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 Melhoramento 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. Sílvio 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 Evellyn 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 deiscentes
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 . . . 8
Abstract . . . 9
1 Introduction . . . 11
References . . . 12
2 Literature Review . . . 13
2.1 *Panicum maximum* Breeding . . . 13
2.2 Molecular Markers . . . 15
2.3 Allelic Dosage Information in Polyploids . . . 16
2.4 Genomic Selection . . . 17
2.5 Statistical Models in Genomic Selection . . . 18
References . . . 19
3 Mixed Modeling and Genetic Parameters Estimation for Forage Traits in *Panicum maximum* 25
3.1 Abstract . . . 25
3.2 Introduction . . . 25
3.3 Material and Methods . . . 27
3.3.1 *Panicum maximum* population . . . 27
3.3.2 Experimental data . . . 27
3.3.3 Longitudinal multivariate linear mixed analysis . . . 28
3.4 Results . . . 31
3.4.1 Model selection . . . 31
3.4.2 Genetic parameters . . . 31
3.4.3 Genotype x harvest interaction . . . 33
3.5 Discussion . . . 33
References . . . 36
4 Development of Statistical Models Including Dosage Information for Genomic Selection in *Panicum maximum* . . . 41
4.1 Abstract . . . 41
4.2 Introduction . . . 41
4.3 Material and Methods . . . 43
4.3.1 *Panicum maximum* population . . . 43
4.3.2 Phenotypic evaluation and longitudinal multivariate linear mixed analysis . 43
4.3.3 Molecular data . . . 44
4.3.4 Genomic prediction models considering dosage . . . 45
4.3.5 Model evaluation . . . 47
4.4 Results . . . 48
4.4.1 Phenotypic models . . . 48
4.4.2 SNP calling . . . 48
4.4.3 Linkage disequilibrium . . . 52
4.4.4 Genomic prediction . . . 53
4.5 Discussion . . . 54

References 56
Appendix 61

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 sucessive 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 (Heffner *et al.*, 2010; 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.

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.

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.

Martuscello, J. A., L. Jank, D. M. Da Fonseca, C. D. a. Cruz, and D. D. N. F. V. Da Cunha, 2007 Repetibilidade de caracteres agronômicos em *Panicum maximum* Jacq. Revista Brasileira de Zootecnia **36**: 1975–1981.

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 LITERATURE REVIEW

2.1 *Panicum maximum* Breeding

Guinea grass (*Panicum maximum* Jacq.) is a tropical perennial grass which stands out among the forage species. It is adapted for light soils, from medium to high fertility, and recommended for more intensive systems of livestock farming (Valle *et al.*, 2009; Jank *et al.*, 2014). Although this species presents vegetative propagation, cultivars produced by seeds are easier to establish, faster to be adopted and much more widespread (Jank *et al.*, 2011).

Panicum maximum belongs to genus *Panicum* L., which is one of the most important within the family Poaceae, subfamily Panicoideae, and tribe Paniceae. This genus comprises more than 500 species, distributed in tropical and subtropical areas of several countries, mainly in Africa. In Brazil, 114 species have been reported (Guglieri *et al.*, 2004). Its center of origin is in Tropical Africa. Among economically important species of this genus are *P. miliaceum* L., *P. virgatum* L., *P. prupurascens* Raddi, and *P. maximum* Jacq (Warmke, 1951; Jank *et al.*, 2008).

The chromosome number in *P. maximum* is $2n = 4x = 32$, being characterized as an autotetraploid species. However it can also be found, in low frequency in nature, as a diploid with chromosome number $2n = 2x = 16$. Most species of the genus *Panicum* have a basic number of $x = 9$, but *P. maximum* has $x = 8$. The DNA content per complete chromosome complement (n) was determined by Akiyama *et al.* (2008) to be approximately 500 Mbp for diploid and 1000 Mbp for tetraploid. Therefore, the DNA content per monoploid chromosome set (x) is approximately 500 Mbp, which suggests that *P. maximum* possesses the smallest genome size in any reported *Panicum* species.

Diploid *Panicum* plants have sexual reproduction whereas tetraploid plants reproduce by apomixis via apospory. In the apospory, meiosis is replaced by mitosis, with formation of an embryo sac and an embryo with the same mother's genotype, that is, the embryo sac is originated from a mitotic division of a somatic cell of the ovum (nucelus or integument) (Savidan *et al.*, 1989; Savidan, 2000; Resende *et al.*, 2008; Jank *et al.*, 2011).

There are several advantages of apomixis, such as: hybrid vigor fixation, simplification in obtaining hybrids, and low costs of hybrid seed production (Savidan *et al.*, 1989). Because there are no apomictic diploid plants, apomixis is associated with polyploidy in this species (Savidan, 2000). Thus, one way to enable genetic breeding is to select diploid plants and, duplicate its chromosome numbers with colchicine, to cross with apomictic tetraploid plants (Savidan *et al.*, 1989). In these crosses, sexual plants are used as females and apomictic ones as donors of pollen.

Apomixis has been determined by a dominant gene or group of genes with simple inheritance (Savidan *et al.*, 1989). Assuming that gene A controls apomixis, Aaaa genotype is associated with apomictic plants and aaaa genotype is associated with sexual plants. Thus, hybrid progenies of these crosses will always segregate in the ratio 1 sexual: 1 apomictic (Savidan *et al.*, 1989). Genetic mapping of apomixis was performed for tropical grasses *P. maximum* (Ebina *et al.*, 2005), *Paspalum notatum* (Stein *et al.*, 2007), and *Brachiaria humidicola* (Vigna *et al.*, 2016), being the apospory mapped in a single linkage group for all species.

The first tropical forage grass to be extensively collected in its center of origin was *P. maximum*, in the late 1960s. The second one was *Brachiaria* spp. in the 1980s (Euclides *et al.*, 2010; Jank *et al.*, 2008). These collections were made with the main purpose of studying the inheritance of apomixis and reproductive strategies of tropical forages. Therefore, tropical grasses are in the early stages of domestication and breeding. *Panicum maximum* was collected in 1967 and 1969 by the French Institution ORSTOM (*Institut Français de Recherche Scientifique pour le Développement en Coopération* - IRD) in East Africa, specifically in Tanzania and Kenya. In 1982, Embrapa Beef Cattle, in Campo Grande, Brazil, signed an agreement with ORSTOM, which allowed the transference to Brazil of 426 apomictic accessions and 417 sexual plants (Savidan *et al.*, 1989). Few years have elapsed since the collection and availability of the germplasm banks to initiate breeding programs.

In Brazil, the first cultivar of *P. maximum* introduced was Colonião, at the time of slavery. It became very well adapted to environmental conditions. Several other cultivars were introduced later, such as Sempre Verde, Guiné, Makueni and Tobiatã. Other cultivars were developed in the country, as a result of selection and breeding programs, in which the most notable were: Centenário, Centauro, Vencedor, Tanzânia, Mombaça, Aruana, Áries, Atlas and Massai (Valle *et al.*, 2009).

The area occupied by this species corresponds to approximately 20% of all cultivated pasture area (around 20 million of hectares), supplying 30% of the forage seed market (Martuscello *et al.*, 2007). Although the number of cultivars available has increased recently, Brazilian pastures can still be characterized as large clonal monocultures, genetically poor and vulnerable to pests and diseases, which makes it fundamental to invest in breeding programs (Valle *et al.*, 2009).

In general, the goals of forage breeding programs are similar to those of other crops, such as increase in productivity, pests and disease resistance, good quality of seed production, efficient use of fertilizers, and adaptation to edaphic and climatic stresses. However, breeding programs still have the additional purpose of animal use. Thus, forage value is measured indirectly by being converted into animal products (such as meat, milk, leather and furs) (Resende *et al.*, 2008; Jank *et al.*, 2011).

Forage breeding programs are traditionally carried out by conventional breeding methods, in which genotypes are selected through several crosses, taking more than a decade. These methods requires three to five years for each cycle of evaluation and, in total, approximately fifteen years are necessary for development, testing and releasing of new cultivars (Resende *et al.*, 2014). Different methods that use genome analysis, such as genomic mapping and marker-assisted selection, have been developed and are widely used in order to improve the eficiency of breeding programs. One promising approach is genomic selection, which can be used to accelerate breeding cycles and increase the accuracy of selection. The implementation of genomic selection in forage breeding was initially presented by Hayes *et al.* (2013).

Posteriorly, Lipka *et al.* (2014) applied genomic selection in *Panicum virgatum* L., with the main objective to evaluate genomic selection efficiency to accelerate breeding cycles. The authors obtained high prediction accuracy analyzing seven morphological characters and thirteen characters related to biomass quality. *Panicum virgatum* is considered a reference

genome for studies with the genus *Panicum*, which means that this research is very important for the genus, as well for *P. maximum*. The authors hope that implementation of genomic selection approaches for breeding programs of *Panicum* spp. will be more advantageous than when compared to traditional breeding programs.

2.2 Molecular Markers

Genetic marker is any inheritable trait that allows to distinguish variations in the genome of different individuals. They are the basis of studies for genome analysis and fundamental for identification of desirable genotypes. New genotyping technologies have been developed and are becoming increasingly important in breeding programs. The main goal is to achieve the required qualities and quantities of markers for a variety of applications in molecular studies. In applications that require extensive genome coverage, the ideal technique should offer not only thousands of molecular markers distributed throughout the genome, but also allow these markers to be obtained preferentially in a single, reliable, and low-cost experiment (Luikart *et al.*, 2003).

With the development of Next Generation Sequencing (NGS), new techniques have been able to discover, sequence, and genotype thousands of markers in any genome of interest in a single step, even in populations where there is little or no genetic information. One promising technique is Genotyping-by-sequencing (GBS) that uses the reduction of genome complexity using specific combinations of restriction enzymes (Elshire *et al.*, 2011). These enzymes produce DNA fragments with cohesive borders, cutting the DNA in non-repetitive regions. These fragments are linked to barcode that allows the identification of sequences generated in each sample. This approach has shown to be consistent in several species and efficient to produce hundreds of thousands of molecular markers (Elshire *et al.*, 2011; Poland *et al.*, 2012a).

High density markers are fundamental in genomic selection. Thus, it is expected that most of QTLs will be in linkage disequilibrium (LD) with at least one marker in the population. In this way, the GBS technique shows promising results in several papers published in the literature (Poland *et al.*, 2012b; Crossa *et al.*, 2013; Lipka *et al.*, 2014; Annicchiarico *et al.*, 2017).

Working with a set of 254 wheat lines, Poland *et al.* (2012b) used GBS approach and obtained 41,371 genotyped SNPs. The authors evaluated four different methods for imputing missing data, compared the performance of two types of markers (GBS and DArT-arrays), and analyzed four phenotypic traits (yield in drought and irrigated conditions, thousand kernel weight, and days to heading). They concluded that GBS technique can be used to generate a large number of markers at an accessible price in addition to allow the development of more accurate genomic selection models.

GBS markers were also used for genomic prediction in maize populations by Crossa *et al.* (2013). The authors analyzed two experiments, the first consisted of 504 double-haploid lines, and the second formed by 296 inbred lines. Three relevant situations for genomic selection were investigated: i) gain in accuracy when using nonimputed, imputed, and GBS-inferred haplotypes methods; ii) accuracy of pedigree models; iii) comparison between parametric and non-parametric models. Their results indicated a slight difference in the level of accuracy for

imputed and nonimputed data, as well as for parametric and non-parametric models. The prediction of maize lines incorporating pedigree information showed better results.

2.3 Allelic Dosage Information in Polyploids

Genetic and statistical analysis of polyploid species have been based basically on the idea of single dose with 1:1 segregation in biparental populations, proposed initially by Wu *et al.* (1992). Currently, new possibilities have arisen to enable accessibility of the complex polyploid genomes for several crops due to advantages in genetic and statistical methods as well as the development of NGS technologies. Thus, the evaluation of Single-Nucleotide Polymorphisms (SNP) throughout the genome allows one to assess the relative abundance of each allele; in other words, to estimate the allelic dosage of SNPs (Serang *et al.*, 2012; Garcia *et al.*, 2013; Mollinari and Serang, 2013).

Diploid dosage is well established in the literature and constitutes three dose classes: nulliplex (aa), simplex (Aa), and duplex (AA), being classified according to reference allele. For tetraploid species, doses are classified into five classes: nulliplex (aaaa), simplex (Aaaa), duplex (AAaa), triplex (AAAa), and quadriplex (AAAA). With increase of ploidy, the number of dose classes increases up to ploidy+1. In spite of the advances in breeding programs and genotyping techniques, genetic studies in polyploids have been limited to the use of markers with only diploid dosage (Poland *et al.*, 2012b; Lipka *et al.*, 2014; Li *et al.*, 2015; Ramstein *et al.*, 2016; Biazzi *et al.*, 2017).

Initially, Rickert *et al.* (2002) reported the use of pyrosequencing^{TM} in polyploids to distinguish different heterozygous states, albeit with some sequence-specific limitations. Later, Bérard *et al.* (2009) and Akhunov *et al.* (2009) used SNPlex^{TM} and Illumina Golden Gate^{TM} assays, respectively, for the genotyping of polyploid wheat. For genotype calling in tetraploid species, Voorrips *et al.* (2011) developed an algorithm, implemented in the R package fitTetra, using mixture models. However, they assumed Hardy–Weinberg equilibrium within the population, which may not occur in all segregating polyploid progeny. Then, Serang *et al.* (2012) presented an algorithm for finding the exact maximum a posteriori (MAP) genotype configuration, in the software SuperMASSA, that allows the classification of the allelic dosage for any ploidy level even when the ploidy is unknown. Furthermore, the population can assume three presuppositions: Hardy-Weinberg model, F_1 model, or Generalized Population model.

A few genetic studies in polyploids have used allelic dosage, such as genetic mapping (with limited segregation of the markers) (Hackett *et al.*, 2013; Massa *et al.*, 2015; Costa *et al.*, 2016; Vigna *et al.*, 2016), genome-wide association (Rosyara *et al.*, 2016), and genomic selection (Slater *et al.*, 2016). Recently, Slater *et al.* (2016) used tetraploid dosage in genomic selection studies with potato. The authors achieved accuracies ranging from 0.2, under conditions of low heritability and small reference populations, to 0.8 in larger reference populations.

The dosage information allows to investigate about its importance to the gain in accuracy of the statistical models in autopolyploid species when all genotypic information is used (Garcia *et al.*, 2013). Therefore, the inclusion of correct allelic dosage is essential for genetic studies in polyploid species.

2.4 Genomic Selection

Genomic Selection (GS) is an effective method to perform predictions of genomic breeding values on genotypes evaluated in breeding programs. Initially proposed by Meuwissen *et al.* (2001), it uses statistical methods to predict these breeding values from markers distributed throughout the genome, with sufficient accuracy to represent the phenotype. GS was superior to former methods such as the traditional method of Marker-Assisted Selection (MAS), mainly for selection of polygenic traits, which are controlled by many loci of small effects (Heffner *et al.*, 2009). GS is different than MAS because it analyzes all markers simultaneously, including both minor and major marker effects, in a population with genome wide coverage of markers. Moreover, this method calculates the genomic estimated breeding values (GEBVs) of individuals (Meuwissen *et al.*, 2001; Bernardo, 2014).

In summary, GS uses a training population, formed by genotyped and phenotyped individuals, to develop a statistical model that estimates GEBVs from genotipic data of a candidate population of untested individuals (Jannink *et al.*, 2010). In this way, the GEBVs are used to select individuals in another population (genotyped only). In order to maximize the GEBV accuracy, the training population must be representative of the selection candidate population (Heffner *et al.*, 2009).

It is common to find the training population splitted in training and validation populations (Resende *et al.*, 2012; Habier *et al.*, 2011), and the candidate population defined as selection population. Thus, according to Resende *et al.* (2012), the three populations are characterized as:

- **Training population:** In this population, individuals are genotyped and their phenotypes evaluated for traits of interest. It can also be called as discovery population or estimation population. Prediction equations of genomic breeding values (GBVs) associate each marker with its effect on the trait of interest. The accuracy of genomic prediction is expected to increase with increasing the size of the population.

- **Validation population:** In this population, individuals are also phenotyped and genotyped. It may be called testing population. The effects of markers, estimated from training population, are used to predict the phenotypes of this population. Thus, the predictive capacity of genomic selection is verified by correlating the observed phenotypic values against the predicted ones. As the validation population was not involved in the prediction of marker effects, the GBVs errors and the phenotypic values are independent. Therefore, this correlation is predominantly genetic and it is equivalent to the predictive ability of genomic selection to estimate the phenotypes.

- **Selection population:** It consists of the population in which individuals will be selected in the breeding program. Individuals will only be genotyped and evaluated through the prediction and selection of GEBVs, using the models that allowed higher accuracy in the prediction of these values.

The success of GS depends of the predictive accuracy to select individuals whose phenotypes are not evaluated. Thus, the existence of a direct relationship between the training

population and selection population is fundamental. According to Desta and Ortiz (2014), the main factors that affect prediction accuracy are: marker density, training population size and relatedness, heritability and genetic architecture of traits, performance of analyzed models, gene effects, and extent and distribution of LD between markers and QTL.

The increase in marker density guarantees the conservation of marker-QTL associations and allows a high predictive accuracy (Desta and Ortiz, 2014). Several authors report that the key to the success of genomic selection is to incorporate all markers in the prediction models, in order to maximize the number of QTLs in linkage disequilibrium (LD) with at least one marker (Heffner *et al.*, 2009; Lorenz *et al.*, 2011; Resende *et al.*, 2012; Desta and Ortiz, 2014). In this way, the number of QTLs whose effects will be captured by markers is maximized, obtaining greater accuracy and avoiding biases in the estimation of marker effects. However, according to Muir (2007), this increase in marker density must be accompanied with a larger training population size in order to decrease the colinearity among markers.

When the genome coverage by molecular markers is sufficiently informative, training population size has more influence on the increase of predictive accuracy (Lorenzana and Bernardo, 2009). If the training population size is large enough, even low heritability traits can be predicted more accurately.

