ISADORA CRISTINA MARTINS OLIVEIRA

ESTUDO DA INTERAÇÃO GENÓTIPO × AMBIENTE E PREDIÇÃO GENÔMICA DE HÍBRIDOS DE SORGO BIOMASSA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

Orientador: José Eustáquio de Souza Carneiro

Coorientadores: Rafael Augusto da Costa Parrella

Pedro Crescêncio Souza Carneiro

VIÇOSA - MINAS GERAIS
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
O48e
2019

Oliveira, Isadora Cristina Martins, 1990-
Estudo da interação genótipo x ambiente e predição genômica de híbridos de sorgo biomassa / Isadora Cristina Martins Oliveira. – Viçosa, MG, 2019.
103f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: José Eustáquio de Souza Carneiro.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Sorgo. 2. Melhoramento genético. 3. Análise fatorial. 4. Genômica. I. Universidade Federal de Viçosa. Departamento de Fitotecnia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22 ed. 633.622

ISADORA CRISTINA MARTINS OLIVEIRA

ESTUDO DA INTERAÇÃO GENÓTIPO × AMBIENTE E PREDIÇÃO GENÔMICA DE HÍBRIDOS DE SORGO BIOMASSA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 06 de setembro de 2019.

Assentimento:

Isadora Cristina Martins Oliveira
Autora

José Eustáquio de Souza Carneiro
Orientador

À minha mãe Berenice, e aos meus irmãos Paulo e Isabela, dedico!

AGRADECIMENTOS

Agradeço primeiramente a Deus, por sempre guiar os meus passos pelos caminhos do bem e me manter integra nos meus princípios. Obrigada por tudo Senhor!

À minha mãe, por sempre me incentivar e lutar comigo para que eu chegasse até aqui. E por me ensinar a ver o mundo de um jeito mais puro, e sempre mantê-lo regado com muito amor. Esta vitória é sua, mãe.

À minha irmã, minha fiel companheira, por estar ao meu lado em todos os momentos e ser minha injeção de coragem diária. Obrigada por tanto.

Ao meu irmão, pela proteção e carinho de sempre.

À Universidade Federal de Viçosa e ao programa de pós-graduação em Genética e Melhoramento pela oportunidade de realizar o doutorado.

Ao meu orientador José Eustáquio pelos ensinamentos e confiança ao me indicar para fazer parte da parceria UFV/Embrapa, o que me fez crescer muito como profissional.

Ao meu coorientador Rafael Parrella, pela orientação e confiança depositada junto ao programa de melhoramento de sorgo bioenergia.

À Dra. Maria Marta, pela orientação nas análises estatísticas e na composição dos artigos. E por ser uma amiga nas horas de desespero. Você alegrava nossos dias no laboratório.

Ao professor Pedro Crescêncio, pela disponibilidade e ensinamentos, você foi um dos meus melhores mestres.

À minha família, pelo amor incondicional. E por sempre estarem ao meu lado me regando de boas energias e tornando meus dias mais leves. Em especial aos “Primos do BOGA”, de onde eu tiro meus espelhos de superação e força de vontade!

Aos colegas e funcionários do Galpão de Melhoramento de Sorgo da Embrapa pela amizade e pelo apoio braçal. Foram momentos que ficarão eternizado na memória e que me fizeram crescer pessoal e profissionalmente. Vocês são demais!

Às minhas amigas, em especial à Isabella, Kamilla, Luciana, Luísa, Marina e Rafaela, que foram compreensivas com minha ausência, e por sempre me darem conforto com palavras. Vocês sabem que mesmo longe estavam sempre presentes no coração. Sou eternamente grata!!!

Ao meus amigos da pós-graduação, pelos dias de luta e alegria que vivemos juntos, em especial ao “bonde” da Bioinformática, que vivenciaram de perto a busca por este título. Agradeço especialmente à Karine e ao Zé Henrique, que estavam sempre presentes e dispostos a ajudar.

Aos órgãos de fomento, uma vez que o presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, do Banco Nacional de Desenvolvimento Econômico e Social (BNDES) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

E a todos que colaboraram direta ou indiretamente para a realização deste trabalho.

MUITO OBRIGADA!

RESUMO

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, setembro de 2019. **Estudo da interação genótipo × ambiente e predição genômica de híbridos de sorgo biomassa**. Orientador: José Eustáquio de Souza Carneiro. Coorientadores: Rafael Augusto da Costa Parrella e Pedro Crescêncio Souza Carneiro.

O sorgo biomassa é uma cultura que vem sendo demandada pelo mercado sucroenergético devido à suas características agronômicas, especialmente tolerância a estresses abióticos e alta acúmulo de biomassa, para a produção de energia. Neste contexto, com base no banco de dados históricos do programa de melhoramento de sorgo bioenergia da Embrapa Milho e Sorgo, buscou-se: i) investigar as fontes de interação genótipo × ambiente (G×A) e selecionar genótipos produtivos e adaptados e/ou estáveis, às diferentes condições ambientais usando análise de fatores; ii) avaliar a acurácia de predição de híbridos usando informação de marcadores SNPs e dados de múltiplos ambientes; iii) comparar a acurácia de modelos aditivos e aditivos-dominantes; e iv) predizer híbridos ainda não testados do programa. Os dados foram genotipados, via genotipagem por sequenciamento (GBS), sendo 17 linhagens R (macho-fértil) e 46 linhagens A (macho-estéril), das quais deram origem à 202 híbridos, que foram fenotipados em diversas regiões no território brasileiro. Devido ao alto desbalanceamento entre os ensaios nos diferentes anos, e entre os ensaios preliminares (APH) e finais (VCU) de avaliação de híbridos, adotou-se os métodos de modelos mistos para análise dos dados, com o auxílio do pacote ASReml disponível no software R. Constatou-se alta interação G×A entre os ambientes avaliados, com correlações genotípicas variando de -0,09 a 0,68. Por meio da análise de PCA os dois primeiros fatores explicaram 48.81% da variação genética, e foram capazes de distinguir os ambientes em avaliação. Por meio do biplot também foi possível selecionar genótipos com ampla adaptabilidade ou adaptabilidade específica à alguns ambientes. Analisando a regressão latente, com base nos genótipos mais produtivos, os híbridos H16 e H64 se mostraram os mais adaptados, e H17 o mais estável, aos ambientes avaliados. No contexto da seleção genômica, após obtidas as matrizes de relacionamento genético, os genitores foram distintos em dois grupos, o primeiro agrupando os genitores masculinos (linhagens R) e o segundo os femininos (linhagens A). Os modelos de predição apresentaram boa acurácia preditiva, exceto os esquemas de validação cruzada para híbridos avaliados em mais de cinco locais (CV3.1 e CV3.2), devido a menor informatividade da população de treinamento. Também observou-se que a adição da matriz de efeitos não-aditivos, proporciona um relativo aumento na acurácia de predição de híbridos em múltiplos ambientes, quando comparado aos modelos aditivos. Já para a predição de híbridos ainda não testados, também foram observadas altas acurácias, no entanto o menor tamanho da população de treinamento, e predições baseadas em informação de

machos, levou a uma redução da acurácia em todos os esquemas testados. Dessa forma, os métodos de modelos mistos podem ser usados com eficiência nos programas de melhoramento de sorgo biomassa, possibilitando selecionar genótipos mais produtivos, e adaptados e/ou estáveis a uma ampla gama de ambientes, ou à ambientes específicos, além de predizer híbridos ainda não avaliados, ou avaliados em múltiplos ambientes, apresentando alta acurácia.

Palavras-chave: *Sorghum bicolor*. Modelos mistos. Melhoramento de plantas. Validação Cruzada.

ABSTRACT

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, September, 2019. **Study of interaction genotype × environment and genomic prediction of biomass sorghum hybrids**. Adviser: José Eustáquio de Souza Carneiro. Co-advisers: Rafael Augusto da Costa Parrella and Pedro Crescêncio Souza Carneiro.

Biomass sorghum is a crop that is gaining high demand in the bioenergy market due to its agronomical characteristics, especially tolerance to abiotic stresses and high biomass yield, aiming to bioenergy production. In this context, this work sought to: i) Investigate the genotype × environment (G×E) interaction sources and select productive, adapted and/or stable genotypes, in different environmental conditions using Factor Analytic models; ii) To evaluate the hybrid prediction accuracy using SNP markers information and multi-environment data; iii) To compare the accuracy of additive and additive-dominant models; and iv) Predict hybrids untested in the program. In our analyses we used the historical database of the Bioenergy Sorghum Breeding Program of Embrapa Milho e Sorgo. We genotyped by genotyping-by-sequencing (GBS) 17 R-lines (male fertile) and 46 A-lines (male sterile). The 202 hybrids generated from these lines were evaluated in various regions of the Brazilian territory. Due to the high unbalanced data from trials of different years, and between the preliminary (APH) and final (VCU) trials for hybrid evaluation, we applied a mixed model method for data analysis using the ASReml package available in the software R. High G × E interaction between the evaluated environments was determined, which showed genotypic correlations varying from -0.09 to 0.68. We were able to distinguish the evaluated environments using a PCA analysis, in which the first two factors explained 48.81% of the genetic variation. Through biplot analysis it was possible to select genotypes showing high adaptability and/or specific adaptability to some environments. The results of the latent regression, based on the most productive genotypes, showed that hybrids H16 and H64 were the most adapted, and hybrid H17 the most stable in the evaluated environments. The genetic relationship matrix from the genomic selection analysis separated the hybrid parents in two groups, the first including the male parents (R-lines) and the second the female parents (A-lines). The predict models presented high prediction accuracy, except the validation schemes for hybrids evaluated in more than five locations (CV3.1 and CV3.2), due to the low informativity of the training population. Also, it was observed that the addition of the non-additive effect matrix provided a relative increase of prediction accuracy to hybrids in multi-environments, when compared to the additive models. Prediction accuracy of unevaluated hybrids were high however the small training population size, and predictions based on male information, led to a reduced prediction accuracy in all

tested scenarios. Thus, our results conclude that mixed model methods can be used with efficiency in biomass sorghum breeding programs, making possible to select more productive genotypes, as well as, more adaptable and/or stable genotypes to specific or to diverse environments. The tested models also allowed high prediction accuracy of unevaluated hybrids or specific environments.

Keywords: *Sorghum bicolor*. Mixed models. Plant breeding. Cross-validation.

SUMÁRIO

INTRODUÇÃO GERAL

O sorgo biomassa [*Sorghum bicolor* (L.) Moench] vem ganhando mercado nos últimos anos, devido principalmente ao seu potencial na produção de energia térmica, biogás e etanol de segunda geração, e às crescentes necessidades de fontes renováveis de energia no Brasil e no mundo (Boyle, 2004; Pao & Fu, 2013; da Silva, 2016). Com isso o programa de melhoramento de sorgo bioenergia da Embrapa Milho e Sorgo, investe constantemente em tecnologias e pesquisas inovadoras que visam a obtenção de genótipos que supram a necessidade do mercado, e que sejam altamente produtivos, com boa qualidade de biomassa, além de ampla adaptabilidade e estabilidade em diferentes regiões/ambientes.

Entre os principais objetivos dos programas de melhoramento, se encontra a obtenção de genótipos superiores, quando comparados aos presentes no mercado, e que sejam possíveis de plantio nas diversas regiões do país (Oliveira et al., 2019). Para isso inúmeros testes de campo são realizados, em diferentes anos em âmbito nacional, os quais são chamados ensaios de valor de cultivo e uso (VCU). Contudo, devido a presença de desbalanceamento entre os ensaios de VCU realizados em diferentes anos, por inclusão de novos tratamentos e exclusão de genótipos inferiores, o uso de modelos mistos para ajuste dos dados é essencial na obtenção de resultados acurados e que forneçam o máximo de informações necessárias para a seleção dos melhores genótipos (de Resende & Thompson, 2004). Para isso, estudos adotando a análise de fatores (FA) estão sendo realizados com alta eficiência no melhoramento de plantas (Sousa et al., 2015), pois possibilitam a captura e o entendimento das principais causas da interação genótipo por ambiente (G × A), e seleção de genótipos mais produtivos, adaptados e estáveis, a uma ampla gama de ambientes, ou a ambientes específicos (Kelly et al., 2007; Cullis et al. 2010; Dias et al., 2018).

