**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

**Statistical models for genomic selection in *Panicum maximum*
considering allelic dosage**

Letícia Aparecida de Castro Lara

Thesis presented to obtain the degree of Doctor in Science.

Area: Genetics and Plant Breeding

**Piracicaba
2017**

Letícia Aparecida de Castro Lara
Bachelor in Biological Sciences

**Statistical models for genomic selection in *Panicum maximum*
considering allelic dosage**

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

Lara, Letícia Aparecida de Castro

Statistical models for genomic selection in *Panicum maximum* considering allelic dosage / Letícia Aparecida de Castro Lara. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2017 .

66 p.

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

1. Melhoramento de plantas 2. Forrageira 3. Autotetraploides 4. Modelos lineares mistos 5. Predição I. Título.

DEDICATORY

To my parents José Maurício and Liberaci,
My angels João Gabriel and Benjamin,
And my fiancé Thiago.
I love you.

ACKNOWLEDGMENTS

I thank God and “Nossa Senhora da Aparecida” for being presence in my life.

To the University of São Paulo, especially “Luiz de Queiroz” College of Agriculture (ESALQ) for my professional qualification.

Financial support from the FAPESP (“Fundação de Amparo à Pesquisa do Estado de São Paulo”), grants 2015/20659-2 and 2016/01279-7; the CNPq (“Conselho Nacional de Desenvolvimento Científico e Tecnológico”); and the CAPES (“Coordenação de Aperfeiçoamento de Pessoal de Nível Superior”) are gratefully acknowledged. Additional support was provided by Embrapa Beef Cattle and by UNIPASTO (“Associação para o Fomento à Pesquisa de Melhoria de Forrageiras”).

Special acknowledgment goes to my advisor Prof. Antonio Augusto Franco Garcia for the support of my Ph.D study and for his academic and friendly advice as well as for being an example of professional for me.

My sincere thanks also goes to Prof. Zhao-Bang Zeng, who provided me an opportunity to join his team as intern and for the several valuable discussions over six months of exchange abroad.

I would also like to acknowledge the entire Embrapa Beef Cattle team, especially Dr. Mateus Figueiredo Santos and Dra. Liana Jank, for always being accessible, for their contributions and assistance. I learned a lot about forage breeding with them.

I am also grateful to all the staff and professors of the courses Genetics and Plant Breeding, and Statistics and Agronomic Experimentation of ESALQ/USP, specially Prof. Gabriel Rodrigues Alves Margarido, Prof. Sílvia Zocchi, Berdan, Fernandinho, Valdir, and Léia. I would like to thank Dr. Mateus F. Santos, Dra. Bianca Baccili Zanotto Vigna, Prof. Roberto Fritsche-Neto, Prof. Gabriel R. A. Margarido, and Prof. A. Augusto F. Garcia for the evaluations and excellent contributions to this thesis.

Thanks to all my friends and my friends from Statistical Genetic Lab (ESALQ/USP). The “old group”: Adriana Cheavegatti, Carina Anoni, Danilo Cursi, Guilherme Pereira, João Ricardo Rosa, Luís Felipe Ferrão, Marcelo Mollinari, Maria Izabel Cavassim; the “new group”: Cristiane Taniguti, Gabriel Gesteira, Jhonathan Pedroso, Kaio Olímpio, Marianella Quezada, Matheus Krause, Rafael Nalin, Rodrigo Amadeu; and the “aggregates”: Amanda Avelar, Elaine Batista, Fernando Correr, and Thiago Oliveira. They were very helpful and contributed to my life in Piracicaba being better. Specially, I would like to thank Guilherme Pereira, Jhonathan Pedroso, Kaio Olímpio, Marcelo Mollinari, Rodrigo Amadeu, and Thiago Oliveira for clarifying my doubts during my analysis. I am also grateful Evelyn Couto, Marianella, Cristiane, Melina Andrade, and Karina Borges for always being good friends.