Models performance will vary according to assumptions and considerations about marker effects and genetic effects that control the trait to be analyzed. Nonlinear models can capture non-additive genetic effects, *i.e.*, dominance and epistatic effects, making them capable of improving the GS accuracy (Kumar *et al.*, 2012; Sun *et al.*, 2012). However, if the traits are purely additive, these models may not produce the expected result, reducing the accuracy (Zhao *et al.*, 2013). Thus, the construction of stable prediction models, which correctly evaluate the gene effects and contemplate all or a large part of the estimation of marker effects, is fundamental for implementation of GS.

2.5 Statistical Models in Genomic Selection

In genomic selection, the number of markers (p) used is generally greater than the number of individuals (n). When this happens, estimates using ordinary least squares (OLS) have high variance and high mean squared error. Therefore, OLS models have a low predictive capacity, because the marker effects are treated as fixed effects, causing multicollinearity among predictors.

Several alternative models have been proposed and the most used in genomic selection can be divided into three main classes:

- **Linear and non-linear regression:** This group includes Penalized and Bayesian methods. The first group uses linear regression and the second one uses non-linear regression. Within Penalized methods are Ridge Regression Best Linear Unbiased Predictor (RR-BLUP), Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net (EN). Within Bayesian methods are Bayesian Ridge Regression (BRR), BayesA, BayesB, BayesC, Bayesian Lasso (BL), and weighted Bayesian Shrinkage Regression (wBSR).

- **Implicit regression:** This group includes semi-parametric and non-parametrtic methods, such as Reproducing Kernel Hilbert Spaces (RKHS) and Neural Networks.
- **Reduced-dimension regression:** In this group, the main methods are: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR).

Penalized estimation methods differ according to penalization functions, which produce different degrees of shrinkage. The main idea is to reduce the mean square error, reducing the estimator variance, and to prevent the super-parameterization of the model (Resende *et al.*, 2012).

Bayesian methods also promote a shrinkage of the model effects, but from the *a priori* distribution assumed for these effects. These methods provide better predictions when the effects of QTLs are not normally distributed because they are associated with nonlinear equations. Bayesian methods overcome penalized estimations when the distribution of QTL effects is leptokurtic (positive kurtosis), due to the presence of large effects genes (Meuwissen *et al.*, 2001; Resende *et al.*, 2012). When the distribution is normal, both methods tend to be equally efficient (Resende *et al.*, 2012).

Implicit regression methods are an alternative to adjust models with many epistatic and dominance interactions. Thus non-parametric regressions are functional representations between a large number of covariates and a dependent variable, generating a less parameterized structure able to accommodate the interactions effects with fewer assumptions (Gianola *et al.*, 2006; Resende *et al.*, 2012). On the other hand, reduced-dimension methods can be applied to marker selection with significant effects on the trait.

The comparison among these prediction methods has been carried out in several studies, such as Gianola *et al.* (2006), de los Campos *et al.* (2009), Heslot *et al.* (2012), Gouy *et al.* (2013), and others. There is no consensus which is the most efficient method because it varies with different species, population, and analyzed traits. Besides, in pratical experiments, the difference among approaches has remained small.

References

Akhunov, E., C. Nicolet, and J. Dvorak, 2009 Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theoretical and Applied Genetics **119**: 507–517.

Akiyama, Y., H. Yamada-Akiyama, H. Yamanouchi, M. Takahara, M. Ebina, T. Takamizo, S.-i. Sugita, and H. Nakagawa, 2008 Estimation of genome size and physical mapping of ribosomal DNA in diploid and tetraploid guineagrass (*Panicum maximum* Jacq.). Japanese Society of Grassland Science **54**: 89–97.

Annicchiarico, P., N. Nazzicari, L. Pecetti, M. Romani, B. Ferrari, Y. Wei, and E. C. Brummer, 2017 GBS-based genomic selection for pea grain yield under severe terminal drought. The Plant Genome **10**: 1–13.

Bernardo, R., 2014 Genomewide selection when major genes are known. Crop Science **54**: 68–75.

Biazzi, E., N. Nazzicari, L. Pecetti, E. C. Brummer, A. Palmonari, A. Tava, and P. Annicchiarico, 2017 Genome-wide association mapping and genomic selection for alfafa (*Medicago sativa*) forage quality Traits. PLoS ONE **12**: e0169234.

Bérard, A., M. C. Le Paslier, M. Dardevet, F. Exbrayat-Vinson, I. Bonnin, A. Cenci, A. Haudry, D. Brunel, and C. Ravel, 2009 High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.). Plant Biotechnology Journal **7**: 364–374.

Costa, E. A., C. O. Anoni, M. C. Mancini, F. R. C. Santos, T. G. Marconi, R. Gazaffi, M. M. Pastina, D. Perecin, M. Mollinari, M. A. Xavier, L. R. Pinto, A. P. Souza, and A. A. F. Garcia, 2016 QTL mapping including codominant SNP markers with ploidy level information in a sugarcane progeny. Euphytica **211**: 1–16.

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.

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**: 375–385.

Desta, Z. A. and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science **19**: 592–601.

Ebina, M., H. Nakagawa, T. Yamamoto, H. Araya, S.-i. Tsuruta, M. Takahara, and K. Nakajima, 2005 Co-segregation of AFLP and RAPD markers to apospory in Guineagrass (*Panicum maximum* Jacq.). Japanese Society of Grassland Science **51**: 71–78.

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.

Euclides, V. P. B., C. B. do Valle, M. C. M. Macedo, R. G. d. Almeida, D. B. Montagner, and R. A. Barbosa, 2010 Brazilian scientific progress in pasture research during the first decade of XXI century. Revista Brasileira de Zootecnia **39**: 151–168.

Garcia, A. A. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva, M. L. C. Vieira, R. Vicentini, E. A. Costa, M. C. Mancini, M. O. S. Garcia, M. M. Pastina, R. Gazaffi, E. R. F. Martins, N. Dahmer, D. A. Sforça, C. B. C. Silva, P. Bundock, R. J. Henry, G. M. Souza, M.-A. van Sluys, M. G. A. Landell, M. S. Carneiro, M. A. G. Vincentz, L. R. Pinto, R. Vencovsky, and A. P. Souza, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific reports **3**: 1–10.

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics **173**: 1761–1776.

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.

Guglieri, A., F. O. Zuloaga, and H. M. Longhi-Wagner, 2004 Sinopse das espécies de *Panicum* L. subg. *Panicum* (Poaceae, Paniceae) ocorrentes no Brasil. Acta Botanica Brasilica **18**: 359–367.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**: 186–197.

Hackett, C. A., K. McLean, and G. J. Bryan, 2013 Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. PLoS ONE **8**: e63939.

Hayes, B. J., N. O. I. Cogan, L. W. Pembleton, M. E. Goddard, J. Wang, G. C. Spangenberg, and J. W. Forster, 2013 Prospects for genomic selection in forage plant species. Plant Breeding **132**: 133–143.

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.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic selection for crop improvement. Crop Science **49**: 1–12.

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.

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., R. M. S. Resende, C. B. Valle, M. D. V. Resende, L. Chiari, L. J. Cançado, and C. Simioni, 2008 Melhoramento genético de *Panicum maximum*. In *Melhoramento de forrageiras tropicais*, edited by R. M. S. Resende, C. B. Valle, and L. Jank, pp. 55–87.

Jank, L., C. B. Valle, and R. M. S. Resende, 2011 Breeding tropical forages. Crop Breeding and Applied Biotechnology **11**: 27–34.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics **9**: 166–177.

Kumar, S., M. C. A. M. Bink, R. K. Volz, V. G. M. Bus, and D. Chagné, 2012 Towards genomic selection in apple (*Malus* × *domestica* Borkh.) breeding programmes: Prospects, challenges and strategies. Tree Genetics & Genomes **8**: 1–14.

Li, X., Y. Wei, A. Acharya, J. L. Hansen, J. L. Crawford, D. R. Viands, R. Michaud, A. Claessens, and E. C. Brummer, 2015 Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. The Plant Genome **8**: 1–10.

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.

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi, H. Iwata, K. P. Smith, M. E. Sorrells, and J.-L. Jannink, 2011 Genomic selection in plant breeding: Knowledge and Prospects. In *Advances in Agronomy*, edited by D. L. Sparks, volume 110, pp. 77–123, Elsevier Inc.

Lorenzana, R. E. and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and Applied Genetics **120**: 151–161.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003 The power and promise of population genomics: from genotyping to genome typing. Nature reviews. Genetics **4**: 981–994.

Martuscello, J. A., L. Jank, D. M. Da Fonseca, C. D. a. Cruz, and D. D. N. F. V. Da Cunha, 2007 Repetibilidade de caracteres agronômicos em *Panicum maximum* Jacq. Revista Brasileira de Zootecnia **36**: 1975–1981.

Massa, A. N., N. C. Manrique-Carpintero, J. J. Coombs, D. G. Zarka, A. E. Boone, W. W. Kirk, C. A. Hackett, G. J. Bryan, and D. S. Douches, 2015 Genetic linkage mapping of economically important traits in cultivated tetraploid potato (*Solanum tuberosum* L.). G3 (Genes|Genomes|Genetics) **5**: 2357–2364.

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.

Mollinari, M. and O. Serang, 2013 Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. In *Methods in Molecular Biology*, edited by J. M. Walker.

Muir, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics **124**: 342–355.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink, 2012a Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE **7**: e32253.

Poland, J. A., 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 **5**: 103–113.

Ramstein, G. P., J. Evans, S. M. Kaeppler, R. B. Mitchell, K. P. Vogel, C. R. Buell, and M. D. Casler, 2016 Accuracy of genomic prediction in switchgrass (*Panicum virgatum* L.) improved by accounting for linkage disequilibrium. G3 (Genes|Genomes|Genetics) **6**: 1049–1062.

Resende, M. D. V., R. M. S. Resende, L. Jank, and C. B. Valle, 2008 Experimentação e análise estatística no melhoramento de forrageiras. In *Melhoramento de forrageiras tropicais*, edited by R. M. S. Resende, C. B. Valle, and L. Jank, pp. 195–287.

Resende, M. D. V. D., F. F. Silva, P. S. Lopes, and C. F. Azevedo, 2012 Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial. p. 291.

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.

Rickert, A. M., A. Premstaller, C. Gebhardt, and P. J. Oefner, 2002 Genotyping of SNPs in a polyploid genome by Pyrosequencing. In *BioTechniques*, volume 32, pp. 592–603.

Rosyara, U. R., W. S. De Jong, D. S. Douches, and J. B. Endelman, 2016 Software for genome-wide association studies in autopolyploids and its application to potato. The Plant Genome **9**: 1–10.

Savidan, Y. H., 2000 Apomixis: genetics and breeding. Plant Breeding Reviews **18**: 10–86.

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.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE **7**: e30906.

Slater, A. T., N. O. I. Cogan, J. W. Forster, B. J. Hayes, and H. D. Daetwyler, 2016 Improving genetic gain with genomic selection in autotetraploid potato. The Plant Genome **9**: 1–15.

Stein, J., S. C. Pessino, E. J. Martínez, M. P. Rodriguez, L. A. Siena, C. L. Quarin, and J. P. A. Ortiz, 2007 A genetic map of tetraploid *Paspalum notatum* Flügge (bahiagrass) based on single-dose molecular markers. Molecular Breeding **20**: 153–166.

Sun, X., P. Ma, and R. H. Mumm, 2012 Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. PLoS ONE **7**: e50604.

Valle, C. B. D., L. Jank, and R. M. S. a. Resende, 2009 O melhoramento de forrageiras tropicais no Brasil. Revista Ceres **56**: 460–472.

Vigna, B. B. Z., J. C. S. Santos, L. Jungmann, C. B. Valle, M. Mollinari, M. M. Pastina, M. S. Pagliarini, A. A. F. Garcia, and A. P. Souza, 2016 Evidence of allopolyploidy in *Urochloa humidicola* based on cytological analysis and genetic linkage mapping. PLoS ONE **11**: e0153764.

Voorrips, R. E., G. Gort, and B. Vosman, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics **12**: 1–11.

Warmke, H. E., 1951 Cytotaxonomic investigations of some varieties of *Panicum maximum* and *P. purpurascens* in Puerto Rico. Agronomy Journal **43**: 143–149.

Wu, K. K., W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore, and S. D. Tanksley, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics **83**: 294–300.

Zhao, Y., J. Zeng, R. Fernando, and J. C. Reif, 2013 Genomic prediction of hybrid wheat performance. Crop Science **53**: 802–810.

3 MIXED MODELING AND GENETIC PARAMETERS ESTIMATION FOR FORAGE TRAITS IN *PANICUM MAXIMUM*

Keywords: Forage Breeding; Guinea Grass; Variance and Covariance Structures; Pedigree Information; Genotype x Harvest Interaction

3.1 Abstract

The production of apomictic guinea grass (*Panicum maximum* Jacq.) hybrid cultivars depends on the availability of improved sexual parents over time, since these hybrids are obtained from sexual × apomictic crosses. A promising strategy to increase the frequency of favorable alleles in sexual parents is by intrapopulation recurrent selection (IRS), which consists of improving the population by successive cycles of evaluation, selection and recombination of sexual plants. Since *P. maximum* is a perennial crop, repeated measures of a given trait are taken in the same individual and linear mixed models can be used to increase selective efficiency and provide more realistic estimates of genetic parameters. The objective of this study was to evaluate nutritional, structural, and yield traits in an outcrossing *P. maximum* multi-parent population of the IRS program of Embrapa Beef Cattle, comparing three different classes of models. Also, we estimated genetic parameters and investigated the interrelationships between these traits. The population consisted of 570 tetraploid sexual genotypes evaluated in an augmented block design, with three checks and six blocks. Yield traits were: total green matter (TGM), total dry matter (TDM), leaf dry matter (LDM), and stem dry matter (SDM); structural trait was: percentage of leaf blade (PLB); and nutritive values were: acid detergent fiber (ADF), crude protein (CP), and in vitro digestibility of organic matter (IVD). The first and second group were analyzed for eight harvests, and the third, for four. We investigated three different classes of models: genetic effects by compound symmetry matrix, genotypes were not correlated, and residual effects were homocedastic and not correlated (Class A); genetic and residual effects with different variance and covariance (VCOV) structures and genotypes not correlated (Class B); genetic and residual effects with different VCOV structures and genotypes correlated by an additive relationship matrix for autotetraploids based on pedigree values (Class C). Correspondence among the 20 best offsprings selected by the different models was evaluated by coincidence index. Class C was selected for all traits, showing that the use of an additive relationship matrix in addition to selected VCOV structures had a goodness of fit. Generalized measure of heritability ranged from 30.65% (IVD) to 87.54% (LDM). In general, traits with higher heritability also had lower divergence between classes and, consequentely, higher coincidence of the 20 best offsprings selected. Classes B and C were, in general, more concordant. Given the best fit of the model besides the difference in ranking between the greater 20 offsprings for the classes, Class C is preferred for selection of superior individuals in an IRS program.

3.2 Introduction

Guinea grass (*Panicum maximum* Jacq.) is one of the most important tropical forage grasses due to its high productivity and good forage quality. It has a wide adaptation on tropical areas, mainly in those of light soils, from medium to high fertility and it is recommended for

more intensive systems of livestock farming (Valle *et al.*, 2009; Jank *et al.*, 2014). This species belongs to genus *Panicum* L., family Poaceae, subfamily Panicoideae, and tribe Paniceae; its chromosome number is $2n = 4x = 32$, being characterized as autotetraploid, but in the nature it can also be found in a low frequency as diploid, with $2n = 2x = 16$. Whereas diploid plants have sexual reproduction, tetraploid plants reproduce by apomixis via apospory. According to Savidan (2000), apomixis is associated with polyploidy, so there are no apomictic diploid plants in nature. Apomixis allows to perpetuate superior genotypes with great precision, although not allowing to achieve genetic recombination by crossing. Thus, to enable genetic breeding in this species, diploid plants are selected and, using colchicine, their chromosome numbers are duplicated to be used in crosses with apomictic tetraploid plants. In these crosses, sexual plants are used as females and apomictic as donors of pollen, and hybrid progenies will always segregate in the ratio 1 sexual: 1 apomictic (Savidan *et al.*, 1989).

Currently, most forage breeding programs are traditionally carried out by conventional breeding, in which the main adopted method is the intrapopulation recurrent selection (IRS). IRS strategy, in apomictic reproducing species, consists of improving the sexual population by recurrent selection cycles; sexual plants are used to obtain the next recurrent selection cycle and/or used in crossings with superior apomictic accessions. These strategies should provide a fast development of improved cultivars, mainly through the exploitation of apomixis in the superior hybrids (Resende *et al.*, 2004). Many goals of forage breeding programs are common to those of other crops, such as increase of productivity, pest and disease resistance, good quality of seed production, efficient use of fertilizers, and adaptation to edaphic and climatic stresses; however, they still have the additional purpose of animal use. Thus, the improvement of forage yield and quality have direct benefits to the farmers by improving animal performance (such as meat, milk, leather, and calves) (Resende *et al.*, 2008; Jank *et al.*, 2011).

In the plant evaluation process, especially in perennial plants, it is common to take repeated measures of a given trait in the same individual. This type of evaluation aims to infer a genotype's ability to repeat its performance over successive evaluations (Braz *et al.*, 2015). In forage experimental designs, half-sibs progenies are commonly evaluated during several harvests and the repeatability of genotypes in successive harvests is estimated by traditional analysis of variance models (Braz *et al.*, 2013, 2015; Fernandes *et al.*, 2017) or by mixed models with variance homogeneity and absence of correlation (Resende *et al.*, 2004; Figueiredo *et al.*, 2012; Simeao *et al.*, 2016). According to Crossa *et al.* (2006), the main feature of mixed linear model methodology is to allow modeling not only independent observations, but also heterogeneous and correlated variance and covariance (VCOV) structures. Moreover, genotype x harvest interaction has been studied in the context where the variances within harvests are assumed to be equal and all pairwise covariances between harvests are zero because harvests are assumed to be independent (Lédo *et al.*, 2008; Figueiredo *et al.*, 2012). This is an unlikely assumption since harvests have a dependence over time and their correlations can differ due to environmental and genetic conditions. Mixed models with VCOV structures such as power and factor analytic can accomodate this in a more realistic way.