Os programas de melhoramento de sorgo, devido a existência da macho-esterilidade genético-citoplasmática, vem focando seus trabalhos na síntese de híbridos. Estes genótipos, de maneira geral, apresentam maiores produtividades e maior adaptação a diferentes ambientes, quando comparados às variedades comerciais. Porém, um dos principais problemas encontrado pelos melhoristas, quando se trabalha com híbridos, é o grande número de possíveis cruzamentos para a síntese de novos híbridos. A determinação de quais destes cruzamentos resultarão em híbridos geneticamente superiores e a viabilidade da avaliação de todos os possíveis cruzamentos são questões importantes de decisão dentro de programas de melhoramento. Além disso, em uma situação ideal, a avaliação de híbridos deveria ser feita em todas as condições edafoclimáticas disponíveis para o plantio da cultura.

Dessa forma, devido ao grande número de possíveis cruzamentos, dado o tamanho do germoplasma de genitores disponível nos programas, se torna impraticável avaliar todas as combinações, o que necessitaria de grande número de mão de obra e de experimentos a serem implementados. Portanto este é um processo economicamente oneroso dada a necessidade de avaliação em campo de todas estas combinações sintetizadas.

Para auxiliar na solução destas questões Hayes et al. (2001) propôs a seleção genômica (GS, do inglês, *genomic selection*). Com o intuito de predizer genótipos ainda não avaliados, com base em banco de dados fenotípicos e genotípicos, a seleção genômica abriu novas perspectivas acerca do melhoramento genético. Diversos estudos vem sendo realizados, o que vem promovendo o aumento da eficiência e poder de acurácia de predição deste método (Heffner et al., 2010; Jarquin et al., 2014; Lopez-Cruz et al., 2015; Crain et al., 2018).

A seleção genômica permite a predição de genótipos não avaliados baseando-se em informações de campo e genotipagem de indivíduos aparentados, o que possibilita a seleções de possíveis cruzamentos superiores sem que haja a necessidade de uma pré-avaliação em

campo. Com isso este método pode guiar os melhoristas para realização de cruzamentos mais promissores e com maior potencial, reduzindo o número de cruzamentos a serem realizados e conduzidos em avaliações de campo. Além disso, a seleção genômica também permite a predição de genótipos em ambientes ainda não avaliados, dada a presença de informações de ensaios anteriormente realizados nestes locais.

Dessa forma, o objetivo deste trabalho foi realizar o estudo da interação G × A, e selecionar genótipos superiores, adaptados e estáveis, e interpretar a influência das diferentes condições ambientais na magnitude da interação. Além de avaliar a acurácia de métodos de predição de híbridos não avaliados e métodos para a predição de genótipos em multi-ambientes para o programa de melhoramento de sorgo biomassa.

REFERÊNCIAS BIBLIOGRÁFICAS

Boyle G (2004) Renewable Energy: Power for a Sustainable Future. 2nd edition. Oxford University Press, Oxford, pp. 456.

Crain J, Mondal S, Rutkoski J, et al (2018) Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. The plant genome 11.

Cullis BR, Smith AB, Beeck CP, et al (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome 53:1002–1016.

Dias KODG, Gezan SA, Guimarães CT, et al (2018) Estimating genotype × environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. Crop Science 58:72–83.

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

Heffner EL, Lorenz AJ, Jannink JL, et al (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop science 50:1681–1690.

Jarquín D, Crossa J, Lacaze X, et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics, 127:595–607.

Lopez-Cruz M, Crossa J, Bonnett D, et al (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3: Genes, Genomes, Genetics 5:569–582.

Oliveira ICM, Damasceno CB, Pastina MM, et al (2019) Desempenho produtivo de híbridos de sorgo biomassa do programa de melhoramento genético da Embrapa. Embrapa Milho e Sorgo-Boletim de Pesquisa e Desenvolvimento (INFOTECA-E).

Pao HT, Fu HC (2013) Renewable energy, non-renewable energy and economic growth in Brazil. Renewable and Sustainable Energy Reviews 25:381–392.

Kelly AM, Smith AB, Eccleston JA, et al (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Science 47:1063–1070.

de Resende MDV, Thompson R (2004) Factor analytic multiplicative mixed models in the analysis of multiple experiments. Revista de Matemática e Estatística 22:31–52.

da Silva RC, de Marchi Neto I, Seifert SS (2016) Electricity supply security and the future role of renewable energy sources in Brazil. Renewable and Sustainable Energy Reviews 59: 328–341.

Sousa LB, Hamawaki OT, Nogueira APO, et al (2015) Evaluation of soybean lines and environmental stratification using the AMMI, GGE biplot, and factor analysis methods. Genet. Mol. Res 14:12660–12674.

CAPÍTULO 1

GENOTYPE-BY-ENVIRONMENT INTERACTION AND YIELD STABILITY ANALYSIS OF BIOMASS SORGHUM HYBRIDS USING FACTOR ANALYTIC MODELS

RESUMO

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, setembro de 2019. **Interação genótipo por ambiente e análise de estabilidade de produção para híbridos de sorgo biomassa usando modelos de fator analítico**. Orientador: José Eustáquio de Souza Carneiro. Coorientadores: Rafael Augusto da Costa Parrella e Pedro Crescêncio Souza Carneiro.

O sorgo biomassa surgiu como uma cultura alternativa para a produção de biocombustíveis e bioeletricidade. A produção de massa verde (PMV) é uma variável quantitativa altamente correlacionada ao poder calorífico das cultivares de sorgo, mas também altamente influenciada pelo ambiente. O objetivo principal desse trabalho foi investigar a interação genótipo por ambiente (G × A) e a estabilidade de 63 híbridos de sorgo avaliados para PMV em 10 locais (ambientes) dos ensaios finais do programa de melhoramento da Embrapa, usando modelos mistos do tipo fator analítico (FA). As correlações entre pares de ambientes variaram entre -0,09 e 0,68, indicando a existência de interação G × A. A análise do *biplot* permitiu a identificação de híbridos exibindo uma ampla adaptabilidade a um conjunto de ambientes, bem como híbridos adaptados a ambientes específicos. Por exemplo os híbridos H22, H38 e H43, foram os mais adaptados a Goiânia, H29 e H30 a Lavras, H21 a Sete Lagoas, Sinop e Planaltina, H5 e H23 a Sinop e Planaltina, H52 a Santa Vitória e H12, H28, H45, H46, H57 e H69 a Campos dos Goytacazes e Pelotas. Análises de regressão latente permitiram a identificação de híbridos estáveis e altamente produtivos, destacando o híbrido H17 como o mais estável entre os ambientes avaliados. Estas informações podem ajudar os melhoristas a selecionar materiais adaptados a ambientes específicos ou estáveis a um conjunto de ambientes.

Palavras-chave: *Sorghum bicolor* (L.) Moench; melhoramento de plantas; bioenergia; ensaios em múltiplos ambientes; modelos mistos.

ABSTRACT

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, setembro de 2019. **Genotype-by-Environment Interaction and Yield Stability Analysis of Biomass Sorghum Hybrids Using Factor Analytic Models**. Adviser: José Eustáquio de Souza Carneiro. Co-advisers: Rafael Augusto da Costa Parrella and Pedro Crescêncio Souza Carneiro.

Biomass sorghum has emerged as an alternative crop for biofuel and bioelectricity production. Fresh biomass yield (FBY) is a quantitative trait highly correlated to the calorific power of energy sorghum cultivars, but also highly affected by the environment. The main goal of this study was to investigate the genotype-by-environments interaction (G × E) and the stability of 63 sorghum hybrids evaluated for FBY across 10 locations (environments) of the late-stage Embrapa's breeding trials, using factor analytic (FA) mixed models. Pairwise correlations of environments ranged from -0.09 to 0.68, indicating the existence of G × E. Biplot analysis allowed for the identification of hybrids exhibiting broad adaptability to a set of environments, as well as hybrids adapted to specific environments. For example, the hybrids H22, H38 and H43, were the most adapted to Goiânia, H29 and H30 to Lavras, H21 to Sete Lagoas, Sinop and Planaltina, H5 and H23 to Sinop and Planaltina, H52 to Santa Vitória and H12, H28, H45, H46, H57 and H69 to Campos dos Goytacazes and Pelotas. Latent regression analysis allowed the identification of stable and highly productive hybrids, highlighting the hybrid H17 as the most stable across the evaluated environments. This information can help breeders to select materials that are adapted to specific environments or stable to a set of environments.

Keywords: *Sorghum bicolor* (L.) Moench; plant breeding; bioenergy; multi-environment-trials; mixed models.

Introduction

Biomass sorghum [*Sorghum bicolor* (L.) Moench] can be used as an alternative feedstock for bioelectricity (Demirbas et al., 2009) and second-generation ethanol production (Reis et al., 2016). For these purposes, one of the most important traits is dry biomass yield (DBY) (de Oliveira et al., 2018), which is highly correlated to the calorific power of sorghum cultivars and, consequently, to the potential to generate energy and heat during the burning process. Because of the large number of required processes for measuring DBY, fresh biomass yield (FBY) can be used to perform indirect selections, due to the high correlation between these traits (Almeida et al., 2019). Besides the high FBY, other attributes also highlight the potential of biomass sorghum as an energy crop in Brazil, such as drought and heat tolerance, water use efficiency, short crop growth cycle (160 to 180 days) (Brenton et al., 2016; Parrella et al., 2011), and high adaptability to tropical and subtropical conditions.

In this context, Embrapa Maize and Sorghum has been focusing in the development of high biomass sorghum cultivars. Annually, v*alue for cultivation and use* (VCU) trials are conducted across different locations from distinct Brazilian regions to evaluate hybrids that showed promising performance in preliminary and/or late-stage breeding trials of previous years. Different sets of hybrids are often tested across years of VCU trials. Thus, commercial cultivars and/or parental lines are included as common checks for the selection of promising hybrids across environments. In breeding programs, the set of connected trials evaluated across multiple years are called multi-environment trials (MET). The joint analysis of MET data can provide information about genotype-by-environment interaction (G × E), yield stability and adaptability of hybrids across distinct environments. This information is very important to release new cultivars showing yield stability across a set of environments, or specifically adapted to a given environment (Bornhofen et al., 2018; Burgueño et al., 2008; Dias et al., 2017).

FBY is a quantitative trait highly affected by the environment. Therefore, the lack of information about G × E can lead to a reduction in the genetic gains across years of a breeding program, reinforcing the usefulness of MET studies (Quintero et al., 2018). Although VCU trials are often balanced between locations within a harvest year, they are highly unbalanced between years, since low-performance hybrids are usually replaced by newly developed elite materials along the years of a breeding program. Another reason of imbalance in MET data sets is the occurrence of missing plots, due to biotic or abiotic stresses during the crop growth cycle. For these reasons, joint analysis of unbalanced experiments cannot be performed using common statistical methods widely used to study G × E, adaptability and stability of hybrids across environments. In MET datasets, these studies require the use of more flexible approaches, such as mixed linear models (de Resende and Thompson, 2004; Kelly et al., 2007; Smith et al., 2005).