I would like to extend my acknowledgments to my family (José Maurício, Liberaci, Renato, Bruna, Lais, and Gabriel) and my family in law (Cido, Rose, and Tati), for their unconditional support during all stages of my life and for believing in my ability.

Finally, I would like to thank my fiancé Thiago Oliveira for his patience and for always believing in my potential.

I hope everything I learned is not an obstacle to humility, but an incentive to achieve great dreams.

CANÇÃO ETERNA

Lá estavam os pássaros, sorrindo sob o som da lua
Era primavera, verão, outono e inverno
Tinham vários olhos nos observando do céu
Ícones indignados com o tempo
Cuja pressa era assustadora!
Imersos íamos, porém, sendo levados pela aurora
Agarrados, como se a alegria fosse deveras única!
Andavam e corriam os lábios
Pureza humana! Natural, sedenta...
Amor é a consequência exata
Reflexo de duas almas livres
Esgueiradas de uma sociedade
Cujas diretrizes
Indubitavelmente insanas!

Dias, semanas, meses e anos
A música que a nós toca
Dança sem um registro de fim
E na solidão inexistente
Cercamos os olhos alheios deiscantes
Apáticos, raivosos, deprimentes...
Sarcásticos... Seria a felicidade cobiçada?
Treinados por regras irreais
Respiram os ares que não são deles!
O lado obscuro que nos cerca, então,
Lá fica, longe e decadente
Amor como este não se entende
Respira-se, sente-se e de repente
As vidas que nos cercam sentem

Author: Thiago de Paula Oliveira

SUMMARY

Resumo	7
Abstract	8
1 Introduction	9
References	10
2 Conclusion	11
References	11

RESUMO

Modelos genético-estatísticos para seleção genômica em *Panicum maximum* com informação de dosagem alélica

Diversas espécies de interesse econômico são autotetraploides, como a forrageira *Panicum maximum*, a qual proporciona alta produtividade e qualidade para pastagens tropicais. Os principais acessos na natureza são plantas apomíticas tetraploides, no entanto pode-se encontrar também plantas sexuais diploides. Embora a apomixia seja vantajosa pela facilidade em fixar o vigor híbrido, a reprodução sexual é fundamental por permitir recombinação genética a partir de cruzamentos entre genótipos superiores. Desta forma, o melhoramento nesta espécie consiste em cruzar plantas apomíticas com plantas sexuais tetraploidizadas. A utilização de parentais sexuais superiores nestes cruzamentos permite aumentar a frequência de alelos favoráveis na progênie. Portanto, programas de seleção recorrente intrapopulacional em populações sexuais tetraploides são fundamentais para programas de melhoramento em *P. maximum*. Além disto, a utilização de estratégias como seleção genômica são promissoras para aumentar os ganhos de seleção, permitindo avançar ciclos de seleção recorrente e lançar cultivares no mercado em menor prazo, quando comparados a programas convencionais. Como *P. maximum* é uma cultura perene, os genótipos são avaliados em sucessivos cortes. Assim, este estudo tem como finalidade avaliar caracteres de produtividade, estruturais e nutricionais em uma população sexual tetraploide de *P. maximum*, investigando diferentes classes de modelos lineares mistos aplicados a dados longitudinais, além de desenvolver modelos de seleção genômica que considerem a natureza tetraploide da população. Este trabalho foi dividido em dois capítulos. No primeiro capítulo, três classes de modelos foram analisados: i) Classe A consiste em modelar a interação genótipos por cortes com correlações homogêneas, genótipos não correlacionados entre si e os efeitos residuais são ajustados com homocedasticidade e ausência de correlação; ii) Classe B consiste em grupos de modelos com diferentes estruturas de variância e covariância (VCOV) para efeitos genéticos e residuais e genótipos não correlacionados; iii) Classe C é similar à Classe B, no entanto os genótipos são correlacionados por uma matriz de parentesco aditivo calculado por pedigree. Para todos os caracteres, os modelos da Classe C tiveram melhor ajuste. Portanto, recomenda-se testar matrizes de VCOV que permitam modelar cortes com diferentes níveis de correlações ao longo do tempo bem como incluir informação de parentesco aditivo e, se disponível, matriz de parentesco genômico. No segundo capítulo, marcadores SNPs, obtidos via genotipagem por sequenciamento, foram aplicados em modelos Bayesianos e GBLUP os quais foram desenvolvidos para incorporar informação de dosagem alélica tetraploide. Uma vez que as acurácias dos modelos Bayesianos não diferiram das acurácias do modelo GBLUP com dosagem alélica, recomenda-se o uso do segundo por requerer menos tempo computacional. A acurácia dos modelos preditivos reforça a vantagem em implementar seleção genômica em programas de melhoramento de *P. maximum*.