Mixed models approach also deals well with unbalanced data and it has been widely used in breeding programs in other crops (Smith *et al.*, 2005; Piepho *et al.*, 2008; Pastina

et al., 2012; MARGARIDO *et al.*, 2015). Another advantage is the possibility to include information from relatives by exploiting genetic correlation. The most common approach consists of the use of a numerator matrix ($\boldsymbol{A}$-matrix) containing an estimate of the pairwise relationship among individuals. This matrix has been widely used for diploid species and was extended for autopolyploids recently to include the ploidy level and double reduction information (KERR *et al.*, 2012).

In forage breeding research, it is important to identify genotype combinations with favorable phenotypic performance for several traits, specially leaf dry matter and nutritive values. According to MARTUSCELLO *et al.* (2009), leaf dry matter is one of the greatest interest for forage breeders, since leaf accumulation is favorable for animal production because of the higher quality of the leaves that are more digestible than the stems. Furthermore, if there is a relationship between groups of yield and quality traits, multivariate analyses such as canonical correlation analysis can provide information for indirect selection (BALKAYA *et al.*, 2011).

The objective of this study was to evaluate nutritional, structural, and yield traits in an outcrossing *P. maximum* multi-parent population, comparing three different classes of longitudinal multivariate linear mixed models. Also, we estimated the heritability and genotypic correlation coefficients between traits, and estimated the interrelationships between sets of yield and quality traits.

3.3 Material and Methods

3.3.1 *Panicum maximum* population

Embrapa Beef Cattle initiated, in 2012, an IRS program, with the objective of improving the average of a tetraploid sexual population, and make this population the basis for new crosses with superior apomictic genotypes. Hereby, 20 tetraploid plants (JA, S7, S13, S16, A42, B87, T103, T4610, A47, A72, B107, C48, C16, B22, Y34, C54, B74, B96, BX4, and B103) were selected based on relevant agronomic performance in forage breeding program of EMBRAPA during the last 30 years. An outcrossing *Panicum maximum* multi-parent population were developed by crossing these 20 parents to get outcrossing half-sib progenies (Figure 3.1). The population was composed of 19 half-sibs progenies (except offsprings of parental B107) and each progeny was formed by around 30 individuals, totalizing 570 tetraploid sexual genotypes in the base population of first recurrent selection cycle.

3.3.2 Experimental data

The phenotypic data was obtained in 2013, 2014 and 2015 (eigth harvests). The design used was an augmented block design (ABD), with 570 sexual regular treatments and three apomitic checks (B107, Mombaça, and Tanzânia), distributed in six blocks. Regular treatments appeared only once in the experiment (95 regular treatments per block), and the checks were repeated in all blocks (5 replications of each check in each block) (Appendix Figure A.6). Experimental plots were formed by transplanted seedlings, with spacing between rows and plants of 2 and 1 meters, respectively. Borders were established around the blocks and were not being considered in the evaluations.

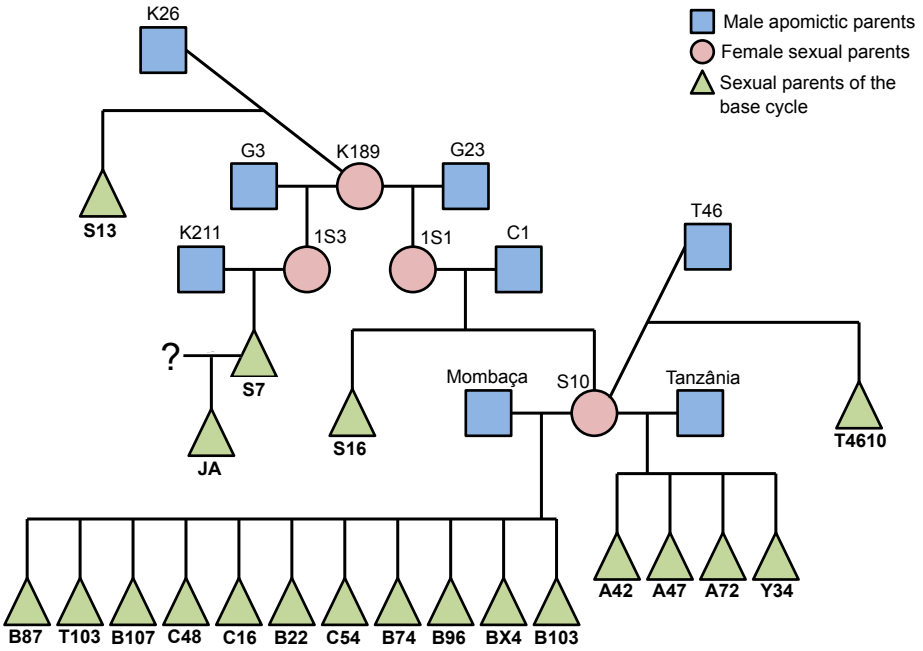

Figure 3.1. Pedigree information of parents (in bold) used to form the base population of intrapopulational recurrent selection program of *P. maximum*.

The first harvest (60 days) was performed to standardize the plants to approximately 20 cm of the ground. The other harvests were carried out subsequently. In the first year, there were three harvests during the rainy season (February, March, and November, 2013) and one harvest during the dry season (October, 2013). In the second year, there was one harvest during the dry season (October, 2014). In the third year, there were three harvests during the rainy season (January, February, and March, 2015).

The experiment was conducted in the experimental field and in the Forage Sample Processing Laboratory (LPAF) of Embrapa Beef Cattle. This center is located in Campo Grande, Mato Grosso do Sul, Brazil (latitude 20°27' S, longitude 54°37' W and altitude of 530 m). The climate is classified as rainy tropical type, characterized by the well defined occurrence of a dry period during the colder months of the year and a rainy period during the summer months.

Response variables were: i) yield traits: total green matter (TGM - t/plant), total dry matter (TDM, g/plant), leaf dry matter (LDM, g/plant), and stem dry matter (SDM, g/plant); ii) structural trait: percentage of leaf blade (PLB, %); and iii) nutritive values of leaf: acid detergent fiber (ADF), crude protein (CP), and in vitro digestibility of organic matter (IVD). These quality traits for nutritive values were performed using near infrared reflectance spectroscopy, according to Marten *et al.* (1985). Yield and structutal traits were evaluated for eight harvests (six during the rainy season and two during the dry season), and nutritive values for four harvests (three during the rainy season and one during the dry season).

3.3.3 Longitudinal multivariate linear mixed analysis

Three different classes of models were used to compare the efficiency in including the relationship matrix. The first class is common in breeding programs and consider, in an explicit

way, the genotype x harvest interaction as an effect in the model, and residual effects were assumed homocedastic and not correlated (Class A). Classes B and C had the genetic and residual effects fitted with different VCOV structures. Class B assumes genotypes as not correlated and the Class C assumes genotypes are correlated by an additive relationship matrix specifically for autotetraploids, based on pedigree values (Amadeu *et al.*, 2016). All models were fitted in the software GenStat, version 17.1 (VSN International, 2015).

Class A (underlining indicates a random variable) was:

$$\underline{y}_{ijlk} = \mu + H_i + \underline{B}_j + P_l + \underline{O}_k + \underline{OH}_{ki} + \underline{\varepsilon}_{ijlk}$$

where $\underline{y}_{ijlk}$ is the phenotype of the k-th genotype, with the l-th parent, at the j-th block and i-th harvest; μ is the overall mean; H_i is the effect of the i-th harvest ($i = 1, \ldots, n_h$, where $n_h = 8$ for yield and structural traits, and $n_h = 4$ for nutritive values); $\underline{B}_j$ is the effect of the j-th block ($j = 1, \ldots, 6$); P_l is the effect of the l-th parent ($l = 1, \ldots, n_s + n_a$, the parents can be separated into two groups, where n_s is the number of sexual parents ($n_s = 1, \ldots, 19$), and n_a is the number of apomictic checks ($n_a = 1, 2, 3$), *i.e.*, $n_s + n_a = 22$); $\underline{O}_k$ is the effect of the k-th genotype ($k = 1, \ldots, 570$); $\underline{OH}_{ki}$ is the genotype by harvest interaction; and $\underline{\varepsilon}_{ijlk}$ is the residual error. Block, genotype, genotype by harvest interaction, and residual effects were assumed to follow a multivariate normal distribution, with zero mean and VCOV matrix $\boldsymbol{I}\sigma^2_B$, $\boldsymbol{I}\sigma^2_O$, $\boldsymbol{I}\sigma^2_{OH}$ and $\boldsymbol{I}\sigma^2_R$, respectively; $\boldsymbol{I}$ is an identity matrix.

Class B and Class C were fitted by a longitudinal multivariate linear mixed (LMLM) model:

$$\underline{y}_{ijlk} = \mu + H_i + \underline{B}_j + P_l + \underline{O}_{ki} + \underline{\varepsilon}_{ijlk}$$

where $\underline{y}_{ijlk}$, H_i, $\underline{B}_j$, P_l, and $\underline{\varepsilon}_{ijlk}$ were described above, and $\underline{O}_{ki}$ is the effect of the k-th genotype ($k = 1, \ldots, 570$), within the i-th harvest.

Block, genotype, and residual effects were assumed to follow a multivariate normal distribution, with zero mean and VCOV matrix $\boldsymbol{I}\sigma^2_B$, $\boldsymbol{G}$, and $\boldsymbol{R}$, respectively. The $\boldsymbol{G}$ matrix is indexed by two factors (harvest and genotype) written as the Kronecker product of matrices: $\boldsymbol{G} = \boldsymbol{G}_H^{i\times i} \otimes \boldsymbol{G}_O^{k\times k}$, in which $\boldsymbol{G}_H$ is relative to harvest effect and $\boldsymbol{G}_O$ is relative to genotype effect. The $\boldsymbol{R}$ matrix is indexed by three factors (plot, harvest, and block): $\boldsymbol{R} = \boldsymbol{I}_{pl}^{n\times n} \otimes \boldsymbol{R}_H^{i\times i} \otimes \boldsymbol{R}_B^{j\times j}$, in which $\boldsymbol{R}_H$ and $\boldsymbol{R}_B$ are relative to harvest and block effects, respectively, and $\boldsymbol{I}_{pl}$ is an identity matrix for plot effects ($n = 95$, which is the number of regular treatments per block).

Matrices $\boldsymbol{G}_H$ and $\boldsymbol{R}_H$ were tested considering six different structures (Table 3.1). $\boldsymbol{R}_B$ was tested for all VCOV matrices, except Po. Matrix $\boldsymbol{G}_O$ was defined as ID matrix for Class B and as additive relationship matrix for Class C. The additive relationship matrix was obtained with the R package AGH-matrix (Amadeu *et al.*, 2016).

Class A is equivalent to Class B, but using a CS structure as $\boldsymbol{G}_H$ matrix and ID for the others effects. For Class B and C, we selected the best model based on goodness of fit of the model, considering the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). The selection was performed into three steps: first we fitted the $\boldsymbol{G}_H$ matrix for different structures; second we fitted the $\boldsymbol{R}_H$ matrix given the selected $\boldsymbol{G}_H$ previously; and third we fitted the $\boldsymbol{R}_B$ matrix given the selected $\boldsymbol{G}_H$ and $\boldsymbol{R}_H$.

Table 3.1. Description and number of parameters (n_{PAR}) of variance and covariance structures examined for genetic ($\boldsymbol{G}$) and residual ($\boldsymbol{R}$) effects.

Model	n_{PAR}	Description
ID	1	Identical residual variation
DIAG	p	Heterogeneous residual variation
CS	2	Compound symmetry with homogeneous residual variation
CS_{Het}	$p + 1$	Compound symmetry with heterogeneous residual variation
Po	2	Power model with homogeneous residual variation
FA1	$2p$	First-order factor analytic model

p is the number of harvests for $\boldsymbol{G}_H$ and $\boldsymbol{R}_H$, and is the number of blocks for $\boldsymbol{R}_B$.

Heritability was calculated using the generalized measure of heritability. The model with CS matrix for genetic effects was considered for all traits (Cullis *et al.*, 2006):

$$\hat{H}^2_C = 1 - \frac{PEV}{2\sigma^2_G}$$

where PEV is the prediction error variance, *i.e.*, the mean variance of a difference of two BLUP (best linear unbiased prediction); σ^2_G is the genetic variance. This was calculated using function VHERITABILITY implemented in GenStat, version 17.1 (VSN International, 2015).

In order to evaluate the correspondence among the 20 best offsprings selected by the different models (3.5% of selection intensity), we estimated the coincidence index (Hamblin and Zimmermann, 1986):

$$CI = \frac{A - C}{M - C} \times 100$$

where M is the total number of offspring selected in each model (*i.e.*, 20), A is the number of offsprings selected in two models, and C is the number of offsprings selected in both models due to chance, using proportion of 5%, implies that $C = 1$.

The response to selection was calculated, for selected model, as follows:

$$R = H^2_C \times S$$

where S is the selection differential.

Genetic and additive correlation were estimated using Pearson correlation among traits for selected models of Class B and C, respectively. These correlations are shown using graphs obtained by R package qgraph (Epskamp *et al.*, 2012). The relationship among sets of forage production and quality traits were investigated using canonical correlation analysis (CCA), which was performed in R package CCA (González *et al.*, 2008). Quality variable set was defined as X-set (ADF, CP, and IVD) with canonical variable U and forage variable set was defined as Y-set (LDM, SDM, and PLB) with canonical variable V. F test was performed for pairs of canonical variables as described by Balkaya *et al.* (2011).

Genotype x harvest interaction was studied using the first-order factor analytic (FA1) as VCOV matrix for $\boldsymbol{G}_H$ and the additive relationship matrix for $\boldsymbol{G}_O$. The graph of correlations among harvests was obtained using the R package corrplot (Wei and Simko, 2016). The graph of performance of five selected offsprings in different harvests was obtained using the R package ggplot2 (Wickham, 2009) (Appendix Figure A.7).

Table 3.2. Comparisons between Class A and the best models of Class B and C, using AIC and BIC, for acid detergent fiber (ADF), crude protein (CP), in vitro digestibility of organic matter (IVD), total green matter (TGM), total dry matter (TDM), leaf dry matter (LDM), stem dry matter (SDM), and percentage of leaf blade (PLB). Classes with smaller AIC and BIC are indicated in bold.

Traits	Class	Genetic effects	Residual effects	AIC	BIC
ADF	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	10358.91	**10382.19**
	B	CS_{Het} $\otimes$ ID	ID $\otimes$ Po $\otimes$ ID	10347.79	10394.34
	C	Po $\otimes$ A	ID $\otimes$ DIAG $\otimes$ ID	**10344.34**	10385.08
CP	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	8395.77	8419.05
	B	FA1 $\otimes$ ID	ID $\otimes$ CS_{Het} $\otimes$ CS	8267.62	8354.92
	C	FA1 $\otimes$ A	ID $\otimes$ FA1 $\otimes$ ID	**8234.65**	**8333.59**
IVD	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	14430.05	14453.32
	B	FA1 $\otimes$ ID	ID $\otimes$ Po $\otimes$ ID	14357.75	14421.77
	C	FA1 $\otimes$ A	ID $\otimes$ ID $\otimes$ ID	**14349.85**	**14408.04**
TGM	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	9210.97	9237.08
	B	Po $\otimes$ ID	ID $\otimes$ DIAG $\otimes$ DIAG	7489.26	7593.72
	C	Po $\otimes$ A	ID $\otimes$ DIAG $\otimes$ DIAG	**7409.11**	**7513.58**
TDM	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	65405.63	65431.72
	B	Po $\otimes$ ID	ID $\otimes$ DIAG $\otimes$ DIAG	62711.48	62815.86
	C	Po $\otimes$ A	ID $\otimes$ DIAG $\otimes$ ID	**62700.68**	**62772.44**
LDM	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	60746.66	60772.76
	B	Po $\otimes$ ID	ID $\otimes$ DIAG $\otimes$ FA1	59178.80	59322.32
	C	Po $\otimes$ A	ID $\otimes$ DIAG $\otimes$ DIAG	**59155.55**	**59259.93**
SDM	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	54513.90	54540.00
	B	DIAG $\otimes$ ID	ID $\otimes$ CS $\otimes$ CS_{Het}	53175.63	53286.54
	C	DIAG $\otimes$ A	ID $\otimes$ CS_{Het} $\otimes$ DIAG	**49462.57**	**49612.61**
PLB	A	CS $\otimes$ ID	ID $\otimes$ ID $\otimes$ ID	38622.21	38648.31
	B	DIAG $\otimes$ ID	ID $\otimes$ FA1 $\otimes$ ID	37800.82	37963.92
	C	CS_{Het} $\otimes$ A	ID $\otimes$ ID $\otimes$ ID	**37752.99**	**37824.75**

3.4 Results

3.4.1 Model selection

The VCOV structures were selected based on AIC and BIC criteria for models of Class B and C (an example can be seen at the Appendix Table A.5 for ADF variable). A comparison between Class A and the best models of Class B and C was performed (Table 3.2) considering AIC and BIC as well. Class C had lower AIC criteria for all traits; for BIC, Class A had goodness of fit for ADF and Class C for the remaining traits. As the difference of AIC values between the first and the second lower values were greater than the respective BIC difference, Class C was also selected for ADF.

3.4.2 Genetic parameters

Generalized measure of heritability ranged from 30.65% to 87.54% (Table 3.3). LDM showed the higher heritability and IVD showed the lower heritability. The response to selection ranged from 0.90% to 70.07% for CP and LDM, respectively. On average, yield traits showed higher heritability and greater response to selection than nutritive values.

Table 3.3. Generalized heritability ($\hat{H}^2_C$%), response to selection of the greater 20 offsprings (R%) for Model 3, and coincidence index (CI%, with absolute number of offsprings selected in parenthesis) among three classes of models, for acid detergent fiber (ADF), crude protein (CP), in vitro digestibility of organic matter (IVD), total green matter (TGM), total dry matter (TDM), leaf dry matter (LDM), stem dry matter (SDM), and percentage of leaf blade (PLB).