Using mixed linear models, different variance and covariance (VCOV) structures can be assumed for genetic and residual effects across environments, allowing for modeling genetic and residual correlations, but also heterogeneity of variances across environments. The most complex VCOV structure is assumed by unstructured (UN) models, which consider heterogeneity of variances across environments and specific covariances for pairs of environments. However, due to the high number of estimated parameters when many environments are analyzed, fitting UN models are often a complex task (Smith et al., 2005), which may lead to high standard errors for the variance components and the predicted genetic effects or frequently to non-convergence of models. In this context, multiplicative factor analytic (FA) structures have been proposed as a more parsimonious approach compared to UN models, allowing the estimation of a smaller number of parameters (Kelly et al., 2007; Piepho, 1998, 1997; Smith et al., 2001b). Additionally, graphical tools such as biplots (Kempton, 1984), latent regression plots (Smith et al., 2015; Thompson et al., 2003) and heatmaps of estimated genetic correlation matrices across environments (Cullis et al., 2014; Smith et al., 2015) can be

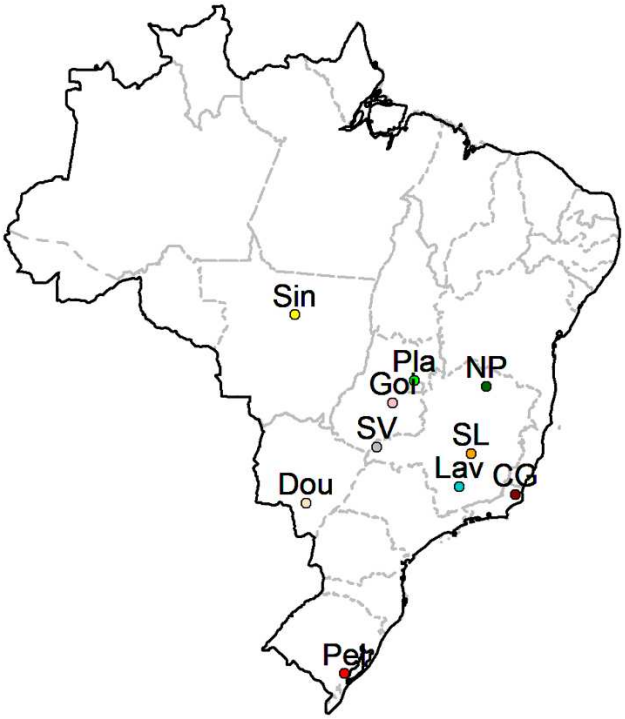
easily implemented to infer about G × E, adaptability and stability of hybrids when using FA models.

The main goals of this study was: i) to infer about G × E, adaptability and stability of Embrapa's biomass sorghum hybrids across environments; ii) to select hybrids stable across a set of environments or specifically adapted to a given environment; using a MET dataset collected over different years and locations of Embrapa Maize and Sorghum breeding program.

Material and Methods

2.1. Genetic Material and Experimental Design

A total of 70 biomass sorghum genotypes were evaluated: 63 experimental hybrids and 7 checks, being three commercial biomass sorghum hybrids (including the cultivar BRS716 from Embrapa), one commercial forage sorghum hybrid (the cultivar BRS655 from Embrapa), and three parental restorer lines. These genotypes were evaluated across 42 VCU trials, carried out over 10 different locations from seven Brazilian states, geographically distributed according to Figure 1. This MET dataset belongs to the historical series of experiments performed by the Embrapa's biomass sorghum breeding program. Trials were named according to the initials of each location followed by the last two digits of year. However, in 2013, two distinct experiments were conducted: the first one during the Brazilian sorghum season, and the second after harvesting the first season crop, being coded by the letter S after the two digits of year (13S).



Years \ Loc.	CG	Dou	Goi	Lav	NP	Pel	Pla	SV	SL	Sin
12	●		●	●	●		●		●	●
13	●	●	●	●	●	●	●		●	
13S	●	●	●	●	●	●			●	●
14		●	●	●	●	●			●	●
15					●			●	●	
16			●	●	●			●	●	●
17					●	●			●	

Figure 1: Geographic coordinates of the locations in which the VCU trials were conducted. Years 2012 to 2017 were coded as 12 to 17. CG: Campos dos Goytacazes-RJ; Dou: Dourados-MS; Goi: Goiânia-GO; Lav: Lavras-MG; NP: Nova Porteirinha-MG; Pel: Pelotas-RS; Pla: Planaltina-DF; SV: Santa Vitória-MG; SL: Sete Lagoas-MG; Sin: Sinop-MT. The code 13S represents the VCU trial conducted during the second season of 2013.

Each trial was arranged as a triple lattice design. Plots consisted of two 5-meters rows spaced by 0.7 meters, considering an initial population density of 140,000 plants ha^{-1}. The number of genotypes in each VCU trial ranged from 13 (2013) to 33 (2014), with an average of 25 treatments per trial (Table 1). Three to six checks were added to each trial, such as the biomass and forage commercial hybrids BRS716 and BRS655, respectively. However, BRS716 and BRS655 were included as common checks across all VCU trials. Although the MET dataset used in this study is highly unbalanced across years, it is balanced across locations within the same harvest year. Table 1 shows the number of genotypes (diagonal) within each of the seven years of VCU trials, and the number of common genotypes between years (above the diagonal).

Table 1: Number of genotypes within each year, and the number of common genotypes between years (above the diagonal) of the VCU trials.

Year	2012	2013	2013(S)°	2014	2015	2016	2017
2012	25(3*)	5	25	8	5	6	6
2013		16(6)	5	12	7	6	6
2013(S)			25(3)	8	5	6	6
2014				36(7)	17	16	14
2015					25(8)	12	11
2016						25(8)	21
2017							25(7)

* number of locations evaluated in each year; (S)° VCU trial conducted after harvesting the first season crop of 2013.

Some of the 63 hybrids were evaluated in all 42 VCU trials, while others were evaluated only across locations within a single year. Hybrids were obtained from crossing six R-lines (restorer) to 32 A-lines (cytoplasmic male-sterile), belonging to the Embrapa's biomass sorghum breeding program. Not all possible hybrid combinations were obtained.

Fertilizer management, weed and pest control, and other agricultural practices were performed as recommended for sorghum cultivation (Borém et al., 2014). In all VCU trials, FBY (tons/ha) was evaluated at grain physiological maturity. All plants of the plot were weighted using a digital suspension scale, and subsequently converted into tons per hectare.

2.2. Phenotypic Data Analyses

2.2.1 Individual Analysis per Trial

Mixed linear models were fitted using the statistical package ASReml-R v.3 (Butler et al., 2009), available for the R software (Team and others, 2015), which estimates the variance components using the restricted maximum likelihood (REML) method through the Average Information algorithm (AI) (Gilmour et al., 1995). Phenotypic data quality, experimental accuracy, coefficient of variation and generalized heritability were estimated by individual analysis of each trial, using the following mixed linear model:

$$\boldsymbol{y = \mu 1_n + Xr + Z_1 b.r + Z_2 t + e}$$

where y is the vector ($n \times 1$) of phenotypes values for i treatments (genotypes), within k blocks of j repetitions, n is the total number of plots in each trial; μ is the general mean; r is the vector ($j \times 1$) of fixed effects of replicates; t is the vector ($i \times 1$) of the random effects of genotypes, with $\boldsymbol{t \sim N(0, I_i \sigma_t^2)}$, in which σ_t^2 is the genetic variance; b is the vector ($jk \times 1$) of random block effects within replicates, with $\boldsymbol{b \sim N(0, I_{jk} \sigma_b^2)}$, in which σ_b^2; is the variance of blocks; and e is the vector of residual effects, with $\boldsymbol{e \sim N(0, I_n \sigma_e^2)}$, in which σ_e^2 is the residual variance. $\boldsymbol{X}$, $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represent the incidence matrices for their respective effects, replicates, genotypes and blocks within replicates, with dimensions $n \times j$, $n \times i$ and $n \times jk$, respectively. 1_n is a vector ($n \times 1$) of ones. And $\boldsymbol{I_i}$, $\boldsymbol{I_{jk}}$ and $\boldsymbol{I_n}$ are identity matrices with their corresponding orders.

Diagnostic plots were used to verify the presence of outliers and the residuals distribution of the fitted models for each trial. Generalized heritability (H^2) (Cullis et al., 2006), and experimental accuracy (Ac) (Mrode, 2014), were estimated for each trial using the following equations:

$H^2 = 1 - [PEV/(2 \times \sigma_g^2)]$ and $Ac = \sqrt{1 - (PEV/\sigma_g^2)}$; in which PEV is the prediction error variance, corresponding to the average variance of the difference between two predicted genetic effects, and σ_g^2 is the genetic variance. The coefficient of variation was calculated using the formula $CV\% = (\sigma_e/\bar{\mu}) \times 100$, in which is σ_e the residual standard deviation, $\bar{\mu}$ is the general mean of each trial.

2.2.2. Joint Analysis of Multiple Trials

The combined analysis of all trials was carried out in two stages. First, in order to obtain the adjusted means of genotypes and the residuals by location, the following model was used:

$$y = \boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{X_1 a} + \boldsymbol{X_2 r.a} + \boldsymbol{X_3 t} + \boldsymbol{Z_1 b.r.a} + \boldsymbol{Z_2 t.a} + \boldsymbol{e}$$

where y is the vector ($n \times 1$) of phenotypic values of i treatments (genotypes), within k blocks and j replicates in m years, in which $\boldsymbol{n = ijkm}$; μ is the general mean; a is the vector ($m \times 1$) of fixed effects of years; $r.a$ is the vector ($jm \times 1$) of fixed effects of replicates within years; t is the vector ($i \times 1$) of fixed effects of genotypes; $t.a$ is the vector ($im \times 1$) of genotype-by-year random interaction effects; $b.r.a$ is the vector ($kjm \times 1$) of random block effects within replicates within years; and e is the vector of residual effects, with $\boldsymbol{e \sim N(0, \sigma^2)}$. $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ represent the incidence matrices for their respective effects, with dimensions $im \times m$, $im \times jm$, $im \times i$, $im \times kjm$ and $im \times im$, respectively. $\mathbf{1}_{\boldsymbol{n}}$ is a vector ($n \times 1$) of ones. At this stage, the adjusted means of genotypes were corrected for the experimental design and year-related effects.

In the second stage, models were fitted across locations, using the adjusted means of genotypes and the residual variance-covariance (VCOV) matrix estimated from the residual effects of the first stage:

$$\boldsymbol{y_1} = \boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{Xs} + \boldsymbol{Zt.s} + \boldsymbol{e}$$

where y is the vector ($il \times 1$) of adjusted means of i treatments in each location (environments); μ is the general means; s is the vector ($l \times 1$) of fixed effects of environments; $\boldsymbol{t.s}$ is the vector ($il \times 1$) of random genetic effects of treatments (genotypes) within environments, with $\boldsymbol{t.s \sim NM(0, G \otimes I_l)}$; and $\boldsymbol{e}$ is the vector ($il \times 1$) of residual effects, with $\boldsymbol{e \sim NM(0, \Sigma)}$, in which $\boldsymbol{\Sigma}$ corresponds to the VCOV matrix associated to $\boldsymbol{e}$, assumed to be known from the first stage. $\boldsymbol{\Sigma}$ is a diagonal matrix, in which the diagonal elements are given by the inverse of the VCOV matrix of the adjusted means of genotypes in each environment (Smith et al., 2001a). $\boldsymbol{G}$ is the genetic VCOV matrix for the effect of genotypes within environments, with dimension $l \times l$. A factor analytic structure (FA) of order k, in which k is the number of multiplicative components,

was considered to model the $\boldsymbol{G}$ matrix. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the incidence matrices for their respective effects, with dimension $il \times l$ and $il \times il$, respectively; 1_n is a vector $(n \times 1)$ of ones; and $\boldsymbol{I_{tja_i}}$, $\boldsymbol{I_m}$ e $\boldsymbol{I_{n_i}}$ identity matrices with their corresponding orders.