Palavras-chave: Melhoramento de plantas; Forrageira; Autotetraploides; Modelos lineares mistos; Predição

ABSTRACT

Statistical models for genomic selection in *Panicum maximum* considering allelic dosage

Several species of economic interest are autotetraploid, such as the forage *Panicum maximum*, which is responsible for high productivity and quality of tropical pastures. The main accessions in nature are autotetraploid apomictic plants, on the other hand, diploid sexual plants may also be found. Although apomixis is advantageous because it fixes hybrid vigor, sexual reproduction is fundamental to allow genetic recombination by crossing among superior genotypes. Thus, genetic breeding consists of crossing apomictic plants with tetraploidized sexual plants. In these crosses, the use of superior sexual parents allows to increase the frequency of favorable alleles in the progeny. Therefore, recurrent selection programs in tetraploid sexual populations are fundamental to *P. maximum* breeding programs and strategies such as genomic selection can increase the accuracy of selection, allowing shorter breeding cycles and release cultivars in the market in the short term when compared to conventional programs. As *P. maximum* is a perennial crop, genotypes are evaluated in successive harvests. Thus, the study goals are to evaluate nutritional, structural, and yield traits in a sexual tetraploid population of *P. maximum*, investigating different classes of linear mixed models applied to longitudinal data, as well as to develop genomic selection models which consider tetraploid allelic dosage. This work was split into two chapters. In the first chapter, three classes of models were analyzed: i) Class A consists in modeling the interaction of genotypes and harvests with homogeneous correlations, genotypes were assumed not correlated, and residual effects were assumed homocedastic and not correlated; ii) Class B consists of groups of models in which genetic and residual effects were fitted with different variance and covariance (VCOV) structures and genotypes were not correlated; and iii) Class C is similar to Class B, however genotypes were correlated by an additive relationship matrix based on pedigree values. For all traits, Class C models performed better based on goodness of fit of the models. Therefore, we recommend to incorporate additive relationship matrix besides to model harvests with different levels of correlations over time. In the second chapter, SNP markers, obtained by genotyping-by-sequencing (GBS) technique, were used to develop Bayesian and GBLUP models that consider tetraploid allelic dosage. Bayesian models accuracies did not differ from the accuracy of GBLUP model and, we recommend the latter because it requires less computational time. The accuracy of genomic selection models reinforces the advantage of implementing this strategy in *P. maximum* breeding programs.

Keywords: Plant breeding; Forage; Autotetraploids; Linear mixed models; Prediction

1 INTRODUCTION

Brazil is a leader in beef production, being the largest or second-largest producer in the world, competing only with the United States (JANK *et al.*, 2014). This position is due to vast pastures and cattle herds present in the country. Its native and cultivated pasture area is equivalent to its agricultural plus planted and natural forest areas (JANK *et al.*, 2011). The greatest cultivated area in the country is represented by *Brachiaria* (Syn. *Urochloa*) spp., where *Brachiaria brizantha* cv. Marandu grass is the predominant forage (JANK *et al.*, 2014). As this cultivar is produced on a large scale, Brazilian pastures are classified as extensive monocultures. Therefore, the use of several species and cultivars is recommended to mitigate problems caused by monoculture, such as break of resistance to known diseases (JANK *et al.*, 2011). *Panicum maximum* (Syn. *Megathyrsus maximum*) Jacq. is an excellent option for diversification and intensification of Brazilian pastures because it is very productive and has excellent nutritive quality, providing high animal production per hectare (JANK *et al.*, 2014).