Trait	$\hat{H}^2_C$	R	CI		
			Class A – Class B	Class A – Class C	Class B – Class C
ADF	36.40	0.96	73.68 (15)	94.74 (19)	68.42 (14)
CP	36.48	0.90	10.53 (3)	47.37 (10)	10.53 (3)
IVD	30.65	1.40	68.42 (14)	57.89 (12)	73.68 (15)
TGM	83.88	56.95	78.95 (16)	78.95 (16)	100.00 (20)
TDM	86.49	58.18	78.95 (16)	78.95 (16)	100.00 (20)
LDM	87.54	70.07	89.47 (18)	89.47 (18)	94.74 (19)
SDM	45.60	7.56	52.63 (11)	15.79 (4)	42.11 (9)
PLB	34.93	3.11	63.16 (13)	52.63 (11)	73.68 (15)

The lower coincidence index was 10.53% for CP, with 36.48% of generalized heritability, between Classes A and B, and the higher index was 100% for TGM and TDM between Classes B and C and with generalized heritability of 83.88% and 86.49%, respectively. Traits with higher heritability also had higher coincidence index among classes. Classes B and C were, in average, more concordant between them.

Genetic and additive correlations can be seen in Figure 3.2A and Figure 3.2B, respectively. The correlations are graded by the size, thickness and color of the traces. Short and thick traces represent high correlations between traits and red and green colors represent negative and positive correlations, respectively.

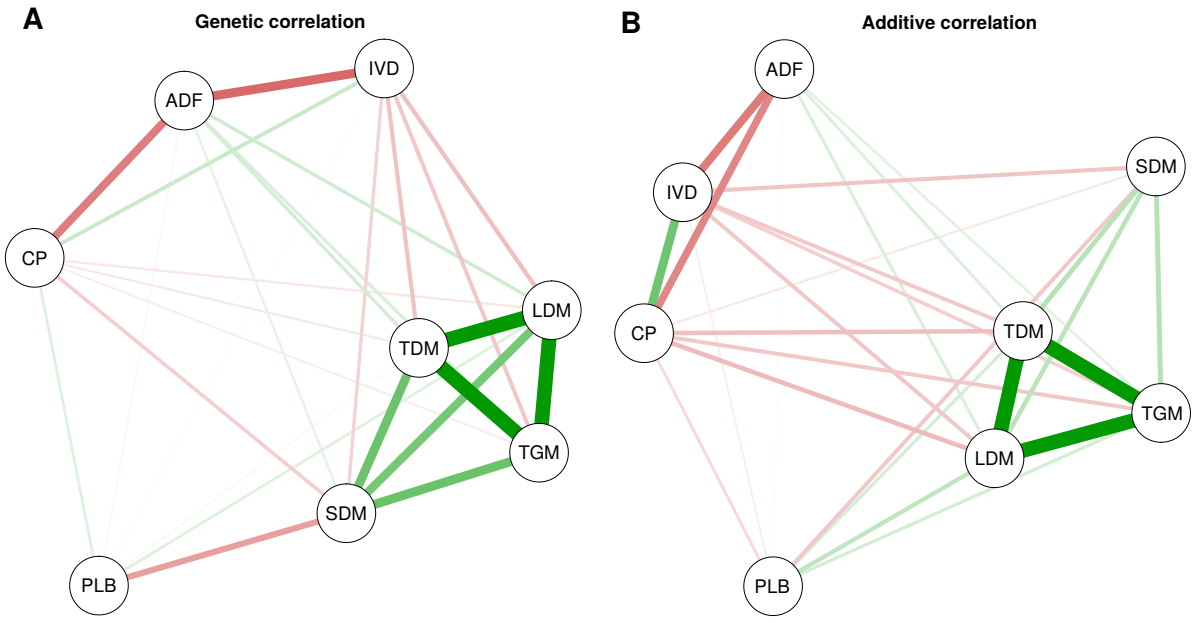

Figure 3.2. (A) Genetic correlation estimated using Class B models. (B) Additive correlation estimated using Class C models. As the shorter and thicker the traces are, the greater are the correlations, being red for negative correlations and green for positive correlations.

Table 3.4. Summary results for the canonical correlation analysis.

Pair of canonical variables	Canonical correlation	Squared canonical correlation	F	df1	df2	p-value
U_1V_1	0.3218	0.1035	10.3065	9	1373	<0.001
U_2V_2	0.2075	0.0431	7.1265	4	1130	<0.001
U_3V_3	0.0761	0.0058	3.3010	1	566	0.0698

Table 3.5. Canonical coefficients.

	X - Variable set				Y - Variable set		
	ADF	CP	IVD		LDM	SDM	PLB
U_1	0.3731	0.5076	0.7963	V_1	-0.5174	-0.6446	-0.3876
U_2	-0.6716	0.8135	-0.9189	V_2	-0.8179	0.8730	0.0784
U_3	-0.9284	-0.7964	0.4116	V_3	-0.5083	0.0291	1.0043

The main difference between genetic and additive correlations is between CP and IVD, in which the genetic correlation was 0.21 and additive correlation was 0.56. The ADF trait had negative correlation between the other nutritive values, no correlation with PLB, and positive and low correlation between yield traits. PLB has negative correlation with SDM, and positive and low correlation with the other yield traits, which have positive and medium-high correlations among them. TGM, TDM, and LDM showed the highest correlations ($r = 0.98$ between TGM and LDM, and $r = 0.99$ between TGM and TDM, and LDM and TDM).

Three canonical correlation coefficients were estimated to explain the interrelationships between the variable sets. First and second canonical coefficients were significant (32.18% and 20.75%, p-value <0.001) while the third one was not significant (7.61%, p-value 0.0698) (Table 3.4). The first optimal linear combination of dependent and independent variables (Table 3.5) is:

$$U_1 = 0.3731(\text{ADF}) + 0.5076(\text{CP}) + 0.7963(\text{IVD})$$

$$V_1 = -0.51741(\text{LDM}) - 0.6446(\text{SDM}) - 0.3876(\text{PLB})$$

3.4.3 Genotype x harvest interaction

Genotype x harvest interaction is visualized from different correlations among pairs of harvests (Figure 3.3) as well as different performances of individuals in successive harvests (Appendix Figure A.7). Results indicate that the performance of the offsprings was not uniform for the different seasons (rainy and dry seasons) as well as among harvests within the same season.

3.5 Discussion

The goal of this study was to evaluate nutritional, structural, and yield traits in an outcrossing *P. maximum* multi-parent population besides comparing three classes of linear mixed models. Furthermore, we estimated the heritability and genotypic correlation coefficients between traits as well as canonical correlation between sets of forage production and quality traits. Given the paucity of information about VCOV structures in *P. maximum* experiments, we fitted

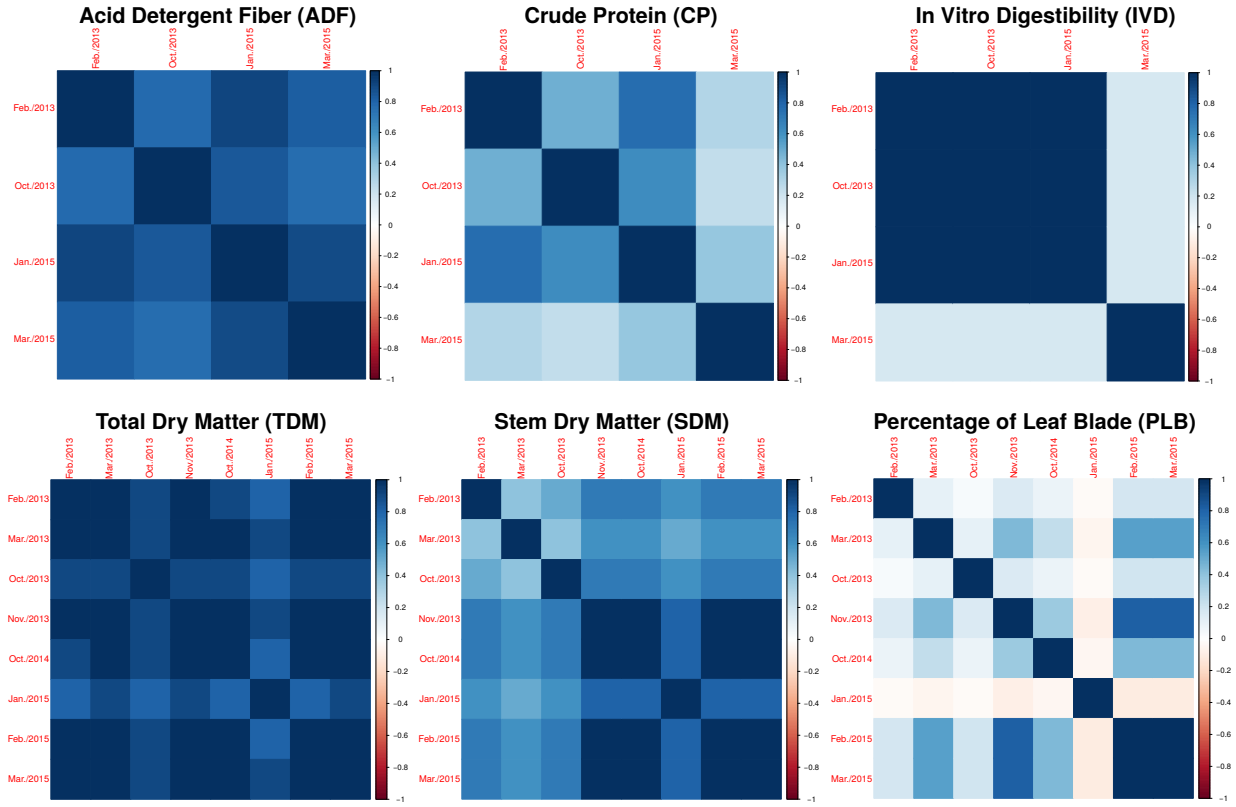

Figure 3.3. Correlation among harvests for acid detergent fiber (ADF), crude protein (CP), in vitro digestibility of organic matter (IVD), total dry matter (TDM), stem dry matter (LDM), and percentage of leaf blade (PLB) using first-order factor analytic as VCOV matrix for genetic effects.

several linear mixed models, selected by AIC and BIC criteria, and compared the best models of each class. First class (Class A) is classified as traditional model, where genotype x harvest interaction was treated as an effect, genotypes were not correlated, and residual effects were assumed homocedastic and not correlated. In the second class (Class B), genetic and residual effects were fitted with different VCOV structures, and genotypes were not correlated. In the third class (Class C), genetic and residual effects were fitted with different VCOV structures, and genotypes were correlated by an additive relationship matrix.

Based on AIC and BIC criteria, Class C was selected for all analyzed traits (Table 3.2). Therefore, the selection of VCOV structures for genetic and residual effects combined with an additive relationship matrix for genotype effect has showed better performance in relation to Classes A and B. In addition, Class C allows to model genotype x harvest interaction indirectly by $\boldsymbol{G}_H$ matrix. Although most studies have incorporated genotype x harvest interaction in mixed models analysis, they usually assume harvests to be independent (Lédo *et al.*, 2008; Figueiredo *et al.*, 2012). However, this is a heavy assumption since the harvests generally have a longitudinal correlation due to environmental and genetic conditions (Figure 3.3). In this work, we assumed different genetic correlation structures among harvests, which were selected based on goodness of fit of the model. For most traits, Po and FA1 structures fitted better than others and should be tested in subsequent experiments. Another advantage of Class C is allowing the

inclusion of correlation among offsprings using an additive relationship matrix for autotetraploids based on pedigree values. According to Crossa *et al.* (2006), the information among relatives should facilitate estimating the association between environments (in this case harvests), as well as modeling the main effects of genotype and genotype x environmental interaction (here, genotype x harvest interaction). Therefore, we recommend to incorporate additive relationship matrix besides modeling genotype x harvest interaction by Po or FA1 structures in longitudinal mixed models in *P. maximum*. Furthermore, an additive relationship matrix estimated by molecular markers can be used instead of pedigree information to increase the accuracy in predicting genetic values (VanRaden, 2008).

Additionally, we estimated the generalized measure of heritability, the response to selection, and the coincidence index among the models. Magnitude of heritability determines the degree of difficulty in improving the trait and indicates the most efficient method to be used. According to Resende *et al.* (2004), it is of utmost importance in the breeding program. In our case, heritability for yield traits was higher than those cited in the literature (Table 3.3). For example, Jank *et al.* (2008) evaluated accessions and hybrids of *P. maximum* and reported means of heritabilities with magnitude between 40% and 68% and from 31% to 76% for TDM and LDM, respectively. In another study, Braz *et al.* (2013) observed 20% and 30% for the same traits. These results are lower than the observed in this article, 86.49% and 87.54% for TDM and LDM, respectively. These differences can be explained based on the type of index used to calculate the heritability. The first two works used the broad sense heritability ($H^2 = \sigma^2_G/\sigma^2_F$, where σ^2_F is the phenotypic variance) and this work used the generalized heritability (H^2_C), which is more indicated for unbalanced designs (Cullis *et al.*, 2006). It is worth mentioning that, in case of balanced designs, the usual broad-sense heritability and the generalized heritability coincide (Piepho and Möhring, 2007). In addition, the broad sense heritability was calculated in this work for comparison. TDM obtained 41.36% and LDM obtained 49.92%, that is closer to reports in the literature.

On average, yield traits showed higher heritability and greater response to selection than nutritive values, since they were evaluated more extensively. In general, traits with lower heritability also had higher divergence among models and, consequentely, lower coincidence of the 20 best offsprings selected (Table 3.3). For CP, 10.53% of the selected offspring were coincident between Classes A and B, and between Classes B and C. For SDM, 15.79% were coincident between Classes A and C. Only TGM and TDM did not differ for selected offsprings between Classes B and C, however these traits have high heritability and are very correlated (Figure 3.2). As expected, Class B and C were more concordant between them, since both allow different VCOV structures for genetic and residual effects. Given the best fit of the model besides the difference in ranking between the greater 20 offsprings for the classes, Class C should be preferred for selection of superior individuals in an IRS program.

The PLB is essential in forage breeding and being a low-heritability trait, it can be genetically improved through strategies such as correlated response. In fact, PLB is slightly correlated with LDM (r = 0.24, Figure 3.2), which has high-heritability and provided greater response to selection in this and in other studies (Martuscello *et al.*, 2007, 2009; Braz *et al.*, 2015). In addition, individuals with good performance in several traits are also desirable such

as plants with high leaf yield, and high percentages of protein and digestibility. One possibility to perform indirect selection of several traits simultaneously is through canonical correlation analysis, which can be a good approach to define selection indices (Cerón-Rojas *et al.*, 2016). Magnitudes of the canonical coefficients (Table 3.5) represent their relative contributions to the correlated variable (Balkaya *et al.*, 2011). Accordingly, if the nutritive values (ADF, CP, and IVD) increase, forage production (LDM, SDM, and PLB) will decrease. However, positive and low canonical correlations were observed for the first (32.18%) and second (20.75%) pairs of canonical variables (Table 3.4). Although it is necessary to give more attention to select, simultaneously, yield and quality traits, it is possible to define a selection index that allows selecting one set of traits without drastically decreasing the other set.

Acknowledgments

This work was supported by FAPESP (São Paulo Research Foundation), grants 2015/20659-2 and 2016/01279-7. Additional support was provided by CNPq ("Conselho Nacional do Desenvolvimento Científico e Tecnológico"), EMBRAPA ("Empresa Brasileira de Pesquisa Agropecuária"), and by UNIPASTO ("Associação para o Fomento à Pesquisa de Melhoramento de Forrageiras"). The author thanks all trainees of Embrapa Beef Cattle that helped the phenotypic evaluations.

References

Akaike, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**: 716–723.

Amadeu, R. R., C. Cellon, J. W. Olmstead, A. A. F. Garcia, M. F. R. Resende, and P. R. Muñoz, 2016 AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. The Plant Genome **9**: 1–10.

Balkaya, A., S. Cankaya, and M. Ozbakir, 2011 Use of canonical correlation analysis for determination of relationships between plant characters and yield components in winter squash (*Cucurbita maxima* Duch.) populations. Bulgarian Journal of Agricultural Science **17**: 606–614.

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.

Braz, T. G. d. S., D. M. D. Fonseca, L. Jank, M. D. V. D. Resende, J. A. Martuscello, and R. M. Simeão, 2013 Genetic parameters of agronomic characters in *Panicum maximum* hybrids. Revista Brasileira de Zootecnia **42**: 231–237.

Cerón-Rojas, J. J., J. Crossa, and J. Sahagún-Castellanos, 2016 Statistical sampling properties of the coefficients of three phenotypic selection indices. Crop Science **56**: 51–58.

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.

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 Enviromental Statistics **11**: 381–393.

Epskamp, S., A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and B. Denny, 2012 qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software **48**: 1–18.

Fernandes, F. D., G. J. Braga, A. K. B. Ramos, L. Jank, M. A. C. Carvalho, G. A. Maciel, C. T. Karia, and C. E. L. Fonseca, 2017 Repeatability, number of harvests, and phenotypic stability of dry matter yield and quality traits of *Panicum maximum* Jacq. Acta Scientiarum **39**: 149–155.

Figueiredo, U. J., J. A. R. Nunes, and C. B. Valle, 2012 Estimation of genetic parameters and selection of *Brachiaria humidicola* progenies using a selection index. Crop Breeding and Applied Biotechnology **12**: 237–244.

González, I., S. Déjean, P. G. P. Martin, and A. Baccini., 2008 CCA: An R package to extend canonical correlation analysis. Journal of Statistical Software **23**: 1–14.

Hamblin, J. E. and M. J. O. Zimmermann, 1986 Breeding common bean for yield in mixtures. Plant Breeding Reviews **4**: 245–272.

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., R. M. S. Resende, C. B. Valle, M. D. V. Resende, L. Chiari, L. J. Cançado, and C. Simioni, 2008 Melhoramento genético de *Panicum maximum*. In *Melhoramento de forrageiras tropicais*, edited by R. M. S. Resende, C. B. Valle, and L. Jank, pp. 55–87.