The overall percentage ($\bar{v}$) of the genetic variance explained by the factors k was obtained by:

$$\bar{v} = \mathbf{100}\, tr\,(\hat{\Lambda}\hat{\Lambda}^T)/\, tr\,(\hat{\Lambda}\hat{\Lambda}^T + \hat{\psi})$$

where $\boldsymbol{\Lambda}$ is the matrix $(l \times k)$ of factor loadings, $\{\boldsymbol{\lambda_{lk}}\}$, in which $\boldsymbol{\lambda_{lk}}$ is the k^{th} factor loading $(k = 1,2, \dots, K)$ for the environment 1; $\boldsymbol{\psi}$ it is a diagonal matrix (l × l) with the specific variances for each environment; and *tr* is the trace of the matrix $(\hat{\Lambda}\hat{\Lambda}^T + \hat{\psi})$. FA models of different orders can be compared based on the overall percentage of the genetic variance explained by the factors in the model, or on the values of AIC (Akaike Information Criterion) (Bozdogan, 1987) and BIC (Bayesian Information Criterion)(Schwarz and others, 1978). In this study, models from the first (FA_1) to fifth orders (FA_5) were compared. The best FA model was selected based on the AIC values.

The VCOV matrix for the effect of genotypes within environments, defined by the FA_k model is:

$$cov(\alpha) = (\Lambda\Lambda^T + \psi) \otimes \boldsymbol{I_l}$$

where $\boldsymbol{\Lambda}$ is the matrix $l \times k$ of factor loadings $\{\lambda_{lk}\}$, in which λ_{lk} is the k^{th} factor loading of the $(k = 1,2, \dots, K)$ for the environment l; ψ is a diagonal matrix $(l \times l)$ with specific variances for each environments; I_l is an identity $(l \times l)$ matrix. The genetic correlations between pairs of environments (ρ_{ij}) were estimated via the FA model in the combined analysis of the

environments. Thus, $\rho_{ij} = COV_{ij} / (\sqrt{\sigma_{li}^2 \sigma_{lj}^2})$, where COV_{ij} is the genetic covariance among trials i and j; and σ_{li}^2 and σ_{lj}^2 are the genetic variances for the trials $\boldsymbol{i}$ and $\boldsymbol{j}$, respectively.

After estimating the variance components and solving the equation of mixed models, the factor scores for genotypes ($\tilde{f}$) and the factor loadings for environments ($\tilde{\delta}$) were obtained as described by Resende and Thompson (2004).

2.2.3. Yield Stability Analysis Across Environments

Latent regression plots were built for 10 genotypes expressing the best yield performance in the joint analysis of environments. This approach can be used to investigate the yield adaptability and stability of genotypes across different environments (Smith and Cullis, 2018). The predicted breeding values reflect the genotype responses to a factor loading of a given environment, and are calculated as the product between the genotypic factor score and the environmental factor loading. According to Cullis et al. (2010), for a meaningful interpretation, environmental factor loadings must be rotated to a principal components solution, maximizing the proportion of the genetic covariance accounted by the first rotated factor loading, while the second rotated factor loading accounts for the next largest proportion and is orthogonal to the first, and so on.

Results

3.1. Individual Analyzes per Trial

Table S1 shows the phenotypic means, estimated genetic, block and residual variances, generalized heritability, experimental accuracy, and coefficients of variation for each trial. The average FBY across trials ranged from 27.66 (GC.15) to 102.97 t ha^{-1} (Dou.15), with an overall

mean of 68.72 t ha^{-1}. Genetic variances ranged from 11.71 (CG.15) to 703.74 (SV.13S) across trials, and differed significantly from zero, based on the Likelihood ratio test (considering α = 0.05). Block variances ranged from 0.00 (SL.12, NP13.S, SV.13S, Goi.13, SV.13, Sin.13, Dou.14, NP.14, Goi.15, Dou.16, Lav.16, Pel.16, Pla.16, Goi.17, Lav.17 and SL.17) to 118.66 (NP.16). Heterogeneity of residual variances (σ_e^2) was observed, with values ranging from 14.02 (GC.15) to 217.43 (Sin.13) across trials. Generalized heritability values ranged from 0.54 (Pel.15) to 0.97 (SL.13S), while experimental accuracy values ranged from low (0.29, Pel.15) to very high (0.97, SL.13S) according to the classification proposed by Resende and Duarte (2007). By contrast, the coefficients of variation (CV%) ranged from 7.45 (Pla.17) to 19.31% (Goi.16), showing high experimental precision. Similar values of CV% were found by Parrella et al. (2016) and Lombardi et al. (Lombardi et al., 2018) for FBY in sweet sorghum, and de Almeida et al. (Almeida et al., 2019) in biomass sorghum.

3.2. Joint Analysis of Multiple Trials

Values of the AIC, REML log-likelihood of the fitted model (logREML), and the number of parameters (NP) across the VCOV models examined for the G matrix are presented in Table 2. The lower the AIC value, the better the model fits. Thus, the models $FA_{(2)}$ and $FA_{(3)}$ presented the best fit to the phenotypic dataset evaluated in this study, due to the lower AIC values compared to the other models. However, due to the lower number of parameters (NP), the $FA_{(2)}$ was selected.

Table 2: Total number of parameters (NP), AIC values and logREML of the VCOC models examined for the G matrix in the combined analysis of environments.

G	NP	AIC	logREML
$FA_{(1)}$	20	23850,44	-11905,22
$FA_{(2)}$	**29**	**23843,16**	**-11892,58**
$FA_{(3)}$	37	23842,10	-11884,05
$FA_{(4)}$	44	23847,34	-11879,67
$FA_{(5)}$	50	23844,89	-11872,44

FA_k: factor analytic model, where k is the order of the model.

Genetic correlations between environment (Figure 2) ranged from -0.09 (Goiânia and Pelotas) to 0.68 (Sete Lagoas and Planaltina), indicating the existence of low to high G × E across environments. Goiânia was the only environment that presented correlations close to zero or negative with other locations, such as Campos dos Goytacazes (-0.06) and Pelotas (-0.09). More than 90% of the genetic correlations between environments were positive. In this study, genetic correlations above 60% were considered high, indicating the occurrence of low G × E between environments, i.e. the genotypes exhibited similar FBY between environments. By contrast, pairs of environments showing low correlation, suggest the occurrence of high G × E, i.e. the performance of genotypes changed across environment. Additionally, the genetic correlation between environments can be used as an index for the definition of mega-environments.

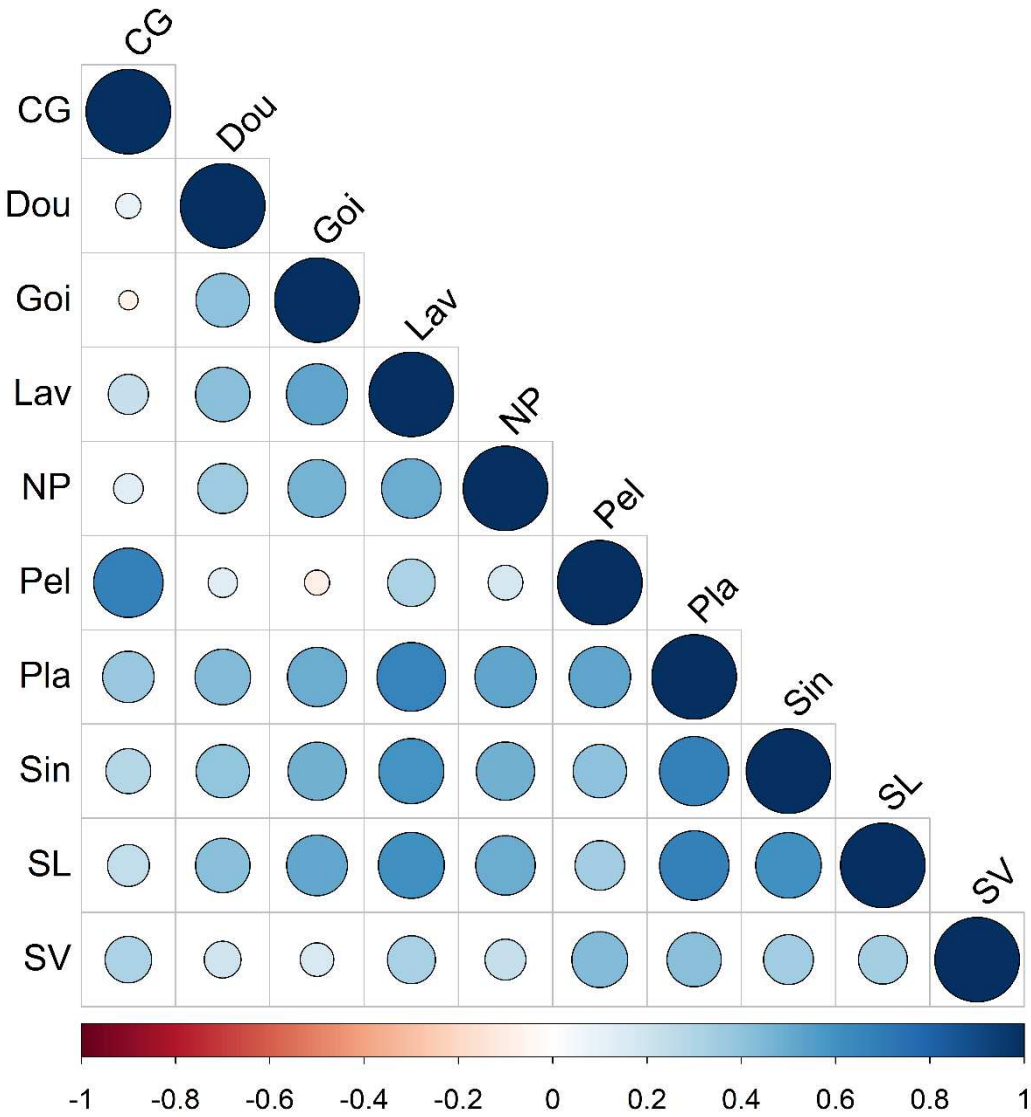


Figure 2: Pairwise genetic correlations for FBY across 10 environments: Campos dos Goytacazes-RJ (CG); Dourados-MS (Dou); Goiânia-GO (Go); Lavras-MG (Lav); Nova Porteirinha-MG (NP); Planaltina-DF (Pla); Santa Vitória-MG (SV); Sete Lagoas-MG (SL); Pelotas-RS (Pel); Sinop-MT (Sin). The size and color of the circles are related to the magnitude and to the sign and magnitude of the genetic correlations between environments, respectively.

The two factors of the $FA_{(2)}$ model jointly explained 48.81% of the observed genetic variance. The first factor captured 38.91% of the genetic variability, and represented mostly the main effect of genotypes, and the environments of CG (FA1 = 9.76, FA2 = 1.02), Pel (FA1 = 8.80, FA2 = 0.79) and SV (FA1 = 8.44, FA2 = 5.92), since these environments had the highest environmental loadings for this factor. The second factor explained 9.90% of the genetic variability, and presented greater representativeness for Dou (FA1 = 0.78, FA2 = 5.57), Goi (FA1 = -1.42, FA2 = 6.78)), Lav (FA1= 2.66, FA2= 7.53), NP (FA1= 1.65, FA2= 8.76), Pla

(FA1= 4.38, FA2= 6.89), Sin (FA1= 5.19, FA2= 10.39) and SL (FA1 = 3.64, FA2 = 9.25), since these environments had the highest environmental loadings for this factor. The genotypic scores ranged from -14.76 (H09) and 22.49 (H19), and from -45.80 (H54) to 28.19 (H27 - data not shown) for the first and second factors, respectively.

Observing the biplot between the two factors of $FA_{(2)}$ model (Figure 3), it is possible to infer about the performances of the genotypes across environments, as well as the genetic correlation between environments. For example, the angles between the vectors of Goi and CG, and between Goi and Pel, are higher than 90°, indicating the existence of negative to zero genetic correlations between Goi and these two environments. Thus, based on this information, it is possible to infer that the performances of genotypes in Goi are distinct from the ones observed in CG e Pel, indicating the occurrence of high G × E between Goi and CG, and between Goi and Pel. Additionally, it is also possible to observe a moderate genetic correlation between SL and Pla (0.68), Pel and CG (0.68), and between Pla and Lav (0.66), showing the existence of a low G × E for these pairs of locations, i.e. the performance of genotypes were more similar between highly correlated environments.

