracy of genomic selection for yield and fruit weight in biparental populations of *Vaccinium macrocarpon* Ait. Front Plant Sci 9:1310. https://doi: 10.3389/fpls.2018.01310

Crossa J, Pérez-Rodríguez P, Cuevas J, et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975. https://doi: 10.1016/J.TPLANTS.2017.08.011

Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. Biology (Basel) 1:460–483. https://doi: 10.3390/biology1030460

Dos Santos JPR, De Castro Vasconcellos RC, Pires LPM, et al (2016) Inclusion of dominance effects in the multivariate GBLUP model. PLoS One 11:1–21. https://doi: 10.1371/journal.pone.0152045

Elshire RJ, Glaubitz JC, Sun Q, et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:1–10. https://doi: 10.1371/journal.pone.0019379

Gomes Torres L, Rodrigues MC, Lima NL, et al (2018) Multi-trait multi-environment Bayesian model reveals G x E interaction for nitrogen use efficiency components in tropical maize. PLoS One 13: 1–15. https://doi: 10.1371/journal.pone.0199492

Gonen S, Wimmer V, Gaynor RC, et al (2018) A heuristic method for fast and accurate phasing and imputation of single nucleotide polymorphism data in bi- parental plant populations. Theor Appl Genet 131:2345–2357. https://doi: https://doi.org/10.1007/s00122-018-3156-9

Gorjanc G, Dumasy J-F, Gonen S, et al (2017) Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. Crop Sci 57:1404. https://doi: 10.2135/cropsci2016.08.0675

Guo G, Zhao F, Wang Y, et al (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15:1–7. https://doi: 10.1186/1471-2156-15-30

Hallauer A., Miranda Filho J. (2010) Quantitative genetics in maize breeding., 2.ed. Iowa State University Press, Ames

Hickey JM, Crossa J, Babu R, de los Campos G (2012) Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci 52:654. https://doi: 10.2135/cropsci2011.07.0358

Hickey JM, Gorjanc G, Varshney RK, Nettelblad C (2015) Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a Hidden Markov Model. Crop Sci 55:1934. https://doi: 10.2135/cropsci2014.09.0648

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529. https://doi: 10.1371/journal.pgen.1000529

Jacobson A, Lian L, Zhong S, Bernardo R (2015) Marker imputation before genomewide selection in biparental maize populations. Plant Genome 8:1–9. https://doi: 10.3835/plantgenome2014.10.0078

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522. https://doi: 10.1534/genetics.112.144246

Lee EA, Tracy WF (2009) Modern maize breeding. In: Bennetzen J.L., Hakes S. (eds) Handbook of Maize. Springer, New York, New York, pp 141–160

Lyra DH, Mendonça L F, Galli G, et al (2017) Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. Mol Breed 37:80. https://doi: 10.1007/s11032-017-0681-1

Marchal A, Legarra A, Sébastien T, et al (2016) Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. Mol Breed 36:2. https://doi: 10.1007/s11032-015-0423-1

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi: 11290733

Montesinos-López OA, Montesinos-López A, Crossa J, et al (2016) A genomic bayesian multi-trait and multi-environment model. G3 Gene Genome Genet 6:2725–2744. https://doi: 10.1534/g3.116.032359

Nazzicari N, Biscarini F, Cozzi P, Brummer EC (2016) Marker imputation efficiency for genotyping-by-sequencing data in rice (*Oryza sativa*) and alfalfa (*Medicago sativa*). Mol Breed 36:1–16. https://doi: 10.1007/s11032-016-0490-y

Poland J a, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome J 5:92–102. https://doi: 10.3835/plantgenome2012.05.0005

Swarts K, Li H, Romero Navarro JA, et al (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12. https://doi: 10.3835/plantgenome2014.05.0023

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: Current status and prospects. Crop J 6:330–340. https://doi: 10.1016/j.cj.2018.03.001

Ward BP, Brown-Guedira G, Tyagi P, et al (2019) Multienvironment and multitrait genomic selection models in unbalanced early-generation wheat yield trials. Crop Sci 59:1–17. https://doi: 10.2135/cropsci2018.03.0189