Jank, L., C. B. Valle, and R. M. S. Resende, 2011 Breeding tropical forages. Crop Breeding and Applied Biotechnology **11**: 27–34.

Kerr, R. J., L. Li, B. Tier, G. W. Dutkowski, and T. A. McRae, 2012 Use of the numerator relationship matrix in genetic analysis of autopolyploid species. Theoretical and Applied Genetics **124**: 1271–1282.

Lédo, F. J. d. S., A. V. Pereira, F. d. S. Sobrinho, A. M. Auad, L. Jank, and J. S. Oliveira, 2008 Estimativas de repetibilidade para caracteres forrageiros de *Panicum maximum*. Ciência e Agrotecnologia **32**: 1299–1303.

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**: 175–189.

Marten, G. C., J. S. Shenk, and F. E. Barton, 1985 Near infrared reflectance spectroscopy (NIRS). p. 96.

Martuscello, J. A., L. Jank, D. M. Da Fonseca, C. D. Cruz, and D. D. N. F. V. Da Cunha, 2009 Among and with family selection and combined half-sib family selection in *Panicum maximum* Jacq. Revista Brasileira de Zootecnia **38**: 1870–1877.

Martuscello, J. A., L. Jank, D. M. Da Fonseca, C. D. a. Cruz, and D. D. N. F. V. Da Cunha, 2007 Repetibilidade de caracteres agronômicos em *Panicum maximum* Jacq. Revista Brasileira de Zootecnia **36**: 1975–1981.

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.

Piepho, H.-P. and J. Möhring, 2007 Computing heritability and selection response from unbalanced plant breeding trials. Genetics **177**: 1881–1888.

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

Resende, M. D. V., R. M. S. Resende, L. Jank, and C. B. Valle, 2008 Experimentação e análise estatística no melhoramento de forrageiras. In *Melhoramento de forrageiras tropicais*, edited by R. M. S. Resende, C. B. Valle, and L. Jank, pp. 195–287.

Resende, R. M. S., L. Jank, C. B. do Valle, and A. L. V. Bonato, 2004 Biometrical analysis and selection of tetraploid progenies of *Panicum maximum* using mixed model methods. Pesquisa agropecuária brasileira **39**: 335–341.

Savidan, Y. H., 2000 Apomixis: genetics and breeding. Plant Breeding Reviews **18**: 10–86.

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.

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

Simeao, R., A. Silva, C. B. D. Valle, M. D. Resende, and S. Medeiros, 2016 Genetic evaluation and selection index in tetraploid *Brachiaria ruziziensis*. Plant Breeding **135**: 246–253.

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. Journal of Agricultural Science **143**: 449–462.

Valle, C. B. D., L. Jank, and R. M. S. a. Resende, 2009 O melhoramento de forrageiras tropicais no Brasil. Revista Ceres **56**: 460–472.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

VSN International, 2015 *GenStat for Windows*. Hemel Hempstead, UK.

Wei, T. and V. Simko, 2016 *corrplot: Visualization of a correlation matrix*. R package version 0.77.

Wickham, H., 2009 *ggplot2: Elegant graphics for data analysis*.

4 DEVELOPMENT OF STATISTICAL MODELS INCLUDING DOSAGE INFORMATION FOR GENOMIC SELECTION IN *PANICUM MAXIMUM*

Keywords: Plant Breeding; Guinea Grass; Autotetraploid Forage; Genotyping-by-sequencing (GBS); Genomic Prediction

4.1 Abstract

Genomic selection is an effective method to predict breeding values based on large set of marker information distributed across the whole genome. Despite advances in breeding programs and genotyping techniques, genetic studies have been limited to the use of markers without allelic dosage information for poliploid species. Since several species of economic interest are autotetraploid, such as the forage *Panicum maximum*, the development of genomic selection models that consider tetraploid allelic dosage is fundamental for breeding programs. Therefore, we developed predictive models that consider tetraploid dosage (GS-TD models) in real data of a recurrent selection population of *P. maximum*. Population consisted of 530 tetraploid sexual plants. Forage production and nutritive value traits were measured by eight and four harvests, respectively, in an augmented block design. A longitudinal multivariate linear mixed model was fitted considering different variance and covariance matrices for residual effects. Genotyping-by-sequencing (GBS) was conducted in NextSeq 500 platform for 96-plex PstI libraries. Raw data was analyzed using Tassel-GBS pipeline and GBS tags were aligned against six pseudo-genomes. Allelic dosage was estimated in SuperMASSA software for ploidy 4. Markers were selected with minimum overall depth of 25 reads and with up to 5% of missing data. Missing data were imputed using random sampling considering the probability of occurrence of each dose. Linkage disequilibrium (LD) was estimated using squared Pearson correlation. Six predictive models were generalized to tetraploid species and predictive ability was estimated by 5-folds cross-validation, repeated 100 times for bayesian models and 1,000 times for frequentist model. A total of 41,424 markers were selected after all filters and a high degree of LD was observed even for extended distances between markers. Mean predictive ability ranged from 0.1691 to 0.4668 among traits. The accuracy of predictive models justifies the implementation of genomic selection in *P. maximum* breeding programs. This is the first work of genomic selection in tropical forages which uses a high throughput genotyping and considers tetraploid allelic dosage in bayesian and frequentist models. Furthermore, the use of tetraploid allelic dosage is a more realistic assumption of the genetic architecture on autotetraploid species. As conclusion, GBS and allelic dosage are promising strategies for genomic analysis in autotetraploid species, in addition, genomic selection may lead to additional gains in recurrent selection program of *P. maximum*.

4.2 Introduction

Many agricultural crops of economic interest are tetraploids, such as potato (*Solanum tuberosum*) (Allard, 1960), alfalfa (*Medicago sativa*) (McCoy and Bingham, 1988), and guinea grass (*Panicum maximum*) (Warmke, 1954). For these species, marker alleles can be represented with differents dosages, ranging from 0 (nulliplex) to 4 (quadruplex). Allele dosage

refers to the number of copies of the reference allele, *e.g.* aaaa for nulliplex and AAAA for quadruplex, in which the A is the reference allele, for a biallelic marker.

Approaches for polyploid genetic analyses historically is based on the idea of using loci in single dose, with 1:1 segregation in biparental crossings, initially proposed by Wu *et al.* (1992). With the development of Next-Generation Sequencing (NGS) technologies and the advance of genetic and statistical methods, new possibilities have arisen to enable studies of the complex polyploid genomes for many crops (Garcia *et al.*, 2013). The evaluation of Single-Nucleotide Polymorphisms (SNP) throughout the genome allows one to assess the relative abundance of each allele, in other words, to estimate the allelic dosage of SNPs (Serang *et al.*, 2012; Garcia *et al.*, 2013; Mollinari and Serang, 2013).

According to Osborn *et al.* (2003), allelic dosage effects are observed in heterozygous genotypes as intermediate gene expression levels and phenotypic effects when compared with low or high expressing alleles in homozygous genotypes. Thus, polyploidy can increase the potential variation in its genic expression, reflecting in phenotypic variation. Therefore, the inclusion of allelic dosage information has become essential for genetic studies in polyploid species. It will allow, in addition to the use of all genotypic information, an investigation about its importance to improve the development of statistical models in autopoliploid species (Garcia *et al.*, 2013), such as in the forage *P. maximum*.

Panicum maximum stands out among the tropical forage species due to its high biomass yield, excellent nutritive quality, and excellent acceptability and digestibility, providing high animal performance (Jank *et al.*, 2011). Forage breeding is a relatively new event, but it has been stimulated due to the commercial interest in tropical pastures. Currently, forage breeding programs are traditionally carried out by conventional breeding methods, such as recurrent selection, in which one cycle requires three to five years for evaluation of productivity, persistence, biomass quality, and others morphological characters. Therewith, approximately fifteen years are necessary for development, testing, and release of new cultivars (Resende *et al.*, 2014).

Initially proposed by Meuwissen *et al.* (2001), genomic selection (GS) is an approach that uses statistical methods to predict breeding values from markers distributed throughout the genome, with sufficient accuracy to represent the phenotype. It is a promising approach to accelerate cycles of recurrent selection and increase the accuracy of selection. GS can be superior when compared to former methods such as the traditional method of Marker-Assisted Selection (MAS), specially for selection of traits with low heritability, which are controlled by many loci of small effects. GS is different from MAS because it analyzes all markers simultaneously, including both minor and major marker effects, in a population with wide genome coverage. Moreover, this method calculates the genomic estimated breeding values (GEBVs) of individuals for ranking and progeny selection (Meuwissen *et al.*, 2001; Bernardo, 2014).

The success of GS depends of the prediction accuracy to select individuals whose phenotypes are not evaluated. To achieve this, it is essential that the training population has a direct relationship with the breeding population. Several other factors can affect the prediction accuracy of GS, like span of linkage disequilibrium (LD), trait heritability, genetic architecture, marker density, size of the training population and performance of the analyzed models (de los Campos *et al.*, 2013; Desta and Ortiz, 2014).

Most of the available GS models were developed for diploids and still is not well established for polyploid species. These models do not include allelic dosage information, and diploid models are commonly used in polyploid species. Therefore, the objective of this study was to develop statistical models in genomic selection which consider allelic dosage information for autotetraploid species, with applications in *P. maximum*. For this, SNP calling was performed to allow the identification of different dose levels. This is the first study that includes tetraploid dosage in frequentist and bayesian models for genomic selection in the *Panicum* genus.

4.3 Material and Methods

4.3.1 *Panicum maximum* population

We generated an outcrossing *P. maximum* multi-parent population using 20 selected plants (JA, S7, S13, S16, A42, B87, T103, T4610, A47, A72, B107, C48, C16, B22, Y34, C54, B74, B96, BX4, and B103) as donors of pollen and 19 sexual plants (all parents except B107) as females. These parents were selected based on relevant agronomic performance in the forage breeding program at EMBRAPA during the last 30 years. After the crossing process, we synthesized 19 half-sibs progenies, each progeny composed by 30 individuals, totalizing 570 tetraploid sexual genotypes.

4.3.2 Phenotypic evaluation and longitudinal multivariate linear mixed analysis

The multi-parent population was evaluated in an augmented block design (ABD) at Embrapa Beef Cattle, located at the Campo Grande city, Mato Grosso do Sul, Brazil (20°27'S, 54°37'W, 530m). We conducted the ABD, with 570 sexual regular treatments and three apomitic checks (B107, Mombaça, and Tanzânia), distributed in six blocks. As the usual procedure in an ABD, all regular treatments appeared only once in a block and the checks repeated in all blocks (the ABD structure can be seen in Appendix Figure A.6).

The evaluated traits were: i) yield trait: leaf dry matter (LDM - g/plant); ii) structural trait: percentage of leaf blade (PLB - %); and iii) nutritive values of leaf: organic matter (OM), crude protein (CP), and in vitro digestibility of organic matter (IVD). Yield and structural traits were evaluated for eight harvests, during the years of 2013 (four harvests), 2014 (one harvest), and 2015 (three harvests), and nutritive values were evaluated for four harvests, being two in 2013, and two in 2015.

We fitted a longitudinal multivariate linear mixed (LMLM) model to obtain adjusted means of the traits free of experimental residual effects, for future genomic selection analysis. Our model, in which the underlining indicates a random effect, is:

$$\underline{y}_{ijlk} = \mu + H_i + \underline{B}_j + \underline{P}_l + O_{ki} + \underline{\varepsilon}_{ijlk}$$

where $\underline{y}_{ijlk}$ is the phenotype of the k-th offspring, with the l-th parent, at the j-th block and i-th harvest; μ is the overall mean; H_i is the effect of the i-th harvest ($i = 1, \ldots, n_h$, where $n_h = 8$ for yield and structural traits, and $n_h = 4$ for nutritive values); $\underline{B}_j$ is the effect of the j-th block ($j = 1, \ldots, 6$); $\underline{P}_l$ is the effect of the l-th parent ($l = 1, \ldots, n_s + n_a$, the parents can be separated into two groups, where n_s is the number of sexual parents ($n_s = 1, \ldots, 19$), and n_a

Table 4.1. Variance and covariance structures examined for residual effects ($\boldsymbol{R}_C$).

Model	n_{PAR}	Description
ID	1	Identical residual variation
DIAG	H	Heterogeneous residual variation
CS	2	Compound symmetry with homogeneous residual variation
CS_{Het}	H + 1	Compound symmetry with heterogeneous residual variation
AR1	2	First-order autoregressive model with homogeneous residual variation
AR1_{Het}	H + 1	First-order autoregressive model with heterogeneous residual variation
Po	2	Power model with homogeneous residual variation
Po_{Het}	H + 1	Power model with heterogeneous residual variation
US	$H(H+1)/2$	Unstructured model

n_{PAR} is the number of parameters for the models and H is the number of harvests.

is the number of apomitic checks ($n_a = 1, \ldots, 3$), *i.e.*, $n_s + n_a = 22$); O_{ki} is the effect of the k-th offspring within the i-th harvest ($k = 1, \ldots, n_a + n_o$, where three apomictic checks are repeated five times in each block, and n_o is the number of offsprings ($n_o = 1, \ldots, 570$)); $\underline{\varepsilon}_{ijlk}$ is the residual error.

Block and parent effects follow a multivariate normal distribution with mean zero and $\boldsymbol{I}\sigma_B^2$ and $\boldsymbol{I}\sigma_P^2$ variance, respectively, where $\boldsymbol{I}$ is an identity matrix. The residual term also was assumed to follow a multivariate normal distribution with mean zero and a variance and covariance (VCOV) matrix indexed by three factors (plot, block, and harvest) written as the Kronecker product of matrices, $\boldsymbol{R} = \boldsymbol{I}_{pl}^{n\times n} \otimes \boldsymbol{I}_B^{j\times j} \otimes \boldsymbol{R}_H^{i\times i}$, in which $\boldsymbol{I}_{pl}$, $\boldsymbol{I}_B$, and $\boldsymbol{R}_H$ are relative to plot, block, and harvest effects, respectively. The $\boldsymbol{I}_{pl}$ and $\boldsymbol{I}_B$ is an identity matrix ($n = 95$, which is the number of regular treatments per block). The $\boldsymbol{R}_H$ was analyzed considering nine different structures of VCOV matrix (Table 4.1): independent (ID), diagonal (DIAG), compound symmetry (CS), compound symmetry heterogeneous (CS_{Het}), first-order autoregressive (AR1), first-order autoregressive heterogeneous(AR1_{Het}), power (Po), power heterogeneous (Po_{Het}), and unstructered (US). The model selection was performed based on the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). Adjusted means were obtained for each trait and used in the predictive genomic models.

These analyses were performed in the R package ASReml (Butler *et al.*, 2009).The heritability for each trait was calculated using the index proposed by Cullis *et al.* (2006).

$$\hat{H}_C^2 = 1 - \frac{PEV}{2\sigma_G^2}$$

where, PEV is the prediction error variance, *i.e.*, the mean variance of a difference of two BLUP (best linear unbiased prediction), and σ_G^2 is the genetic variance.

4.3.3 Molecular data

Due to losses of individuals in field, we sequenced a total of 530 offsprings and used it in genomic selection model. DNA of these 530 offsprings were extracted using the DNeasy Plant kit (QIAGEN) and sequenced along with the multi-parents repeated twice. To provide a higher sequence depth, genotyping-by-sequencing (GBS) was conducted in NextSeq 500 platform

for 96-plex Pst1 libraries and following the protocol from Genomic Diversity Facility, Cornell University. Genomic libraries were prepared following Elshire *et al.* (2011).

Raw data was analyzed using Tassel-GBS pipeline (Glaubitz *et al.*, 2014) modified to obtain the original count of the number of reads for each SNP allele. As this pipeline requires a reference genome and *P. maximum* does not have one, we proposed the alignment of GBS tags using six pseudo-genome: (i) *Panicum hallii* genome (v. 2.0; ∼554 Mb arranged in 9 chromosomes and 8,405 scaffolds; diploid forage); (ii) *Panicum virgatum* genome (v 1.0; ∼1,230 Mb arranged in total of 18 chromosomes, 9 chromosomes named as A and B, and 220,646 contigs; tetraploid forage); (iii) *Setaria italica* genome (v 2.2; ∼405.7 Mb arranged in 336 scaffolds; diploid forage); (iv) *Setaria viridis* genome (v 1.0; ∼394.9 Mb arranged in 9 chromosomes and 724 scaffolds; diploid forage); (v) transcriptome of *P. maximum* obtained by EMBRAPA research group (43,803 sequences with width mean of 841.334); and (vi) transcriptome of *P. maximum* obtained by UNICAMP research group (138,853 sequences with width mean of 695.658). The genomes are available in Phytozome website (http://www.phytozome.net/) (Goodstein *et al.*, 2011). The transcriptomes were provided directly by research groups. The transcriptome obtained by UNICAMP group was published by Toledo-Silva *et al.* (2013). The transcriptome obtained by EMBRAPA group has not yet been published. The Bowtie2 algorithm (Langmead and Salzberg, 2012) was used to align tags against each reference with -D and -R parameters defined as 20 and 4, respectively, and with very-sensitive-local argument.

In Tassel-GBS pipeline, the minimum minor allele frequency (mnMAF) considered was 1%. Count information from it was used in SuperMASSA software (Serang *et al.*, 2012) to estimate the correct tetraploid allelic dosage of the individuals. In SuperMASSA software, the minimum overall depth considered was 25 reads and the model used was Generalized Population Model. Markers were fitted and filtered to ploidy 4. Triallelic SNPs were eliminated in this step. Markers were selected manually with up to 5% of missing data. The imputation was made using random sampling considering the probability of occurence of each dose within each marker.

As linkage disequilibrium (LD) can affect the prediction accuracy of GS, the LD was estimated using squared Pearson correlation, r^2 (Vos *et al.*, 2017). Correlations were calculated on tetraploid dosage (0, 1, 2, 3, and 4) among marker pairs for three reference genomes, *i.e.*, *P. hallii*, *P. virgatum*, and *S. viridis*, which have chromosome information. The average r^2 between adjacent markers were calculated and, subsequently, the pairwise correlations were pooled over all chromosomes for each reference genome.