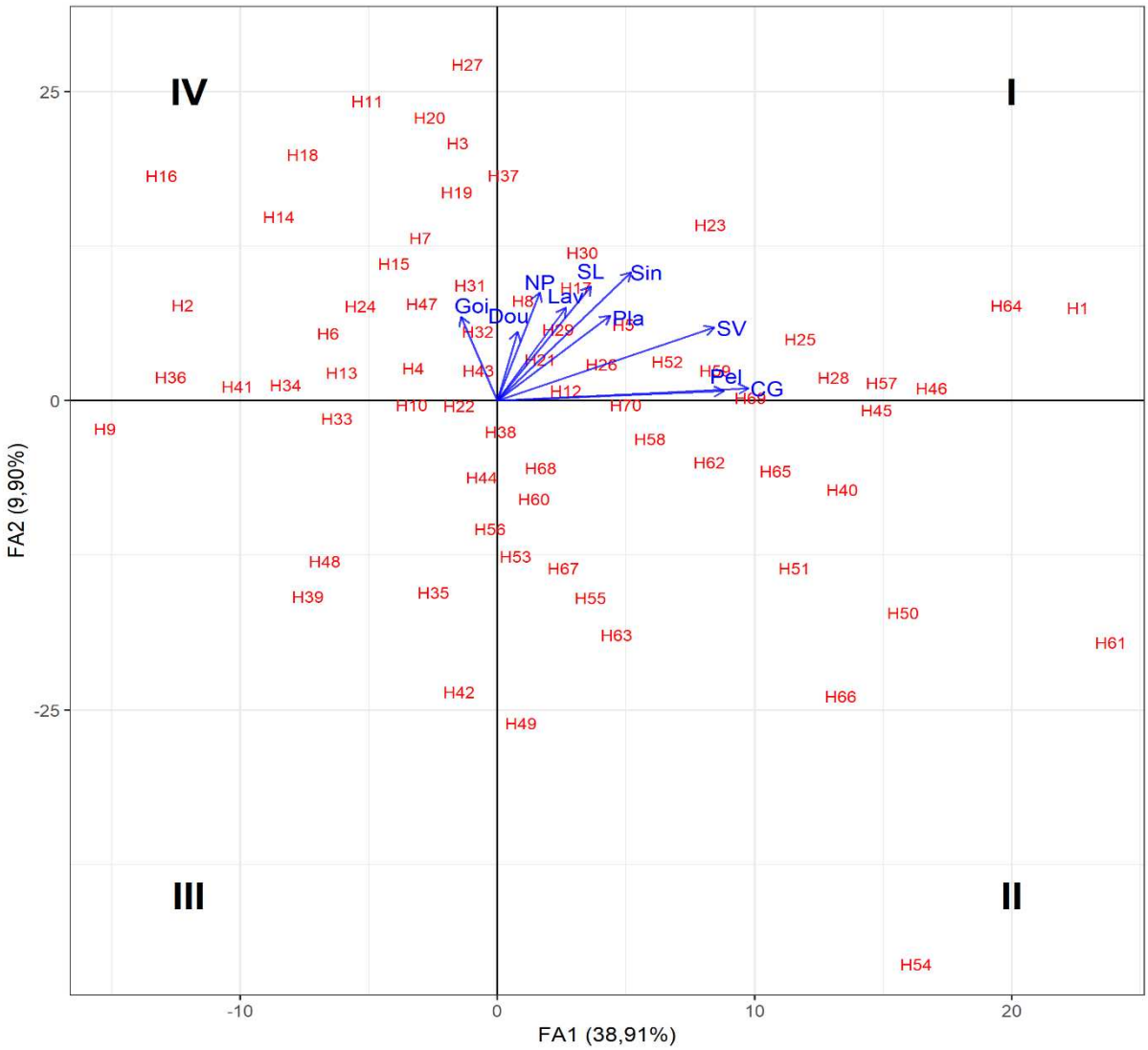


Figure 3: Biplot of the scores of the 70 genotypes and the loadings of the 10 environments for the first (FA1) and second factors (FA2) of $FA_{(2)}$ model. Campos dos Goytacazes-RJ: CG; Dourados-MS: Dou; Goiânia-GO: Go; Lavras-MG: Lav; Nova Porteirinha-MG: NP; Planaltina-DF: Pla; Santa Vitória-MG: SV; Sete Lagoas-MG: SL; Pelotas-RS: Pel; Sinop-MT: Sin.

According to Murakami & Cruz (2004), genotypes located in the same quadrant, in which one or few environments were allocated, have specific adaptability to these environments. Thus, the genotypes located in quadrant IV have specific adaptability to the environment Goi. Among them, the treatments H31, H32, and H47 were the most adapted to this specific environment, since they presented scores close to the loadings of Goiânia for both FA1 and FA2. Additionally, as proposed by Cullis et al. (2010), the interpretation of the FA biplots may be similar to the GGE biplot. Thus, genotypes showing scores close to zero are

broadly adapted across environments, i.e. with environments-specific performances close to the general mean, such as hybrids H12 (FA1 = 1.90, FA2 = -0.53), H22 (FA1 = 0.60, FA2 = 0.27), H38 (FA1 = -0.37, FA2 = -1.33) and H43 (FA1 = -0.21, FA2 = 1.42). Genotypes located in the same direction to the environments, i.e. having angular coefficient similar to the environments, tend to be more adapted to the respective environments. Thus, the genotypes H7 (FA1 = -2.49, FA2 = 14.18) and H11 (FA1 = -4.48, FA2 = 23.12) exhibited higher adaptability to Goi; H29 (FA1 = 1.83, FA2= 6.77) and H30 (FA1= 4.14; FA2= 13.07) to Lav; H21 (FA1 = 1.20, FA2 = 22.37) adaptability to SL, Sin and Pla; H5 (FA1 = 1.20, FA2 = 22.37) and H23 (FA1 = 8.83, FA2 = 15.27) to Sin and Pla; H52 (FA1 = 6.56, FA2 = 4.07) to SV and H12 (FA1 = 2.05, FA2 = -0.33), H28 (FA1 = 12.48, FA2 = 3.08), H45 (FA1 = 14.27, FA2 = 0.27), H46 = 17.16, FA2 = 2.07), H57 (FA1 = 14.08, FA2 = 1.64) and H69 (FA1 = 9.29, FA2 = 1.34), with higher adaptability to CG and Pel. By contrast, genotypes located in the opposite direction to the environment, i.e. having an angle close to 180 with the vector of the environment, tend to be less adapted to this respective environment. Thus, in this study H35 can be classified as a less adapted genotype to Dou; H68, H55 and H63 to Goi, H48 to Sin, H39 to SL, H33 and H9 to Pel and CG.

3.3. Yield Stability Analysis Across Environments

Smith et al. (Smith et al., 2015), proposed the use of latent regression plots to study yield stability and the adaptability, of genotypes across environments. In this approach, the predicted breeding values of genotypes are regressed on the environmental factor loadings of the FA model. Thus, in the present study, latent regression plots were built for 10 genotypes, evaluated in at least eight locations, showing the highest overall predicted means by the $FA_{(2)}$ model, such as proposed by Smith and Cullis (2018). For the other genotypes, latent regression plots are not

shown for reasons of brevity. First, the latent regressions plots were built for the first factor (FA1), regressing the predicted breeding values on the rotated environmental loadings of FA1 (Figure 4). Then, for the second factor, the predicted breeding values were regressed on the rotated environmental loadings of FA2 (Figure 5). In each latent regression plot (Figures 4 and 5), the circles and triangles correspond to the predicted breeding values of genotypes in tested and untested locations, respectively. The overall predicted means obtained by the $FA_{(2)}$ model ranged from 41.55 t ha^{-1} (H25) to 84.60 t ha^{-1} (H19), with an overall mean of 67.79 t ha^{-1}. Additionally, the overall predicted means of the 10 highest and the 10 smallest-yielding genotypes, and their respective factor scores from the $FA_{(2)}$ model, are presented in the supplementary tables S2 and S3, respectively.

In the latent regression plots for the first and the second factors (Figures 4 and 5), the slope of the latent regression line correspond to the genotype score for the factor. Thus, genotypes showing high positive slopes are more responsive to environmental improvements, i.e. exhibit higher predicted breeding values in environments having higher factor loadings. For example, among all environments, CG (FA1 = 9.75, black points), Pel (FA1 = 8.80, light-blue points) and SV (FA1 = 8.44, dark-green points) presented higher loadings for the first factors. Thus, the hybrids H64 and H66 can be pointed as the best adapted genotypes to these environments. However, H64 exhibited the highest factor score and the best fit to the regression line. On the other hand, slopes close to zero are observed for genotypes with yield stability across the set of environments. For example, the hybrid H17 did not responded to the environmental changes, i.e. showed stable performance across environments. Additionally, H17 also showed the highest predicted breeding values among all the evaluated genotypes.

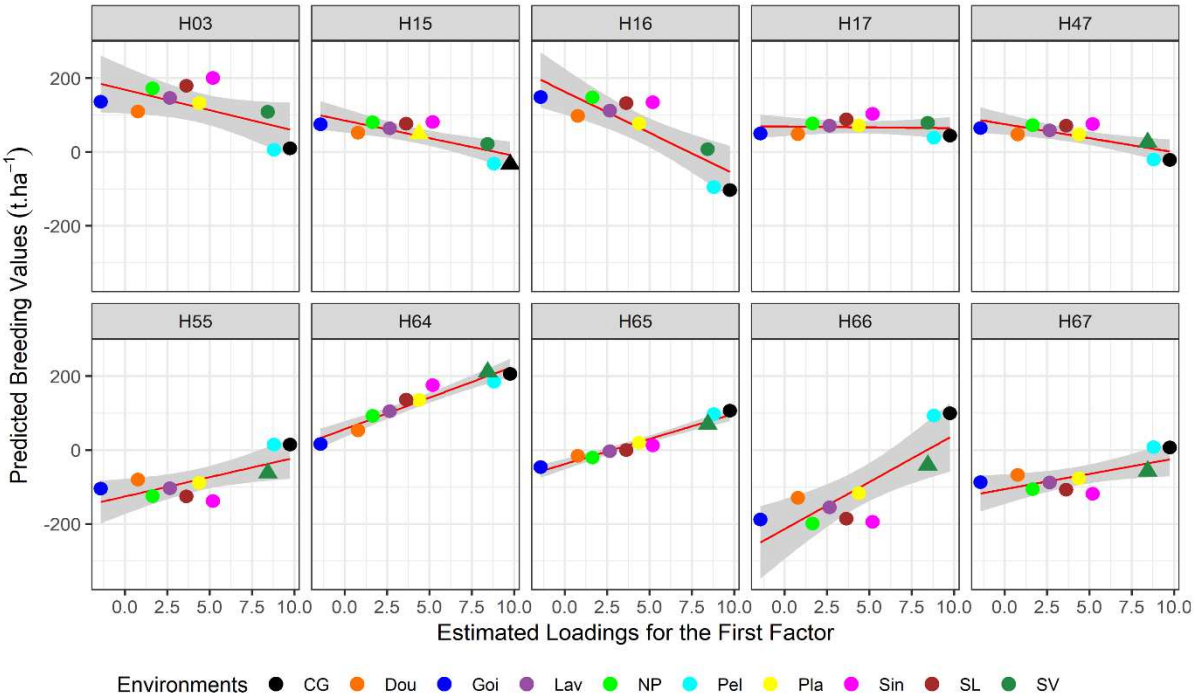


Figure 4: Latent regression plots of the first factor (FA1) for the best 10 genotypes grown in at least eight locations. The circles correspond to predicted breeding values of genotypes in tested locations, and the triangles to predicted breeding values of genotypes in untested locations. The solid red line and the gray shade correspond to the latent regression line and to the confidence interval of 95%, respectively.