The main reproduction strategies is apomixis (autotetraploid plants), but sexuality occurs sporadically in diploid plants. Apomixis is a clonal propagation by seeds, in which offsprings genetically identical to the female parent are produced. According to SAVIDAN *et al.* (1989), the apomixis has several advantages in breeding programs, such as hybrid vigor fixation, simplification in obtaining hybrids, and low cost of seed production. However, the main disadvantage is that it does not allow recombination of superior individuals, avoiding the exploration of genetic variability. Thus, the improvement in this species was made possible from the chromosomic duplication of sexual diploid plants and, later, crossing with apomictic plants (SAVIDAN *et al.*, 1989).

Main tropical forage breeding programs in Brazil are at Embrapa Centers, which hold the main germplasm banks in the country. The *P. maximum* forage breeding is mainly coordinated by Embrapa Beef Cattle, with the goals of increasing leaf and seed yield, disease resistance, and nutritive quality (JANK *et al.*, 2011). The breeding program uses recurrent selection methods where each cycle requires three to five years of evaluation. According to RESENDE *et al.* (2014), the process of development, testing, and recommendation of a new cultivars span over approximately fifteen years.

Genomic selection is an effective method to explore genetic variation in breeding programs, from the prediction of breeding values based on markers distributed throughout the genome. It can increase the accuracy of selection, reduce evaluation costs per genotype and get shorter breeding cycles than phenotypic selection (LIPKA *et al.*, 2014; RESENDE *et al.*, 2014). Its potential to increase the efficiency of breeding programs has been shown in several crops (?CROSSA *et al.*, 2013; GOUY *et al.*, 2013; LIPKA *et al.*, 2014). Therefore, the application of genomic selection in forage breeding is promising, since many of the main traits have high assessment costs, as well as evaluation after flowering time in the breeding cycle.

This work consists of a partnership between the Embrapa Beef Cattle (Campo Grande, MS, Brazil), the Graduate Program in Genetics and Plant Breeding at ESALQ / USP (“Luiz de Queiroz” College of Agriculture / University of São Paulo - Piracicaba, SP, Brazil), and the Bioinformatics Research Center at NC State (North Carolina State University - Raleigh, NC, USA). The work was organized in two chapters. The first one has the goal of to evaluate

nutritional, structural, and yield traits in a tetraploid sexual *P. maximum* population using linear mixed models and to estimate genetic parameters and canonical correlation between sets of traits. The second one aims to develop statistical models in genomic selection considering tetraploid allelic dosage for the same population of *P. maximum*.

References

- 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.
- GOUY, M., Y. ROUSSELLE, D. BASTIANELLI, P. LECOMTE, L. BONNAL, D. ROQUES, J.-C. EFILE, S. ROCHER, J. DAUGROIS, L. TOUBI, S. NABENEZA, C. HERVOUET, H. TELISMART, M. DENIS, A. THONG-CHANE, J. C. GLASZMANN, J. Y. HOARAU, S. NIBOUCHE, and L. COSTET, 2013 Experimental assessment of the accuracy of genomic selection in sugarcane. *Theoretical and Applied Genetics* **126**: 2575–2586.
- JANK, L., S. C. BARRIOS, C. B. VALLE, R. M. SIMEÃO, and G. F. ALVES, 2014 The value of improved pastures to Brazilian beef production. *Crop and Pasture Science* **65**: 1132–1137.
- JANK, L., C. B. VALLE, and R. M. S. RESENDE, 2011 Breeding tropical forages. *Crop Breeding and Applied Biotechnology* **11**: 27–34.
- LIPKA, A. E., F. LU, J. H. CHERNEY, E. S. BUCKLER, M. D. CASLER, and D. E. COSTICH, 2014 Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. *PLoS ONE* **9**: e112227.
- RESENDE, R. M. S., M. D. CASLER, and M. D. V. DE RESENDE, 2014 Genomic selection in forage breeding: Accuracy and methods. *Crop Science* **54**: 143–156.
- SAVIDAN, Y. H., L. JANK, J. C. G. COSTA, and C. B. D. VALLE, 1989 Breeding *Panicum maximum* in Brazil. 1. Genetic resources, modes of reproduction and breeding procedures. *Euphytica* **41**: 107–112.