4.3.4 Genomic prediction models considering dosage

Here we generalized well known GS models for diploid to tetraploid species using the information of tetraploid allelic dosage. We evaluated both bayesian and frequentist approaches: Bayesian Ridge Regression (BRR) (Whittaker *et al.*, 2000; Meuwissen *et al.*, 2001); Bayes A (Meuwissen *et al.*, 2001); Bayes B (Meuwissen *et al.*, 2001); Bayes C (Habier *et al.*, 2011); Bayesian LASSO (BL) (Park and Casella, 2008); and Genomic Best Linear Unbiased Predictor (GBLUP) (VanRaden, 2008).

All bayesian models expanded for tetraploid allelic dosage (TD models) share the same predictive multiple linear regression model,

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

where, $\boldsymbol{y}$ $(n\times1)$ is the n adjusted entry mean response vector (from the phenotypic analysis in the previous step); 1_n is a vector of 1's; μ is a scalar representing the population mean; $\boldsymbol{X}$ $(n \times p)$ is the tetraploid allelic dosage incidence matrix of p marker loci coded as $x_{ij} \subset \{0, 1, 2, 3, 4\}$ according to the copy number of reference allele; $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of (unknown) marker with tetraploid dosage genetic (TDG) effects; and $\boldsymbol{\varepsilon}$ $(n \times 1)$ is the vector of residual effects, $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2_\varepsilon)$.

Different assumptions of TDG effects were evaluated. BRR-TD model (Whittaker *et al.*, 2000; Meuwissen *et al.*, 2001) assumes that all marker loci share the same normal prior distribution, $\beta_j|\sigma^2_\beta \sim N(0, \sigma^2_\beta)$, where the common genetic variance hyperparameter (σ^2_β) follows a scaled inverse chi-squared hyperprior distribution $\sigma^2_\beta|d.f._\beta, S_\beta \sim \chi^{-2}(d.f._\beta, S_\beta)$, in which $d.f._\beta$ is the number of degrees of freedom and S_β is the scale parameter of the distribution.

BA-TD model (Meuwissen *et al.*, 2001) is an extension of the above model, which assumes that each TDG prior effects follows specific normal densities, $\beta_j|\sigma^2_{\beta_j} \sim N(0, \sigma^2_{\beta_j})$. As before, each specific genetic variance hyperparameter $(\sigma^2_{\beta_j})$ follows a scaled inverse chi-squared distribution. Due to its property, we expect that BA-TD model tends to shrink TDG effects with different prior strength (desirable for highly parameterized models, $p >> n$), as opposed to the BRR-TD that assumes a common genetic variance hyperparameter. Genetically, these assumptions mean that the analyzed traits are controlled by many genes of small effects and few genes of large effects.

BB-TD model (Meuwissen *et al.*, 2001) is an extension of BA-TD model, that takes into account the TDG effects prior as a mixture of two normal densities, $\beta_j|\delta, \sigma^2_{\beta_j}, \pi \sim \pi\, N(0, \delta) + (1-\pi)\, N(0, \sigma^2_{\beta_j})$. In this model, we can interpret the mixture proportion (π) as a known expectation of a Bernoulli random variable, that is, the expectation of which mixture component best describes the TDG effects (Dos Santos *et al.*, 2016). This first mixture component is indexed by the δ genetic variance hyperparameter, which BB-TD assumes as a known infinitesimal small value. The second mixture component describes the hyperprior component of markers with strong genetic signals, which is indexed again by specific genetic variance hyperparameters that follow scaled inverse chi-squared hyperprior distributions.

BC-TD model (Habier *et al.*, 2011) is a parsimonious variant of the BB-TD model, which considers that all TDG effects follows a common mixture of two normal distributions, $\beta_j|\delta, \sigma^2_\beta, \pi \sim \pi\, N(0, \delta) + (1-\pi)\, N(0, \sigma^2_\beta)$. BC-TD model has the same property of the BRR-TD model, that is, all TDG effects follow the same prior distribution, but with a mixture of two normals like BB-TD model. All hyperparameters of this prior have the same interpretation of the ones described above, as well as the same hyperprior assumptions.

BL-TD model (Park and Casella, 2008) assumes that all markers follows specific normal priors, with genetic variance hyperprior given by the product $\sigma^2_\varepsilon\sigma^2_{\beta_j}$. However, the key difference of BL-TD is the assumption that one component of genetic variance hyperparameters follows exponential distributions, $\sigma^2_{\beta_j}|\lambda_\beta \sim Exp(\lambda_\beta)$. The hyperparameter λ_β measures the knowledge (precision) about the genetic variance hyperparameter. Finally, as the usual proce-

dure for all the above models, we assume the residual genetic variance hyperparameter follows $\sigma^2_\varepsilon | d.f._\varepsilon, S_\varepsilon \sim \chi^{-2}(d.f._\varepsilon, S_\varepsilon)$.

To fit all bayesian models, we used the R package BGLR (de los Campos and Rodriguez, 2015), choosing the default package settings for all known hyperparameters. To obtain the posterior distribution of the unknown parameters and hyperparameters, we used the Gibbs sampler with 20,000 iterations; the first 2,000 cycles were discarded as burn in.

We also evaluated the frequentist model GBLUP-TD (VanRaden, 2008):

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

where $\boldsymbol{y}$, 1_n, μ, and ε_{ij} were the same; $\boldsymbol{Z}$ $(n \times n)$ is the indicence matrix, and $\boldsymbol{g}$ $(n \times 1)$ is the vector mapping the individuals total dosage genetic effects (random effect). This model assumes that the random variable $\boldsymbol{g}$ follows a multivariate normal distribution, $\boldsymbol{g} \sim MVN(\boldsymbol{0}_n, \boldsymbol{K}\sigma^2_g)$, where σ^2_g represents the genetic variance of the population and $\boldsymbol{K}$ is the genomic relationship matrix (VanRaden, 2008), which for tetraploid organisms we derive as $\boldsymbol{K} = \dfrac{\boldsymbol{WW^T}}{tr(\boldsymbol{WW^T})/n}$ and w_{ij} is:

$$w_{ij} \begin{cases} 4-4\,p_j & \text{for AAAA} \\ 3-4\,p_j & \text{for AAAa} \\ 2-4\,p_j & \text{for AAaa} \\ 1-4\,p_j & \text{for Aaaa} \\ 0-4\,p_j & \text{for aaaa} \end{cases}$$

where p_j is the reference allelic frequency at loci j.

Each $k_{i'i}$ on $\boldsymbol{K}$ can be interpreted as a correlation between genotypes of different individuals (genomic relationship), end each k_{ii} as the correlation of the genotypes of a individual with himself (inbreeding).

The GBLUP-TD model was analyzed using R package sommer (Covarrubias-Pazaran, 2016), considering the Newton-Raphson (NR) algorithm for estimating variance components.

4.3.5 Model evaluation

The best statistical model was selected for each of the phenotypic trait analyzed. Cross-validation with 5-folds has been repeated 100 times for bayesian approaches (computationally intensive) and 1,000 times for the frequentist approach, to obtain an asymptotic empirical distribution of the predictive ability. In each replication, the population was randomly split into 5 disjoint subsets of genotypes. Whereas one subset was used as validation population (20% or 106 individuals), the remaining four were combined as training population (80% or 424 individuals) to predict the left-out genotypes in the first population. Subsequently, another subset was used as validation population and the left-out genotypes of this set were predicted. These steps were repeated until all five subsets were used as validation population once.

We calculated Pearson correlation between observed ($\boldsymbol{y}$) and predicted ($\hat{\boldsymbol{y}}$) adjusted entry means considering, simultaneously, all five cross-validations of each replication. Predictive ability was calculated as the mean of these correlations. In addition, we also derived the empirical

distribution of the predicted residual error sums of squares (PRESS), given by the sum of squares of the differences between the predicted and observed adjusted entry means.

4.4 Results

4.4.1 Phenotypic models

The selected VCOV model for the $\boldsymbol{R}_C$ matrix based on AIC and BIC criteria coincided for all nutritive values ($AR1_{Het}$) and for leaf dry matter (UNST) (Table 4.2). However, different VCOVs were selected for PLB (UNST was selected by AIC and $AR1_{Het}$ by BIC). As the differences of AIC values between $AR1_{Het}$ and UNST were greater than the respective BIC differences, the UNST model was selected for PLB. Despite UNST model requires estimation of a larger number of parameters, it had the smallest AIC for PLB. Furthermore, both $AR1_{Het}$ and UNST account for residual correlations and heterogeneous residual variation across harvests. Generalized heritability ranged from 0.3162 to 0.8930. Leaf dry matter showed the higher heritability and in vitro digestibility of organic matter showed the lower heritability. On average, forage production traits showed higher heritability than nutritive values.

4.4.2 SNP calling

Approximately 485,195,807 reads per lane were obtained with the genomic sequence, in which 81.73% had good quality (mean of good, barcoded reads per lane). Initially, the number of tags per lane on average was 15,121,979, and after merge multiple tags, the number of tags was 6,596,939. The alignment results using *P. hallii*, *P. virgatum*, *S. italica*, *S. viridis*, and two transcriptomes of *P. maximum*, obtained by EMBRAPA and UNICAMP research groups (Table 4.3) showed that the overall alignment ranged from 19.05% to 24.24% aligned tags. Although transcriptomes obtained by UNICAMP had the highest overall alignment rate, transcriptomes obtained by EMBRAPA had the highest unique alignment rate, followed by *S. viridis* and *S. italica* (Appendix Table A.6). The number of haplotypes on average was 5,062,583, and the number of unique tags retained was 6,374,151. In the end of this step, 476,904 markers were obtained, with minimum minor allele frequency of 0.01 and minimum minor allele count of 1000. The final number of markers in Tassel-GBS pipeline for each reference genome (Table 4.3) was different in each case.

Due to the nature of GBS technique, the sequencing coverage of different samples is random. It is possible that the same genomic region was not sequenced for all samples. Furthermore, sequences that have mutation in the restriction site of the enzyme also are not observed (Elshire *et al.*, 2011). Therefore, large amount of missing data is expected. The proportion of missing data for the markers (Figure 4.1) showed that the most part of selected markers had 0% of missing data. The reference genome of *S. italica*, *S. viridis*, and transcriptome obtained by EMBRAPA had more than 15,000 markers with 0% of missing data. *P. hallii*, *P. virgatum*, *S. italica*, and *S. viridis* had around 5,000 markers with 87% of missing data. Markers with high proportion of missing values were eliminated in subsequent steps.

Table 4.2. Values of AIC and BIC criteria for the $\boldsymbol{R}_C$ matrix, considering different VCOV structures, as well as generalized heritabilities for all evaluated traits. Traits were organic matter (OM), crude protein (CP), in vitro digestibility of organic matter (IVD), leaf dry matter (LDM), and percentage of leaf blade (PLB).

Trait	H^2_C	$\boldsymbol{R}_C$ matrix	n_{PAR}	AIC	BIC
OM	0.6237	ID	1	2641.55	2658.26
		DIAG	4	2573.20	2606.62
		CS	2	2640.54	2662.82
		CS_{Het}	5	2573.33	2612.32
		AR1	2	2636.48	2658.76
		AR1_{Het}	5	**2553.31**	**2592.30**
		Po	2	2614.54	2636.82
		Po_{Het}	5	2563.48	2602.47
		UNST	10	2555.51	2622.36
CP	0.3658	ID	1	3457.56	3474.27
		DIAG	4	3388.20	3421.62
		CS	2	3450.74	3473.02
		CS_{Het}	5	3379.22	3418.21
		AR1	2	3430.66	3452.94
		AR1_{Het}	5	**3352.82**	**3391.81**
		Po	2	3411.95	3434.23
		Po_{Het}	5	3374.20	3413.19
		UNST	10	3358.94	3425.79
IVD	0.3162	ID	1	8199.83	8216.54
		DIAG	4	8159.68	8193.10
		CS	2	8198.89	8221.18
		CS_{Het}	5	8158.56	8197.55
		AR1	2	8179.69	8201.97
		AR1_{Het}	5	**8143.20**	**8182.19**
		Po	2	8191.85	8214.14
		Po_{Het}	5	8145.05	8184.04
		UNST	10	8147.61	8214.45
LDM	0.8930	ID	1	45909.45	45928.68
		DIAG	8	44985.31	45049.39
		CS	2	45727.46	45753.09
		CS_{Het}	9	44203.65	44274.14
		AR1	2	44776.89	44802.52
		AR1_{Het}	9	43625.89	43696.38
		Po	2	44600.72	44626.35
		Po_{Het}	9	43721.82	43792.31
		UNST	36	**43328.46**	**43571.97**
PLB	0.5037	ID	1	27036.63	27055.85
		DIAG	8	26406.88	26470.96
		CS	2	27029.13	27054.77
		CS_{Het}	9	26403.98	26474.47
		AR1	2	27016.46	27042.10
		AR1_{Het}	9	26400.12	**26470.61**
		Po	2	27025.02	27050.66
		Po_{Het}	9	26408.88	26479.37
		UNST	36	**26306.32**	26549.84

Table 4.3. Total of markers obtained in the different cenarios.

Reference Genome	Overall alignment rate	Tassel-GBS	SuperMASSA	Filter NA
Panicum hallii	19.05%	77,105	12,835	6,945
Panicum virgatum	22.66%	84,119	11,230	5,598
Setaria italica	22.04%	92,494	15,047	8,066
Setaria viridis	22.07%	92,591	15,271	8,118
Transcriptome (EMBRAPA)	20.11%	74,049	14,129	7,665
Transcriptome (UNICAMP)	24.24%	56,546	9,777	5,032
Total	–	476,904	78,289	41,424

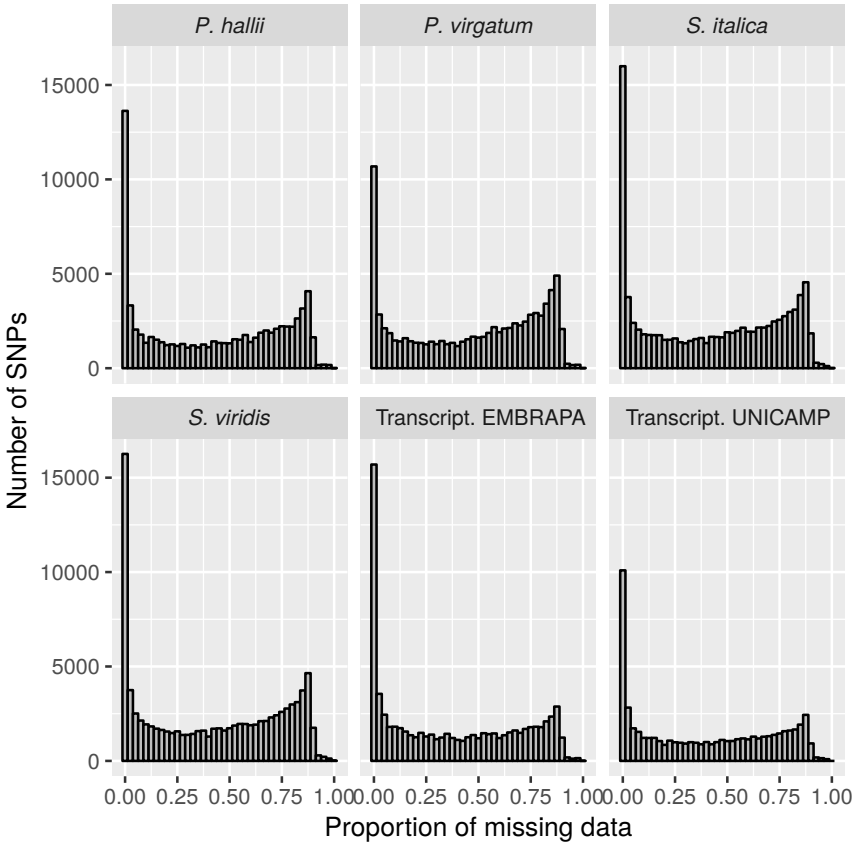

Figure 4.1. Proportion of missing data in GBS markers using *P. hallii*, *P. virgatum*, *S. italica*, *S. viridis*, and two transcriptomes of *P. maximum* (obtained by EMBRAPA and UNICAMP research groups) as reference genomes.

As mentioned before, the output of Tassel-GBS pipeline was used as input in the SuperMASSA software, considering the Generalized Population Model, fiting and filtering for ploidy 4. A total of 78,289 markers was selected with minimum overall depth of 25 reads (Table 4.3). From this, 32,619 markers had more than 100 of minimum overall depth. For example, marker Hallii.1_3461595 (Figure 4.2) shows the intuition of how SuperMASSA uses the ratio of the intensity of two alleles to classify individuals according to their genotype using a probabilistic graphical model (Serang *et al.*, 2012). The name of the markers was formed by: reference genome plus number of chromosome plus position of SNP in the chromosome.