For the second factor, Dou (FA2 = 5.57, orange points), Goi (FA2 = 6.78, dark-blue points), Lav (FA2 = 7.53, purple points), NP (FA2 = 8.76, light-green points), Pla (FA2 = 6.89, yellow points), Sin (FA2 = 10.39, pink points) and SL (FA2 = 9.25, magenta points) showed the highest loadings. The latent regression plots on the second factor highlight the hybrid H65 as the most stable genotype, showing the smallest factor score, i.e. the smallest slope for the latent regression line. By contrast, H3 and H16 were more responsive genotypes in the above environments. Additionally, Figures 4 and 5 also show changes from downward to upward direction for the genotypes H3 and H16, indicating a positive response of these hybrids to the environmental loadings of the second factor, and from upward to downward direction for H55, H66 and H67, showing a negative response of these hybrids to second factor environmental loadings.

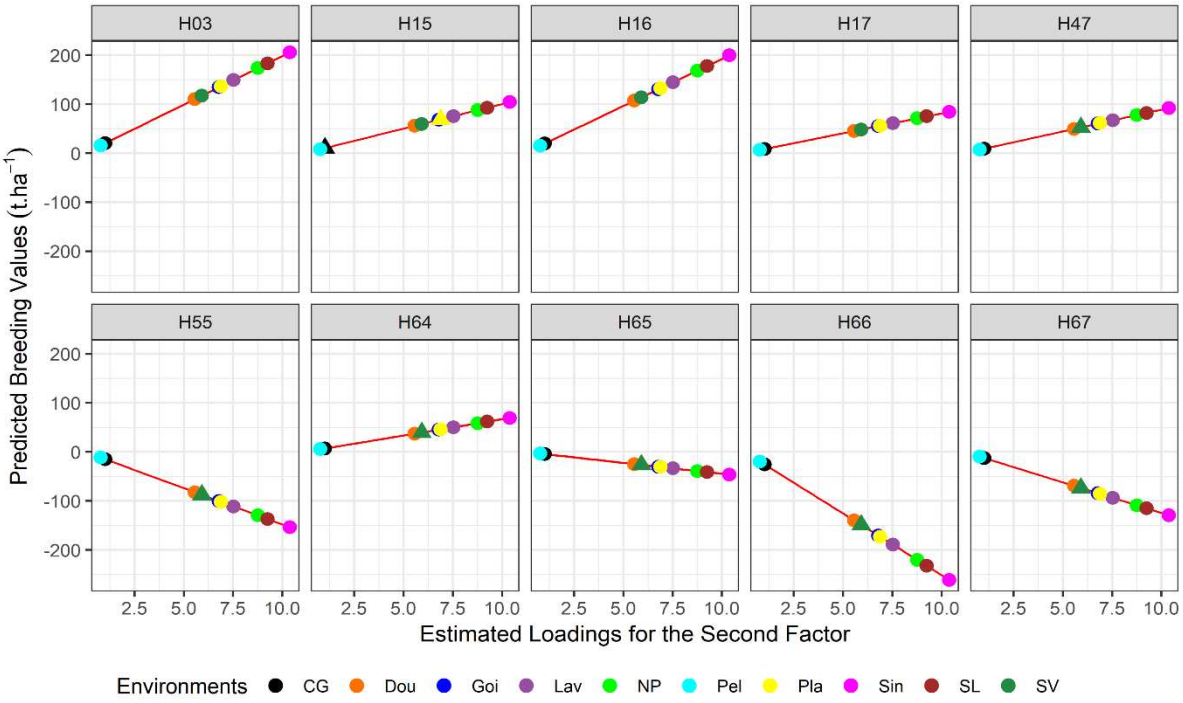


Figure 5: Latent regression plots of the second factor (FA2), for the best 10 genotypes grown in at least eight locations. The circles correspond to predicted breeding values of genotypes in tested locations, and the triangles to predicted breeding values of genotypes in untested locations. The solid red line corresponds to the latent regression line.

Table 3 presents the 10 hybrids presenting the highest overall predicted means, which were evaluated in at least eight trials. Among the most productive materials, the hybrid H17 was the most stable, presenting the lowest slope for latent regression lines on factor FA1 and FA2 rotated loadings. It is also possible to highlight the hybrid H64 as one of the highly adapted genotypes across the set of evaluated environments, given the high and positive slope for latent regression lines of the two factor rotated loadings.

Table 3: Rank of the best 10 genotypes grown in at least eight environments, based on their overall predicted means ($\bar{\mu}$); the slope (genotype score) of the latent regression lines for the first [CA (FA1)] and the second [CA (FA2)] factors.

Hybrids	$\bar{\mu}$	CA(FA1)	CA(FA2)
H03	75.80	-11.08	19.80
H15	74.04	-9.56	10.02
H16	77.71	-22.29	19.22
H17	76.59	-0.45	8.11
H47	76.18	-7.62	8.87
H55	74.56	10.47	-14.81
H64	76.48	17.02	6.66
H65	74.95	13.64	-4.49
H66	79.58	25.46	-25.15
H67	75.69	8.26	-12.46

Discussion

The Embrapa's biomass sorghum breeding program constantly seeks for more productive, more adapted and more stable cultivars for a wide range of environments. In this sense, multi-environment trials (MET) are annually executed to evaluate yield performance of genotypes in different environments, distributed across the Brazilian territory, to cover distinct regions and edaphoclimatic conditions. Moreover, multi-location multi-year trials, called Value for Cultivation and Use (VCU) trials, are required by the Ministry of Agriculture, Livestock and Supply, to register new cultivars for commercial-use. The release of new cultivars only occurs when they present specific features according to the Cultivar Registration form, highlighting the importance of MET studies in breeding programs.

Factor analytic (FA) mixed model has emerged as a flexible and robust approach for modeling genetic variance-covariance (VCOV) matrices, being more parsimonious for MET analyses than unstructured models (Smith et al., 2001a). Additionally, it allows the analysis of unbalanced data, which is extremely important for breeding programs since low-performance genotypes are often replaced by newly-developed materials over the

years. G × E studies have been performed using FA models to understand the adaptability and stability of genotypes across environments (Dias et al., 2018; Li et al., 2017; Peixouto et al., 2016), and also to define mega-environments in plant breeding programs (Monteverde et al., 2018; Smith et al., 2015; Smith and Cullis, 2018).

In the present study, factor analysis allowed identifying genotypes specifically adapted to a given environment, such as the hybrids H31, H32 and H47 to Goi, as well as genotypes exhibiting yield stability across distinct environments, such as the H12, H22, H38 and H43 to the environments evaluated in this study. In breeding programs, stable hybrids are highly desired, showing the relevance of these results. Additionally, latent regression plots also allowed selecting, among the best 10 genotypes, those specifically adapted to a given set of environments. For example, the genotypes H64 and H66 were selected for the environments CG, Pel and SV, H3 and H16 for Dou, Goi, Lav, NP, Pla, Sin and SL. Moreover, it was also possible to select genotypes showing yield stability across some environments, such as the hybrids H17 to CG, Pel and SV, and H65 to all other environments.

Although FA analysis allowed an efficient identification of hybrids exhibiting adaptability to a given environment, as well as hybrids showing yield stability across a set of environments, no clear patterns were observed to cluster locations in mega-environments. However, high genetic correlations were observed among some of the environments, suggesting that the evaluated hybrids exhibited very similar genetic responses across these environments. In this case, the hybrid performance observed in a given environment can help to predict untested hybrids in correlated environments (Smith et al., 2015). Several studies have been focused on the prediction of untested genotypes, such as Dias et al. (2019) e Burgueño et al. (2012). These authors presented interesting results about the accuracy of MET genomic selection (MET-GS) models to predict

genotypes not tested in any environment, or genotypes tested in some locations but not in others, based on the genetic relationship among tested and untested genotypes. MET-GS models provided better accuracies when predicting the performance of genotypes that were evaluated in other environments.

Latent regression plots of the predicted breeding values on FA1 and FA2 identified the hybrid H17 exhibiting yield stability to most the evaluated environments, besides being the fourth highest productive genotype. Additionally, this hybrid was located in the first quadrant of the FA biplot (having positive scores for FA1 and FA2), showing adaptability to the majority of environments, which were also located in the first quadrant. H17 is an experimental hybrid, listed as a promising new-developed cultivar to be released by Embrapa in the next years, confirming its superiority. By contrast, H64 and H65 were considered as responsive and non-responsive genotypes to the FA1 and FA2 environmental loadings, respectively, suggesting that these hybrids exhibited high adaptability to environments with the highest FA1 loadings (CG, Pel and SV), and stable performance across the environments with the highest FA2 loadings (Sin, NP and SL). Based on the FA biplot, H64 also showed adaptability to most of the evaluated environments, besides being the fifth highest productive genotype, according to its overall predicted mean through $FA_{(2)}$ model. Moreover, based on the latent regression plots on the second factor, the hybrids H03, H16 and H17 were the most responsive genotypes to the environmental loadings of this factor, and also presented the highest FBY. The H16 is a commercial hybrid of biomass sorghum, named BRS 716, released by the Embrapa in 2014, which confirms its superiority. This hybrid is indicated for cultivation in a wide range of Brazilian regions, mainly in the Southeast and Central-West regions, being able to reach fresh biomass productivities of up to 150 t ha^{-1}.

Overall low productivities were observed for the checks H68 and H24 (Table S3), cultivars commercially released in Brazil by two distinct private companies as biomass and forage hybrids, respectively. Other checks, the parental R-line used to produce H21, H22 and H23, and the hybrid H25 (BRS 655), also presented an overall low FBY performance. However, the low performance observed for BRS 655 can be explained by the fact that it is a non-photosensitive forage cultivar, with an expected lower FBY (approximately 55.00 t ha^{-1}) compared to biomass sorghum. The above mentioned genotypes were among the 20 treatments with the lowest overall predicted means (Table S3). Moreover, H68 did not showed adaptability to any of the evaluated environments. The lack of fit by the latent regression models (Figure 4) for some hybrids, such as H66, shows that the present G × E study was not conclusive for them, suggesting the need to expand the stage of hybrid testing to other Brazilian regions.

Latent regression plots allowed studying the genotype performance across different environments, showing their responses to environmental changes. According to Smith et al. (Smith et al., 2015), environmental loadings are difficult to interpret, being more consistent when environmental covariates are included in the FA models. Thus, the FA models applied in the present study can be extended to include some edaphoclimatic covariates, such as temperature, solar radiation, air moisture, rainfall, soil type, soil physical and chemical composition, among others. However, these covariates are not available for the evaluated trials, but they can be collected and used to analyze future biomass sorghum breeding trials. This information can be used to investigate the environmental factors affecting genotypes performance of biomass sorghum, helping breeders to identify the most suitable genotypes for each environment. Moreover, the FA models developed in this study can be easily extended to incorporate genomic relationship matrices, which can be used to predict the performance of untested genotypes across environments via MET-GS, being an interesting approach to accelerate the genetic gains in biomass sorghum breeding programs.

Conclusion

The hybrids H16, H17 and H64 stood out for their high fresh biomass yield and for the best response to the environmental improvements, showing adaptability across the evaluated environments, and stability in some environments. High genetic correlations were found among environments, despite the presence of high G × E interaction. FA models can be a useful tool to provide estimates of genetic parameters in sorghum breeding programs using MET analyses.

Acknowledgments

The authors thank all the scientists, research assistants, undergraduate and graduate students enrolled with the experimental design, management and data collection at Embrapa Maize and Sorghum, and all the graduate students of the Federal University of Viçosa, who indirectly collaborated in carrying out this study.

Financing

This work was supported by BNDES (Brazilian National Bank for Economic and Social Development), CNPq (National Science and Technology Development Council), CAPES (National Council for the Improvement of Higher Education), Fapemig (Research Support Foundation of Minas Gerais) and Embrapa (Brazilian Agricultural Research Corporation).

References

Almeida, L.G.F. de, Parrella, R.A. da C., Simeone, M.L.F., Ribeiro, P.C.D.O., Santos, A.S. dos, da Costa, A.S.V., Guimarães, A.G., Schaffert, R.E., 2019. Biomass and Bioenergy Composition and growth of sorghum biomass genotypes for ethanol production 122, 343–348.