2 CONCLUSION

Recurrent selection (RS) is an approach that aims to improve breeding populations by increasing the frequency of favorable alleles, while maintaining genetic variability (HALLAUER and CARENA, 2012). In perennial crops, such as *Panicum maximum*, the evaluation process is performed by repeated measures in the same individual, in order to infer the genotype's ability to repeat its performance over successive evaluations (BRAZ *et al.*, 2015). Analysis of mixed models allows modeling different structures of variance and covariance (VCOV) to investigate different assumptions of heterogeneity and dependence between observations (CROSSA *et al.*, 2006). In addition, these analyzes allow the inclusion of information from relatives, exploring the genetic correlation between individuals by an additive relationship matrix estimated by pedigree values or, when available, estimated by molecular markers.

In the first chapter, three classes of models were considered to evaluate nutritional, structural, and yield traits. For all traits, models of the class that consider different VCOV matrices and include pedigree information had a better fit. Therefore, incorporating the additive relationship matrix and modeling the genotype by harvest interaction with VCOV structures allows better selection of individuals of the breeding population.

Genomic selection (MEUWISSEN *et al.*, 2001) is an approach that can accelerate future genetic progress in plant breeding (HEFFNER *et al.*, 2010), as well as accelerate recurrent selection cycles (LIPKA *et al.*, 2014). Although some crops have benefited from this methodology, studies in autotetraploids (and polyploids in general) are still modest. Thus, the main objective of the second chapter was to develop models that consider tetraploid allelic dosage to estimate genetic values for genomic selection in *P. maximum*, an autotetraploid forage. These models can be applied to other autotetraploid species, as well as extended to species with other ploidy level.

Therefore, we compared six different GS models and presented strategies for modeling phenotypic residual effects and performing SNP calling in autotetraploids. The GBS and allelic dosage strategies shown to be promising for genomic analyzes in *P. maximum* and the accuracy of predictive models justifies the implementation of this approach in breeding programs.

References

- BRAZ, T. G. D. S., D. M. D. FONSECA, L. JANK, C. D. CRUZ, and J. A. MARTUSCELLO, 2015 Repeatability of agronomic traits in *Panicum maximum* (Jacq.) hybrids. *Genetics and Molecular Research* **14**: 19282–19294.
- CROSSA, J., J. BURGUEÑO, P. L. CORNELIUS, G. McLAREN, R. TRETHOWAN, and A. KRISHNAMACHARI, 2006 Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science* **46**: 1722–1733.
- HALLAUER, A. R. and M. J. CARENA, 2012 Recurrent selection methods to improve germplasm in maize. *Maydica* **57**: 266–283.
- HEFFNER, E. L., A. J. LORENZ, J.-L. JANNINK, and M. E. SORRELLS, 2010 Plant breeding with genomic selection: Gain per unit time and cost. *Crop Science* **50**: 1681–1690.

- LIPKA, A. E., F. LU, J. H. CHERNEY, E. S. BUCKLER, M. D. CASLER, and D. E. COSTICH, 2014 Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. PLoS ONE **9**: e112227.
- MEUWISSEN, T. H. E., B. J. HAYES, and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.