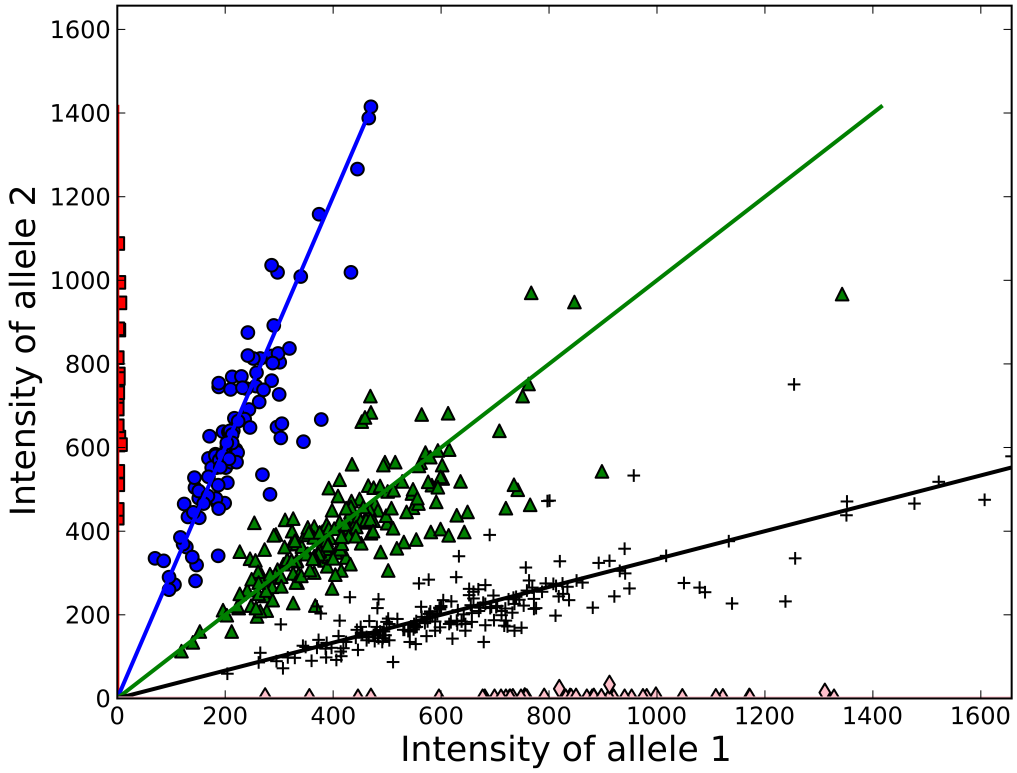

Figure 4.2. Marker Hallii.1_3461595. Red squares represent offsprings with the allele A in homozygous, *i.e.* AAAA (quadriplex or dose 4). Blue circles represent offsprings AAAa (triplex). Green triangles represent offsprings AAaa (duplex). Black crosses represent offsprings Aaaa (simplex). And pink rhombus represent offsprings aaaa (nulliplex, or dose 0 of A).

Handling datasets with high level of missing data is complex and relies heavily on imputation methods; in addition, the accuracy of which with high missing genotypes can be variable (Fu, 2014). Therefore, markers were selected with up to 5% of missing data, aiming to reduce imputation bias. The final number of markers was 41,424, which were used in GS models (Table 4.3). Subsequently, imputation was made using random sampling and considering the dose proportion for each marker.

The redundancy among markers was inspected (Figure 4.3) and a greater similarity was verified between three specific groups of reference: *Panicum* genus, *Setaria* genus, and transcriptomes. This result is expected due to phylogenetic proximity of groups. More than half of the markers identified by the different reference genomes have non-redundant information, and 31,046 markers were classified as unique. This may due to the great genomic variability still persistent in each genome, since they are species with a broad genetic base, and relatively new to breeding programs.

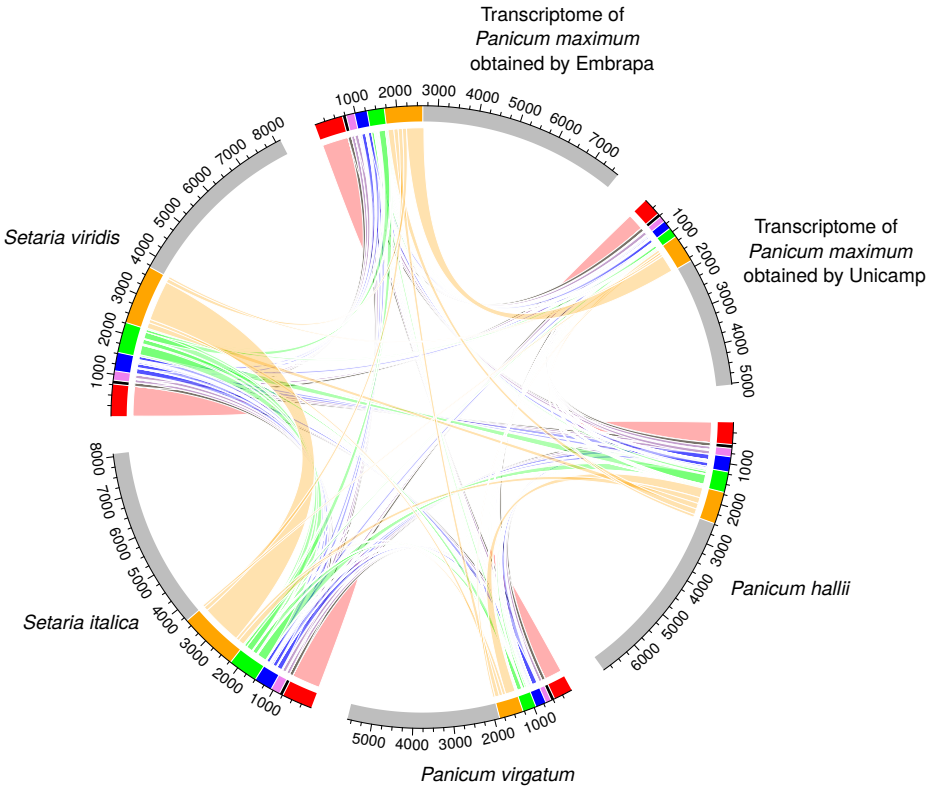

Figure 4.3. Circular graph showing redundancy among markers (Gu *et al.*, 2014). Regions in red represent redundant markers within each reference, while regions in black, pink, blue, green, and orange represent, respectively, redundant markers among six, five, four, three, and two references. Gray regions represent markers with unique information for each reference.

4.4.3 Linkage disequilibrium

Average linkage disequilibrium (LD) between adjacent markers was 0.2422, 0.2176, and 0.2457 for markers identified by alignment with *P. hallii* (Figure 4.4A), *P. virgatum* (Figure 4.4B), and *S. viridis* (Figure 4.4C), respectively. A high degree of LD was observed even for extended distances between markers, being more dispersed for *P. virgatum* and less for *S. viridis*. This is expected since *P. virgatum* has a higher genome size and *S. viridis* obtained a higher density of markers by pairs of base in relation to the others.

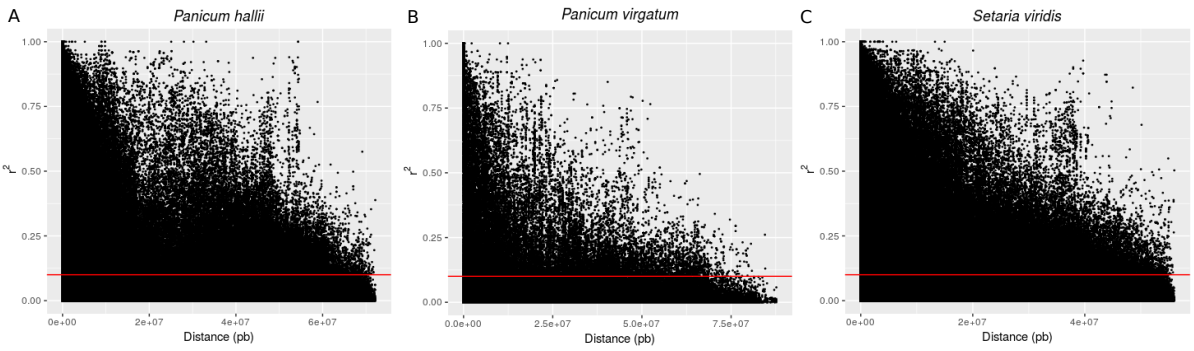

Figure 4.4. Linkage disequilibrium calculated as squared Pearson correlation, r^2, for three reference genomes: (A) *P. hallii*. (B) *P. virgatum*. (C) *S. viridis*. Red lines are $r^2 = 0.1$.

Table 4.4. Mean predictive ability of GS-TD models considering tetraploid allelic dosage and applied to prediction of organic matter (OM), crude protein (CP), in vitro digestibility of organic matter (IVD), leaf dry matter (LDM), and percentage of leaf blade (PLB). Lower and upper limits are in parentheses.

	Nutritive values			Yield trait	Structural trait
Model	OM	CP	IVD	LDM	PLB
GBLUP-TD	0.4591	0.3332	0.3310	0.2101	0.3015
	(0.4316; 0.4933)	(0.3025; 0.3653)	(0.2944; 0.3693)	(0.1707; 0.2504)	(0.2635; 0.3409)
BRR-TD	0.4601	0.3331	0.3305	0.2109	0.3315
	(0.4336; 0.4939)	(0.3047; 0.3609)	(0.2891; 0.3672)	(0.1759; 0.2531)	(0.2879; 0.3653)
BA-TD	0.4629	0.3312	0.3279	0.2030	0.3316
	(0.4342; 0.4950)	(0.2980; 0.3581)	(0.2917; 0.3717)	(0.1610; 0.2377)	(0.2866; 0.3644)
BB-TD	0.4668	0.3331	0.3289	0.2092	0.3311
	(0.4425; 0.4988)	(0.3053; 0.3608)	(0.2929; 0.3741)	(0.1711; 0.2482)	(0.2808; 0.3636)
BC-TD	0.4609	0.3332	0.3302	0.2113	0.3309
	(0.4337; 0.4934)	(0.3049; 0.3597)	(0.2883; 0.3682)	(0.1750; 0.2506)	(0.2828; 0.3627)
BL-TD	0.4585	0.3335	0.3311	0.1691	0.3293
	(0.4347; 0.4932)	(0.3070; 0.3583)	(0.2973; 0.3734)	(0.1262; 0.2104)	(0.2768; 0.3646)
Average	0.4614	0.3329	0.3299	0.2021	0.3260

4.4.4 Genomic prediction

Genomic selection models were analyzed using all markers (41,424 markers) and only non-redundant ones (31,046 markers). The predictive accuracy did not differ between these two data sets (results not shown), since the predictive models deal well with multicollinearity. So, only predictive models using all markers will be presented.

Mean values of predictive ability ranged from 0.1691 (BL-TD for leaf dry matter) to 0.4668 (BB-TD for organic matter) (Table 4.4). The predictive ability of structural and nutritive values traits were higher than those of yield trait. Leaf dry matter showed the lowest accuracies for all analyzed GS-TD models and organic matter showed the highest ones (Figure 4.5).

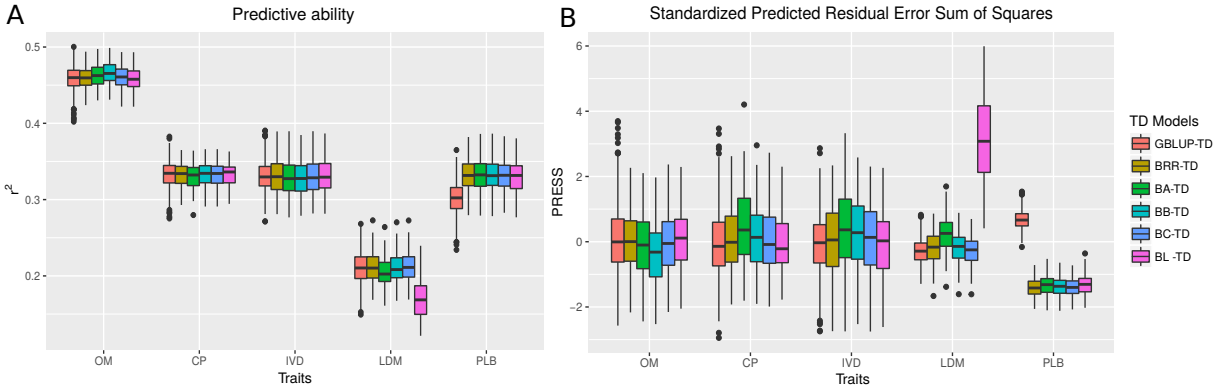

Figure 4.5. Comparison among six GS-TD models which consider tetraploid allelic dosage and applied to prediction of organic matter (OM), in vitro digestibility of organic matter (IVD), crude protein (CP), leaf dry matter (LDM) and percentage of leaf blade (PLB). (A) Predictive ability. (B) Standardized Predicted Residual Error Sum of Squares.

The variation results from 100 and 1,000 replications of 5-folds cross-validation for Bayesian and GBLUP-TD models, respectively. No clear difference in predictive ability was

observed among models for nutritive values (Figure 4.5). BL-TD was different from others for leaf dry matter, and GBLUP-TD for percentage of leaf blade (Figure 4.5A); accordingly, these models obtained higher estimates of standardized PRESS for the corresponding traits (Figure 4.5B). Standardized PRESS were obtained subtracting each values from the mean and dividing by the standard deviation (mean PRESS in its natural scale for each trait and each model are shown in Appendix Table A.7).

4.5 Discussion

The aim of this study was to develop predictive models which consider tetraploid allelic dosage to estimate breeding values for genomic selection in *P. maximum*. We compared the accuracy of predicted breeding values using six different models, and evaluated strategies for modeling residual effects and performing SNP calling. This methodology can be applied to other autotetraploid species and can be extended to species with other ploidy levels.

Before the development of GS models considering tetraploid allelic dosage (GS-TD models), it is important to perform a precise phenotypic analysis and high throughput genotyping to achieve high levels of predictive accuracy. As highlighted by Cabrera-Bosquet *et al.* (2012), the sucess of GS in breeding for quantitative traits largely depends on a reliable phenotyping process. Therefore, the precise phenotypic data is one of the key components to train GS models for accurately predicting GEBV of the breeding population.

In order to obtain adjusted means of the traits free of experimental residual effects for the genomic selection analysis, we performed a two stage approach for the analysis. In the first stage, for each trait we fitted a longitudinal multivariate linear mixed (LMLM) model. Considering all nutritive values analyzed, the first-order autoregressive heterogeneous (AR_{Het}) structure provided a better fit than other models (Table 4.2). This structure allows to model residual correlations and heterogeneity of variance, wherein the correlations between harvests decay with time and each harvest has its own residual variance. For LDM and PLB, the selected model has an unstructured (UNST) variance and covariance matrix (Table 4.2), which allows specific residual variances and covariances for each harvest. In the second stage, molecular markers were considered in the predictive models. For this, the SNP calling took into consideration the allelic dosage, discriminating among the five possible genotypes. According to Uitdewilligen *et al.* (2013), a high sequence depth is required to identify the correct genotypic class accurately, where 60–80 coverage leads to 98.4% accuracy in genotypic calls. A test was performed using a minimum overall depth of 25 and 100 reads, and the predictive ability of GS-TD models was similar for both criteria (results not shown), probably because most markers had a good genotype quality; approximately 78.7% of markers selected with minimum overall depth of 25 reads were also selected with overall depth of 100 reads.

The underlying assumption of genomic selection is the presence of SNPs at some loci in linkage disequilibrium (LD) with QTL alleles that affect the traits that are subject to selection (Calus *et al.*, 2008). According to Vos *et al.* (2017), LD is the non-random association between alleles at different loci in a breeding population. It can be estimated using the correlation between markers when the SNP alleles at those loci have numerical values. These authors calculated several estimators for LD in a simulated and real panel of tetraploid potato and concluded

that $LD_{1/2,90}$ values provides the most consistent estimates of LD decay. This estimator consists of 90% percentile of r^2 the short-range LD. Short-range LD is calculated across a defined interval of genetic distances between marker pairs (Vos *et al.*, 2017). Here we used the squared Pearson correlation (r^2) as an estimate of LD.

According to Riedelsheimer *et al.* (2012), one major reason of the minor differences in prediction accuracies among prediction models is the high level of LD found in breeding population. The authors obtained similar accuracies of GS models in elite maize germoplasm, which had high level of LD. Accuracies did not differ independently whether the effect of large QTL were precisely captured or spread over a larger region. In this work, we also obtained a high level of LD between marker pairs ($r^2 = 0.22$), which can explain the similarity among GS-TD models for prediction purposes.

The prediction of the breeding values was made using estimated means considering all harvests simultaneously. The goal was to select individuals in the present recurrent selection cycle (already phenotypically evaluated) as well as to select non-phenotyped individuals from the next generation recurrent cycle. Since one cycle requires three to five years of evaluations (Resende *et al.*, 2014), *P. maximum* breeding program with genomic selection will reduce approximately four years for each recurrent cycle. Therefore, superior sexual plants can be selected every year to cross with apomictic plants, obtaining new apomictic hybrid combinations to test as new cultivars and to release the best one in agricultural marketing. Lipka *et al.* (2014) applied genomic selection in *P. virgatum* L. species considering diploid dosage with the objective of evaluating genomic selection efficiency to accelerate breeding cycles in this species. The authors obtained high prediction accuracy for most of the traits, in which seven were morphological traits and thirteen were traits related to biomass quality. Similar to our results, they observed the same prediction accuracies across GS approaches. Although analyzed traits were different from ours, the range of values were similar as well. The higher mean prediction accuracy obtained by Lipka *et al.* (2014) was 0.52 for standability and the lower was -0.08 for minerals. Our higher mean predictive ability was 0.46 for organic matter and the lower was 0.20 for leaf dry matter (Table 4.4).

A comparison of GS-TD models with GS models considering the usual diploid allelic dosage (GS-DD models) was also performed here. To do so, the three possible heterozygotes (Aaaa, AAaa, and AAAa) were coded as diploid heterozygote (Aa), while the two tetraploid homozygotes (AAAA and aaaa) as diploid homozygotes (AA and aa). Usual GBLUP model for diploid species was evaluated with R package sommer (Covarrubias-Pazaran, 2016), considering the genomics relationship matrix as described by VanRaden (2008). The mean predictive ability (Appendix Table A.8) showed no clear difference between GBLUP-TD and GBLUP models. The lack of differences between these two models might be due to the genetic structure of the breeding population. The individuals analyzed have a high level of relationship because they constitute families of half-sibs; also the majority of genotypes were classified as aaaa or Aaaa (84.27% and 12.57%, respectively). Hence, we recommend repeating our study in more heterozygous populations to investigate how much predictive accuracy increases when considering tetraploid allelic dosage. It is worth mentioning that using tetraploid allelic dosage relies on more realistic assumptions of the genetic architecture on autotetraploid species.

To our knowledge, this is the first study that applied genomic selection in *P. maximum* including dosage in predictive models. Previously, Slater *et al.* (2016) developed an extension of genomic relationship matrix proposed by Yang *et al.* (2010) for autotetraploids and applied genomic selection in potato. The authors achieved accuracies ranging from 0.2, under conditions of low heritability and small reference populations, to 0.8 in larger reference populations.