Borém, A., Pimentel, L., Parrella, R., 2014. Sorgo: do plantio à colheita. Univ. Fed. Viçosa.

Bornhofen, E., Todeschini, M.H., Stoco, M.G., Madureira, A., Marchioro, V.S., Storck, L., Benin, G., 2018. Wheat Yield Improvements in Brazil: Roles of Genetics and Environment. Crop Sci.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370.

Brenton, Z.W., Cooper, E.A., Myers, M.T., Boyles, R.E., Shakoor, N., Zielinski, K.J., Rauh, B.L., Bridges, W.C., Morris, G.P., Kresovich, S., 2016. A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. Genetics genetics-115.

Burgueño, J., Crossa, J., Cornelius, P.L., Yang, R.C., 2008. Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. Crop Sci. 48, 1291–1305.

Burgueño, J., de los Campos, G., Weigel, K., Crossa, J., 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52, 707–719.

Butler, D.G., Cullis, B.R., Gilmour, A.R., Gogel, B.J., 2009. ASReml-R reference manual. State Queensland, Dep. Prim. Ind. Fish. Brisbane.

Cullis, B.R., Jefferson, P., Thompson, R., Smith, A.B., 2014. Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in

outcrossing plant species with application to a Pinus radiata breeding programme. Theor. Appl. Genet. 127, 2193–2210.

Cullis, B.R., Smith, A.B., Beeck, C.P., Cowling, W.A., 2010. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome 53, 1002–1016.

Cullis, B.R., Smith, A.B., Coombes, N.E., 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11, 381.

de Oliveira, A.A., Pastina, M.M., da Costa Parrella, R.A., Noda, R.W., Simeone, M.L.F., Schaffert, R.E., de Magalhães, J.V., Damasceno, C.M.B., Margarido, G.R.A., others, 2018. Genomic prediction applied to high-biomass sorghum for bioenergy production. Mol. Breed. 38, 49.

de Resende, M.D.V., Thompson, R., 2004. Factor analytic multiplicative mixed models in the analysis of multiple experiments. Rev. Mat. Estat 22, 31–52.

Demirbas, M.F., Balat, M., Balat, H., 2009. Potential contribution of biomass to the sustainable energy development. Energy Convers. Manag. 50, 1746–1760.

Dias, K.O.D.G., Gezan, S.A., Guimarães, C.T., Parentoni, S.N., Guimarães, P.E. de O., Carneiro, N.P., Portugal, A.F., Bastos, E.A., Cardoso, M.J., Anoni, C. de O., de Magalhães, J.V., de Souza, J.C., Guimarães, L.J.M., Pastina, M.M., 2018. Estimating genotype × environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. Crop Sci. 58, 72–83.

Dias, K.O.D.G., Gezan, S.A., Guimarães, C.T., Parentoni, S.N., Guimarães, P.E. de O., Carneiro, N.P., Portugal, A.F., Bastos, E.A., Cardoso, M.J., Anoni, C. de O., others, 2017. Estimating Genotype × Environment Interaction for and Genetic Correlations among Drought Tolerance Traits in Maize via Factor Analytic Multiplicative Mixed Models. Crop Sci.

Dias, K.O.G., Piepho, H.P., Guimarães, L.J.M., Guimarães, P.E.O., Parentoni, S.N., Pinto, M.O., Noda, R.W., Guimarães, C.T., Garcia, A.A.F., Pastina, M.M., 2019. Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. Theor. Appl. Genet. 1–22.

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 1440–1450.

Kelly, A.M., Smith, A.B., Eccleston, J.A., Cullis, B.R., 2007. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Sci. 47, 1063–1070.

Kempton, R.A., 1984. The use of biplots in interpreting variety by environment interactions. J. Agric. Sci. 103, 123–135.

Li, Y., Suontama, M., Burdon, R.D., Dungey, H.S., 2017. Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. Tree Genet. Genomes 13, 60.

Lombardi, G.M.R., Navegantes, P.C.A., Pereira, C.H., Fonseca, J.M.O., Parrella, R.A. da C., Castro, F.M.R., Rocha, M.J. da, Ornelas, D.O., Bruzi, A.T., Nunes, J.A.R., 2018. Heterosis in sweet sorghum. Pesqui. Agropecuária Bras. 53, 593–601.

Monteverde, E., Rosas, J.E., Blanco, P., de Vida, F., Bonnecarrère, V., Quero, G., Gutierrez, L., McCouch, S., 2018. Multienvironment Models Increase Prediction Accuracy of Complex Traits in Advanced Breeding Lines of Rice. Crop Sci. 58, 1519–1530.

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

Murakami, D.M., Cruz, C.D., 2004. Proposal of methodologies for environment stratification and analysis of genotype adaptability. Crop Breed. Appl. Biotechnol. 4.

Parrella, R.A. da C., Schaffert, R.E., May, A., Emygdio, B., Portugual, A.F., Damasceno,

C.M.B., 2011. Desempenho agronômico de híbridos de sorgo biomassa. Embrapa Milho e Sorgo-Boletim Pesqui. e Desenvolv.

Parrella, R.A. da C., Souza, V.F. de, Parrella, N.N.L.D., others, 2016. Maturation curves of sweet sorghum genotypes. Ciência e Agrotecnologia 40, 46–56.

Peixouto, L.S., Nunes, J.A.R., Furtado, D.F., 2016. Factor analysis applied to the G+ GE matrix via REML/BLUP for multi-environment data. Crop Breed. Appl. Biotechnol. 16, 1–6.

Piepho, H. P., 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor. Appl. Genet. 97, 195–201.

Piepho, H. P., 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics 761–766.

Quintero, A., Molero, G., Reynolds, M.P., Calderini, D.F., 2018. Trade-off between grain weight and grain number in wheat depends on GxE interaction: A case study of an elite CIMMYT panel (CIMCOG). Eur. J. Agron. 92, 17–29.

Reis, A.L.S., Damilano, E.D., Menezes, R.S.C., de Morais Jr, M.A., 2016. Second-generation ethanol from sugarcane and sweet sorghum bagasses using the yeast Dekkera bruxellensis. Ind. Crops Prod. 92, 255–262.

Resende, M.D.V. de, Duarte, J.B., 2007. Precisão e controle de qualidade em experimentos de avaliação de cultivares 37, 182–194.

Schwarz, G., others, 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Smith, A., Cullis, B., Gilmour, A., 2001a. The analysis of crop variety evaluation data in Australia. Aust. New Zeal. J. Stat. 43, 129–145.

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

Smith, A.B., Cullis, B.R., 2018. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. Euphytica 214, 143.

Smith, A.B., Cullis, B.R., Thompson, R., 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J. Agric. Sci. 143, 449–462.

Smith, A.B., Ganesalingam, A., Kuchel, H., Cullis, B.R., 2015. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128, 55–72.

Team, R.C., others, 2015. R: A language and environment for statistical computing.

Thompson, R., Cullis, B., Smith, A., Gilmour, A., 2003. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust. N. Z. J. Stat. 45, 445–459.

Supplementary Information

Table S1: Estimates of the phenotypic mean (Mean), genetic (σ_g^2) and residual (σ_e^2) variance components, generalized heritability (H^2), experimental accuracy (Ac) and coefficient of variation (CV%) for the 42 trials conducted by the Embrapa's biomass sorghum breeding program to evaluate fresh biomass yield (FBY).

Environment	Mean	σ_b^2	σ_g^2	σ_e^2	H^2	Ac	CV%
NP.12	66.22	0.85	116.10	45.00	0.88	0.88	10.13
Pel.12	62.23	0.88	172.94	51.70	0.91	0.90	11.55
SL.12	84.30	0.00	339.55	104.15	0.91	0.90	12.11
NP.13S*	61.55	0.00	183.88	100.70	0.85	0.83	16.30
SV.13S	77.29	0.00	703.74	104.63	0.95	0.95	13.23
SL.13S	84.34	13.36	604.83	51.14	0.97	0.97	8.48
Goi.13	50.64	0.00	92.32	86.22	0.76	0.72	18.34
Lav.13	74.93	73.04	187.36	143.81	0.77	0.74	16.00
NP.13	74.45	42.27	190.27	33.52	0.93	0.93	7.78
SV.13	60.69	0.00	43.06	57.89	0.69	0.62	12.54
SL.13	64.12	37.40	238.51	38.28	0.94	0.94	9.65
Sin.13	82.06	0.00	524.51	217.43	0.88	0.87	17.97
Dou.14	84.53	0.00	309.89	155.32	0.86	0.84	14.74
Goi.14	45.73	9.31	17.28	25.46	0.65	0.55	11.03
Lav.14	82.79	0.31	256.31	119.95	0.86	0.85	13.23
NP.14	70.50	0.00	96.35	78.36	0.79	0.76	12.56
Pel.14	35.60	8.70	27.14	20.62	0.78	0.75	12.76
SL.14	41.61	4.26	82.05	62.00	0.79	0.77	18.93
Sin.14	68.70	3.01	104.21	36.01	0.89	0.89	8.74
CG.15	27.66	1.05	11.71	14.01	0.71	0.64	13.53
Dou.15	102.97	41.53	89.43	79.72	0.75	0.71	8.67
Goi.15	73.06	0.00	69.19	73.59	0.74	0.69	11.74
Lav.15	68.11	9.31	234.98	81.61	0.89	0.88	13.26
NP.15	102.23	86.19	366.16	103.03	0.90	0.90	9.93
Pel.15	33.21	4.55	16.35	39.30	0.54	0.29	18.88
SL.15	66.03	5.12	173.94	76.30	0.87	0.86	13.23
Sin.15	74.16	17.65	125.98	39.71	0.89	0.89	8.50
CG.16	79.12	3.32	107.21	46.10	0.87	0.86	8.58
Dou.16	94.08	0.00	95.52	164.81	0.63	0.52	13.65
Goi.16	52.89	10.07	83.90	104.32	0.70	0.63	19.31
NP.16	93.79	118.66	334.92	164.28	0.84	0.83	13.67
Lav.16	64.54	0.00	126.72	68.60	0.85	0.83	12.83
Pel.16	52.66	0.00	71.53	47.43	0.82	0.80	13.08
Pla.16	60.91	0.00	155.63	22.24	0.95	0.95	7.74
SL.16	75.09	11.51	192.30	55.32	0.91	0.90	9.91
CG.17	82.42	55.56	157.20	118.16	0.78	0.75	13.19
Goi.17	65.51	0.00	199.23	54.05	0.92	0.91	11.22

... continue on next page

Lav.17	50.73	0.00	131.78	45.01	0.90	0.89	13.22
NP.17	88.25	49.14	493.92	150.61	0.90	0.89	13.91
Pla.17	54.50	2.57	123.56	16.46	0.95	0.95	7.45
SL.17	81.09	0.00	231.03	59.58	0.92	0.92	9.52
Sin.17	70.99	5.01	448.13	104.20	0.93	0.92	14.38

* S: 2nd crop of the respective site and year. CG.; Dou.; Goi.; Lav.; NP.; Pla.; SV.; SL.; Pel. and Sin. refer to the municipalities of Campos dos Goytacazes-RJ; Dourados-MS; Goiânia-GO; Lavras-MG; Nova Porteirinha-MG; Planaltina-DF; Santa Vitória-MG; Sete Lagoas-MG; Pelotas-RS; Sinop-MT, respectively. The numbers 12; 13; 14; 15; 16; and 17 are the evaluation years between 2012 to 2017, respectively.

Table S2: Overall predicted mean ($\bar{\mu}$) of genotypes and their respective scores of the first two factors [FA(1) and FA(2)] for the 20 best biomass sorghum genotypes in the MET analysis.