In our work, mean predictive ability ranged from 0.1691 (BL-TD for LDM) to 0.4668 (BB-TD for OM) (Table 4.4). Similar accuracy was obtained for oats (Asoro *et al.*, 2011), maize (González-Camacho *et al.*, 2012), rice (Spindel *et al.*, 2015), and potato (Slater *et al.*, 2016). Despite this difference among traits, the accuracy did not differ among models. Besides that, we suggest to use BRR-TD model for PLB because it presented lower PRESS (Appendix Table A.7) and to use GBLUP-TD model for all remaining traits (since it requires less computational time).

This is the first work of genomic selection in *P. maximum* which uses a high throughput genotyping and considers tetraploid allelic dosage in bayesian and frequentist models. GBS and allelic dosage showed to be promising strategies for genomic analysis in autotetraploid species. Furthermore, the accuracy of predictive models justifies the implementation of genomic selection in *P. maximum* breeding programs.

Acknowledgments

This work was supported by FAPESP (São Paulo Research Foundation), grants 2015/20659-2 and 2016/01279-7. Additional support was provided by CNPq ("Conselho Nacional do Desenvolvimento Científico e Tecnológico"), EMBRAPA ("Empresa Brasileira de Pesquisa Agropecuária"), and by UNIPASTO ("Associação para o Fomento à Pesquisa de Melhoramento de Forrageiras"). The authors thanks Mariane de Mendonça Vilela (Embrapa Beef Cattle) by the assistance in the DNA extraction step, and all trainees of Embrapa Beef Cattle that helped on the phenotypic evaluations.

References

Akaike, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**: 716–723.

Allard, R. W., 1960 Inheritance in autotetraploids. In *Principles of Plant Breeding*, edited by J. W. . Sons, pp. 385–399.

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 **4**: 132–144.

Bernardo, R., 2014 Genomewide selection when major genes are known. Crop Science **54**: 68–75.

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel, 2009 *ASReml-R reference manual*.

Cabrera-Bosquet, L., J. Crossa, J. von Zitzewitz, M. D. Serret, and J. L. Araus, 2012 High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. Journal of Integrative Plant Biology **54**: 312–320.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. Roos, and R. F. Veerkamp, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178**: 553–561.

Covarrubias-Pazaran, G., 2016 Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE **11**: e0156744.

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 Enviromental Statistics **11**: 381–393.

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. and P. P. Rodriguez, 2015 BGLR: Bayesian generalized linear regression. R package version 1.0.4 .

Desta, Z. A. and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science **19**: 592–601.

Dos Santos, J. P. R., L. P. M. Pires, R. C. d. C. Vasconcellos, G. S. Pereira, R. G. V. Pinho, and M. Balestre, 2016 Genomic selection to resistance to *Stenocarpella maydis* in maize lines using DArTseq markers. BMC Genetics **17**: 1–10.

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.

Fu, Y.-B., 2014 Genetic diversity analysis of highly incomplete SNP genotype data with imputations: An empirical assessment. G3 (Genes|Genomes|Genetics) **4**: 891–900.

Garcia, A. A. F., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva, M. L. C. Vieira, R. Vicentini, E. A. Costa, M. C. Mancini, M. O. S. Garcia, M. M. Pastina, R. Gazaffi, E. R. F. Martins, N. Dahmer, D. A. Sforça, C. B. C. Silva, P. Bundock, R. J. Henry, G. M. Souza, M.-A. van Sluys, M. G. A. Landell, M. S. Carneiro, M. A. G. Vincentz, L. R. Pinto, R. Vencovsky, and A. P. Souza, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Scientific reports **3**: 1–10.

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.

González-Camacho, J. M., G. De los Campos, P. Pérez, D. Gianola, J. E. Cairns, G. Mahuku, R. Babu, and J. Crossa, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics **125**: 759–771.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar, 2011 Phytozome: a comparative platform for green plant genommics. Nucleic Acids Research **40**: D1178–D1186.

Gu, Z., L. Gu, R. Eils, M. Schlesner, and B. Brors, 2014 *circlize* implements and enhances circular visualization in R. Bioinformatics **30**: 2811–2812.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**: 186–197.

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.

Langmead, B. and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nature Methods **9**: 357–359.

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.

McCoy, T. J. and E. T. Bingham, 1988 Cytology and cytogenetics of alfalfa. In *Alfalfa and alfalfa improvement*, edited by B. D. K. H. R. R. Hanson, A. A., volume 29, pp. 737–776.

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.

Mollinari, M. and O. Serang, 2013 Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. In *Methods in Molecular Biology*, edited by J. M. Walker.

Osborn, T. C., J. Chris Pires, J. A. Birchler, D. L. Auger, Z. Jeffery Chen, H.-S. Lee, L. Comai, A. Madlung, R. W. Doerge, V. Colot, and R. A. Martienssen, 2003 Understanding mechanisms of novel gene expression in polyploids. Trends in Genetics **19**: 141–147.

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

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.

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**: 1–9.

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

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS ONE **7**: e30906.

Slater, A. T., N. O. I. Cogan, J. W. Forster, B. J. Hayes, and H. D. Daetwyler, 2016 Improving genetic gain with genomic selection in autotetraploid potato. The Plant Genome **9**: 1–15.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña, 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 line. PLoS ONE **11**: e1004982.

Toledo-Silva, G., C. B. Cardoso-Silva, L. Jank, and A. P. Souza, 2013 De Novo transcriptome assembly for the tropical grass *Panicum maximum* Jacq. PLoS ONE **8**: e70781.

Uitdewilligen, J. G. A. M. L., A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser, and H. J. van Eck, 2013 A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE **8**: e62355.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Vos, P. G., M. João Paulo, R. E. Voorips, R. G. F. Visser, H. J. van Eck, and F. A. van Eeuwijk, 2017 Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theoretical and Applied Genetics **130**: 123–135.

Warmke, H. E., 1954 Apomixis in *Panicum maximum*. American Journal of Botany **41**: 5–11.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genetical research **75**: 249–252.

Wu, K. K., W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore, and S. D. Tanksley, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics **83**: 294–300.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics **42**: 565–569.

APPENDIX

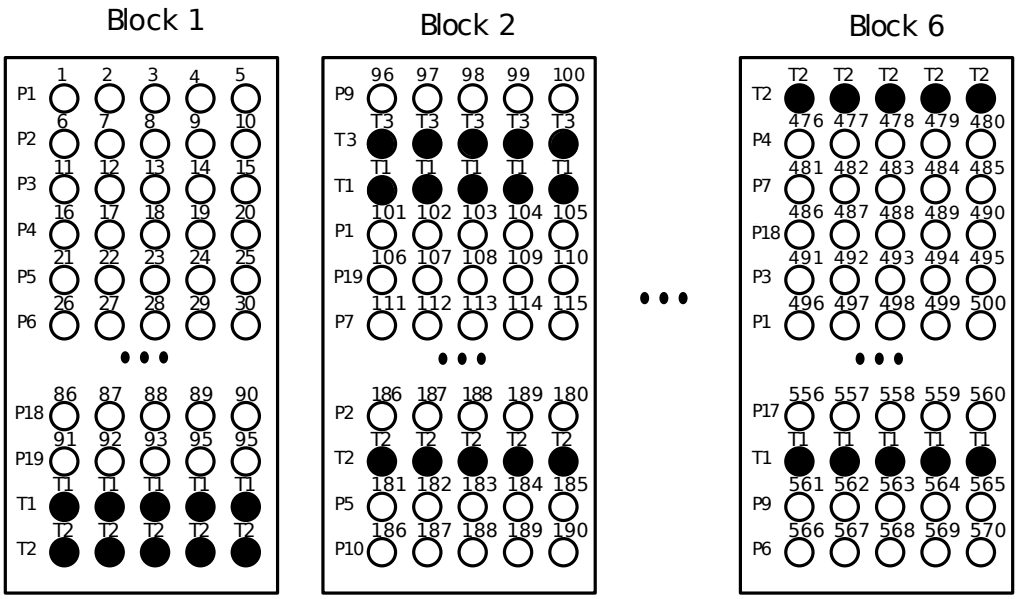

Figure A.6. Illustative scheme to demonstrate the augmented block design (ABD). Empty circles are tetraploid sexual offsprings and solid circles are tetraploid apomitic checks. Parents are from P1 to P19 and checks are from T1 to T3.

Table A.5. Variance and covariance (VCOV) structures tested for genetic ($\boldsymbol{G}_C$) and error ($\boldsymbol{R}_C$ and $\boldsymbol{R}_B$) effects to the models 2 and 3 for acid detergent fiber (ADF). The selection was performed into three steps: first we fitted the $\boldsymbol{G}_C$ matrix for differents structures; second we fitted the $\boldsymbol{R}_C$ matrix given the selected $\boldsymbol{G}_C$ previously; and third we fitted the $\boldsymbol{R}_B$ matrix given the selected $\boldsymbol{G}_C$ and $\boldsymbol{R}_C$. The n_{PAR} and NC are the number of parameters for each VCOV structure and the not convergence of the model, respectively.

Model 2: BLUP/REML model without additive relationship matrix

$\boldsymbol{G}_C$					$\boldsymbol{R}_C$					$\boldsymbol{R}_B$				
Matrix	n_{PAR}	AIC	BIC	Deviance	Matrix	n_{PAR}	AIC	BIC	Deviance	Matrix	n_{PAR}	AIC	BIC	Deviance
ID	1	10400.11	10417.57	5819.64	ID	1	10349.89	**10390.62**	5761.41	ID	1	**10347.79**	**10394.34**	5757.31
DIAG	4	10388.93	10423.84	5802.45	DIAG	4	10351.41	10409.61	5756.93	DIAG	NC	–	–	–
CS	2	10358.91	**10382.19**	5776.43	CS	2	10351.60	10398.15	5761.12	CS	2	10349.66	10402.04	5757.19
CS_{Het}	5	**10349.89**	10390.62	5761.41	CS_{Het}	NC	–	–	–	CS_{Het}	NC	–	–	–
Po	2	10358.16	10385.43	5775.68	Po	2	**10347.79**	10394.34	5757.31	FA1	NC	–	–	–
FA1	8	10353.75	10411.95	5759.27	FA1	NC	–	–	–					

Model 3: BLUP/REML model using additive relationship matrix

$\boldsymbol{G}_C$					$\boldsymbol{R}_C$					$\boldsymbol{R}_B$				
Matrix	n_{PAR}	AIC	BIC	Deviance	Matrix	n_{PAR}	AIC	BIC	Deviance	Matrix	n_{PAR}	AIC	BIC	Deviance
ID	1	10399.71	10417.17	5819.24	ID	1	10354.88	**10378.16**	5772.40	ID	1	**10344.34**	10385.08	5755.86
DIAG	4	10394.36	10429.28	5807.88	DIAG	4	**10344.34**	10385.08	5755.86	DIAG	6	10390.68	**10360.52**	5692.20
CS	2	10355.38	10378.66	5772.91	CS	2	10356.88	10385.97	5772.40	CS	2	10346.32	10392.88	5755.85
CS_{Het}	5	10357.77	10388.51	5759.30	CS_{Het}	5	10346.29	10392.84	5755.81	CS_{Het}	7	10392.67	10368.32	5692.19
Po	2	10354.88	**10378.16**	5772.40	Po	2	10356.18	10385.28	5771.70	FA1	NC	–	–	–
FA1	8	**10353.44**	10411.63	5758.96	FA1	NC	–	–	–					

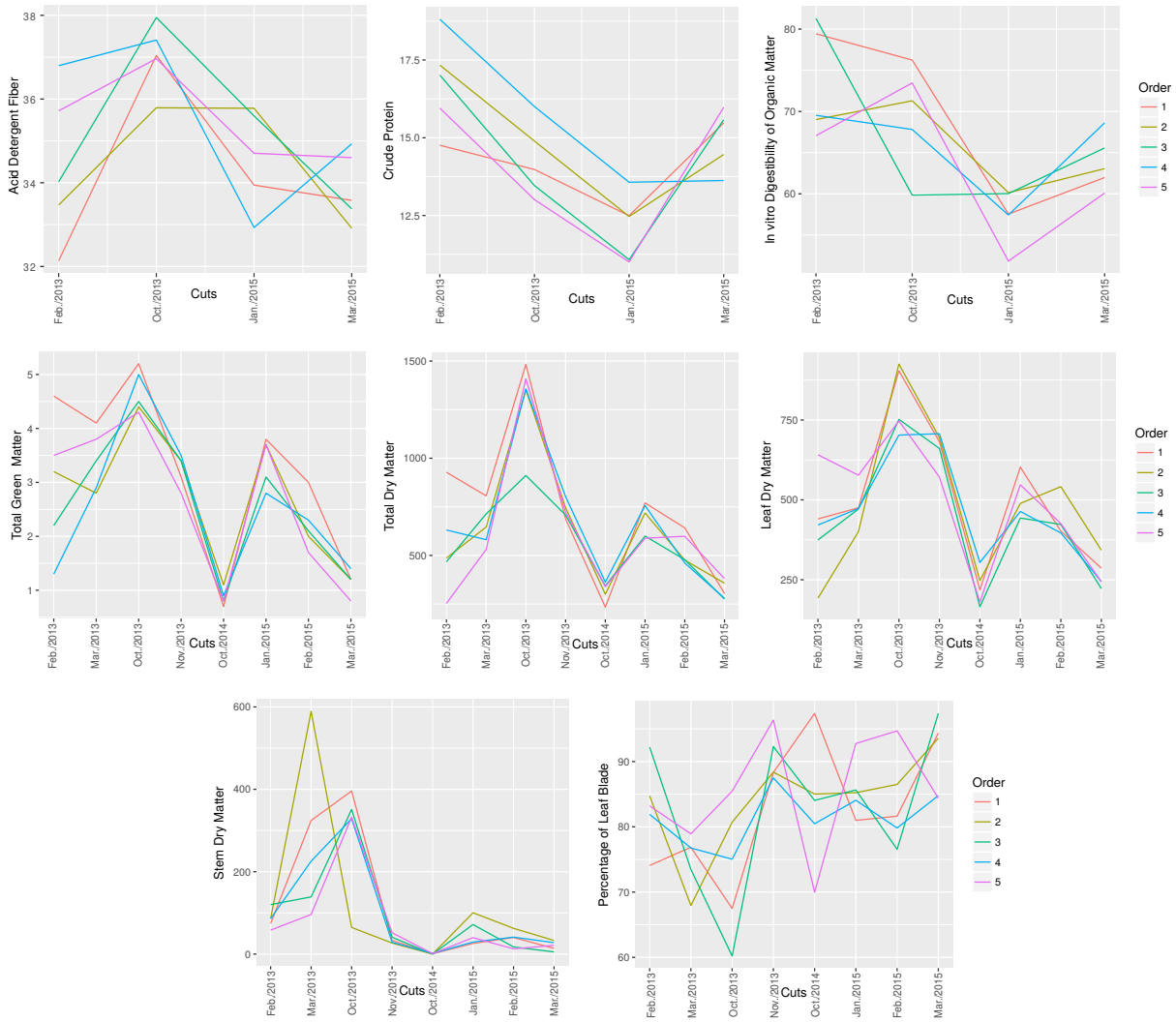

Figure A.7. Response of five selected offsprings by Model 3 in each harvest for acid detergent fiber (ADF), crude protein (CP), in vitro digestibility of organic matter (IVD), total green matter (TGM), total dry matter (TDM), leaf dry matter (LDM), stem dry matter (SDM), and percentage of leaf blade (PLB).

Table A.6. Bowtie2 alignment results of 6,596,939 GBS tags in absolute and relative (in parentheses) values

Reference Genome	Non-aligned tags	Aligned tags		
		Overall alignment	Unique alignment	Non-unique alignment
Panicum hallii genome	5,340,535 (80.95%)	1,256,404 (19.05%)	1,002,261 (15.19%)	254,143 (3.85%)
Panicum virgatum genome	5,101,776 (77.34%)	1,495,163 (22.66%)	503,124 (7.63%)	992,039 (15.04%)
Setaria italica genome	5,143,121 (77.96%)	1,453,818 (22.04%)	1,149,693 (17.43%)	304,125 (4.61%)
Setaria viridis genome	5,141,196 (77.93%)	1,455,743 (22.07%)	1,164,462 (17.65%)	291,281 (4.42%)
Transcriptome (EMBRAPA)	4,997,950 (75.76%)	1,326,602 (20.11%)	1,244,861 (18.87%)	81,741 (1.24%)
Transcriptome (UNICAMP)	5,270,337 (79.89%)	1,598,989 (24.24%)	839,084 (12.72%)	759,905 (11.52%)

Table A.7. Predicted residual error sum of squares (PRESS) of GS-TD models considering tetraploid allelic dosage and applied to prediction of organic matter (OM), crude protein (CP), in vitro digestibility of organic matter (IVD), leaf dry matter (LDM) and percentage of leaf blade (PLB). Upper and lower limits are in parentheses.

	Nutritional values			Production values	
Model	OM	CP	IVD	LDM	PLB
GBLUP-TD	67.0595	76.8397	758.0740	1332705	10186.02
BRR-TD	66.9673	76.9616	760.0915	1336232	8957.235
BA-TD	66.8326	77.3960	765.5939	1349922	9004.154
BB-TD	66.4725	77.0446	762.2685	1336829	8981.552
BC-TD	66.9092	76.9274	759.9575	1333633	8961.889
BL-TD	67.0827	76.8337	758.5738	1436145	9005.046

Table A.8. Mean predictive ability of GBLUP-TD and GBLUP models for organic matter (OM), crude protein (CP), in vitro digestibility of organic matter (IVD), leaf dry matter (LDM), and percentage of leaf blade (PLB). Lower and upper limits are in parentheses.

Model	GBLUP-TD	GBLUP
OM	0.4591	0.4382
	(0.4316; 0.4933)	(0.4074; 0.4701)
CP	0.3332	0.3258
	(0.3025; 0.3609)	(0.2908; 0.3556)
IVD	0.3310	0.3395
	(0.2944; 0.3693)	(0.3082; 0.3771)
LDM	0.2101	0.2138
	(0.1707; 0.2504)	(0.1741; 0.2528)
PLB	0.3015	0.3106
	(0.2635; 0.3409)	(0.2707; 0.3534)