Treatment	$\bar{\mu}$	FA1	FA2
H19	84.60	-0.72	15.77
H66	79.58	10.44	-26.22
H16	77.71	-10.74	20.31
H17	76.59	4.40	7.70
H64	76.48	20.91	4.53
H29	76.48	2.45	6.57
H01	76.27	22.49	4.27
H47	76.18	-2.28	9.12
H03	75.80	0.86	19.81
H67	75.69	0.82	-12.58
H65	74.95	10.89	-5.57
H55	74.56	1.62	-15.01
H15	74.04	-3.48	10.39
H54	73.67	12.65	-45.80
H39	73.59	-9.49	-16.06
H14	73.15	-6.45	16.64
H63	73.03	2.15	-20.38
H61	71.67	21.20	-22.90
H51	71.54	9.58	-15.52
H56	71.39	-0.35	-9.82

Table S3: Overall predicted mean ($\bar{\mu}$) of genotypes and their respective scores of the first two factors [FA(1) and FA(2)] for the twenty biomass sorghum genotypes exhibiting the smallest fresh biomass yield (FBY) in the MET analysis.

Treatment	$\bar{\mu}$	FA1	FA2
H09	64.35	-14.76	0.23
H62	64.19	7.34	-4.66
H53	64.11	-0.55	-11.51
H22	64.01	-0.58	0.33
H60	63.87	-0.08	-8.96
H36	63.66	-13.04	2.05
H68	63.53	0.62	-4.78
H11	63.49	-2.33	23.43
H52	63.48	6.90	3.44
H58	63.05	6.23	-2.53
H23	62.32	10.21	14.39
H50	61.71	14.71	-17.57
H43	61.35	-0.08	1.43
H21	61.06	1.42	2.25
H31	60.39	0.38	8.06
H42	60.29	-4.28	-24.31
H70	59.40	4.24	-1.77
H69	55.70	9.37	0.47
H24	49.67	-3.86	9.34
H25	41.55	12.61	2.65

CAPÍTULO 2

USO DE DADOS HISTÓRICOS DE PROGRAMAS DE MELHORAMENTO PARA A PREDIÇÃO GENÔMICA PARA PRODUÇÃO DE MASSA VERDE EM HÍBRIDOS DE SORGO BIOMASSA

RESUMO

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, setembro de 2019. **Uso de dados históricos de programas de melhoramento para a predição genômica para produção de massa verde em híbridos de sorgo biomassa**. Orientador: José Eustáquio de Souza Carneiro. Coorientador: Rafael Augusto da Costa Parrella e Pedro Crescêncio Souza Carneiro.

O sorgo biomassa é uma fonte de energia renovável, com grande potencial para promover a redução do consumo de combustíveis fósseis e de outras fontes energéticas não-renováveis. Os programas de melhoramento genético buscam constantemente por materiais com bom desempenho produtivo em múltiplos ambientes. Esse processo de desenvolvimento de cultivares pode ser ainda mais rápido e efetivo com o auxílio de técnicas de seleção genômica. No presente trabalho, a acurácia da predição genômica de híbridos em múltiplos ambientes, e também, de híbridos ainda não sintetizados foi examinada para diferentes esquemas de validação cruzada, considerando modelos com a inclusão de efeitos genéticos aditivos e aditivo-dominantes. Para isso, foi utilizado um conjunto de dados fenotípicos composto por 202 híbridos de sorgo biomassa avaliados em ensaios de valor de cultivo e uso e em ensaios preliminares de avaliação de híbridos, cujos genitores foram genotipados via GBS (do inglês, Genotyping-by-Sequencing), durante sete safras ao longo de dez diferentes locais. Primeiramente, os genótipos dos híbridos foram inferidos a partir dos genótipos dos seus genitores. Foram testados seis esquemas de validação cruzada para a predição de híbridos em múltiplos ambientes, que apresentaram acurácia de predição variando de 0,26 a 0,77 com média de 0,50, e mostraram que o relacionamento entre a população de treinamento e a população de validação podem ser um dos principais fatores que influenciam a acurácia de predição dos modelos de seleção genômica. Nos esquemas de validação para a predição de híbridos ainda não sintetizados, as acurácias médias obtidas variaram de 0,49 a 0,86 e média de 0,74, dado que o aumento do número de genitores na população de validação que não participaram de outros cruzamentos presentes na população de treinamento, causou uma queda na médias das acurácias de predição. Além disso, observou-se que, em geral, a inclusão da matriz de efeitos de dominância levaram a um aumento da acurácia de predição dos modelos propostos. A partir desses resultados, é possível afirmar que a seleção genômica pode ser utilizada de forma eficiente para acelerar o desenvolvimento de novos cultivares em programas de melhoramento de sorgo biomassa.

Palavras-chave: *Sorghum bicolor* (L.) Moench; bioenergia; seleção genômica; múltiplos ambientes; híbridos não sintetizados; modelos mistos.

Abreviações: *APH:* avaliação preliminar de híbridos*;* G × A, interação genótipo × ambiente; PMV: produção de massa verde; PMS: produção de massa seca; VC: validação cruzada; VCOV: variância-covariância; VCU: valor de cultivo e uso.

ABSTRACT

OLIVEIRA, Isadora Cristina Martins, D.Sc., Universidade Federal de Viçosa, September, 2019. **Use of historical data from breeding programs for genomic prediction for fresh mass production in biomass sorghum hybrids**. Advisor: José Eustáquio de Souza Carneiro. Co-advisor: Rafael Augusto da Costa Parrella and Pedro Crescêncio Souza Carneiro.

Biomass sorghum is a promising renewable energy source with great potential to promote the reduction in the use of fossil fuels and other non-renewable energetic materials. Breeding programs are constantly looking for high-yielding cultivars for different environments. This process of developing new cultivars can be even more rapid and efficient with the use of genomic selection (GS) techniques. In this study, the prediction accuracy of hybrids in multiple environments and of non-synthesized hybrids were evaluated in a practical context of a biomass sorghum breeding program, considering predictive models incorporating only additive (A) as well as additive-dominant (AD) genetic effects. For this, a phenotypic dataset comprised of 202 hybrids evaluated in the v*alue* for *cultivation and use* trials and in the preliminary trials of hybrid testing along seven harvests across ten distinct locations. The genotypes of the hybrids were inferred through the genotypes of their parents (inbred lines). Six cross validation schemes for predicting hybrids in multiple environments were examined, which had a mean prediction accuracy ranging from 0.26 to 0.77 with a mean of 0.50, and the results showed that the relationship between the training and the validation set can be one of the main factors influencing the predictive accuracy of the models. In the cross-validation schemes for the prediction of un-synthesized hybrids, the average accuracy obtained ranged from 0.49 to 0.86 and an average of 0.74, given that the increase in the number of unrepresented parents in the training set caused a decrease in the average prediction accuracy. Moreover, it was also observed that the inclusion of the dominance effects leads to an increase in the accuracy of the predictive models. Based on these results, it is possible to affirm that the genomic selection can be used efficiently to accelerate the process of new cultivars development in biomass sorghum breeding programs.

Keywords: *Sorghum bicolor* (L.) Moench; bioenergy; genomic selection; multiple environments; non-synthesized hybrids; mixed models.

Abbreviations: APH: preliminary evaluation of hybrids; G × E, genotype × environment interaction; PMV: fresh biomass yield; PMS: dry biomass yield; CV: cross validation; VCOV: variance-covariance; VCU: *value* for *cultivation and use*.

Introdução

As altas produtividades de massa verde e o alto poder de combustão fazem do sorgo biomassa [*Sorghum bicolor* (L.) Moench] uma matéria-prima de grande potencial para a produção de vapor e cogeração de energia [1,2]. Com o intuito de buscar genótipos de sorgo mais produtivos, adaptados e estáveis a uma ampla gama de ambientes, ensaios em múltiplos ambientes são anualmente conduzidos em programas de melhoramento para a avaliação de híbridos de sorgo biomassa ao longo de várias regiões do país, locais e anos [3]. Mas devido à necessidade de avaliações em vários anos e locais, que são requisitos para a certificação e registro de cultivares em órgãos governamentais, este processo têm-se mostrado moroso e de alto custo para os programas de melhoramento [4,5].

Outra dificuldade encontrada pelos melhoristas de sorgo é a baixa eficiência da seleção de genitores para a síntese de híbridos em estágios iniciais do programa. Por exemplo, Oliveira et al. [2], relatam que os efeitos de epistasia dos genes de nanismo (*dw1-dw4*) no controle genético de características relacionadas à produção de biomassa promovem um baixo desempenho das linhagens A (macho-estéreis), o que não possibilita a seleção de híbridos pelo desempenho "*per se*" dos seus genitores. Dessa forma, é essencial o estudo da capacidade de combinação dos genitores em análises dialélicas, o que também aumenta os gastos e o tempo de avaliação em campo. Além disso, a avaliação de todas as combinações híbridas em campo é inviável [6], devido à ampla gama de híbridos possíveis de serem sintetizados dado um pequeno grupo de genitores.

Neste contexto, a seleção genômica (GS, do inglês *genomic selection*) [7] vem ganhando espaço nos programas de melhoramento genético de plantas, e tem promovido o aumento do ganho genético por unidade de tempo [4,8–10]. Dado que o objetivo principal da seleção genômica é predizer o desempenho de genótipos ainda não avaliados, com base em dados genotípicos e fenotípicos de indivíduos geneticamente relacionados [11], essa estratégia pode

promover o descarte precoce de materiais geneticamente inferiores, diminuindo o número de materiais de baixo potencial que serão testados em campo, e consequentemente, resultando em uma redução no tempo de duração dos ciclos de melhoramento e nos custos de fenotipagem [8,12,13].

A seleção genômica também tem mostrado resultados favoráveis no melhoramento de plantas a partir da incorporação de efeitos genéticos aditivos e não-aditivos em modelos genético-estatísticos de predição genômica. Híbridos de sorgo biomassa apresentam produções de massa verde superiores em relação aos seus genitores, podendo em alguns casos corresponder ao dobro da média dos pais [14]. Duas hipóteses sobre os efeitos gênicos responsáveis por essa superioridade são estudadas. A primeira sugere a presença de heterose devido aos efeitos de dominância [15], e a segunda considera a existência de efeitos epistáticos em linhagens genitoras (Linhagens A macho-estéreis), como o principal causador da superioridade dos híbridos [2]. No entanto, nenhuma delas se baseia em informações genômicas, o que mostra a importância da interpretação destes efeitos para os programas de melhoramento, a partir de modelos que incorporam informações de marcadores moleculares.

Estudos de seleção genômica já foram realizados para características de produção de massa verde em sorgo biomassa, mostrando alta acurácia de predição [10,13,16,17]. Entretanto, nenhum destes trabalhos considera os efeitos da interação entre genótipos e ambientes (G × A). Dado que a incorporação dos efeitos ambientais aumenta a acurácia de predição dos modelos de GS [18], Hunt et al. [4] relataram a necessidade da inclusão de dados de múltiplos ambientes e dos efeitos de G × A nas análises de seleção genômica em sorgo, uma vez que a produção de massa verde é controlada por muitos genes de efeitos pequenos e, consequentemente, muito influenciada pelo ambiente [5,19].

A análise conjunta de dados de múltiplos ambientes é um dos maiores desafios para a utilização de dados fenotípicos históricos para o ajuste de modelos de seleção genômica em

programas de melhoramento [20]. Nesse contexto, algumas estratégias têm sido propostas para facilitar o processamento dessas análises, por exemplo, análises fenotípicas em dois estágios [21], que reduz a demanda computacional, devido à simplificação dos dados de entrada nas análises de seleção genômica, diminuindo o tempo para convergência dos modelos e estimativa dos parâmetros genético-estatísticos [22]. Assim, a partir de análises fenotípicas em dois estágios, é possível ajustar com maior facilidade estruturas complexas para as matrizes de variâncias e covariâncias genéticas entre ambientes, além de permitir a inclusão de matrizes de relacionamento genético entre indivíduos, via informações de pedigree e/ou de marcadores moleculares. Tal estratégia de análise em dois estágios tem resultados tão acurados quanto às análises fenotípicas realizadas em um único estágio, ou seja, diretamente a partir dos dados experimentais, desde que as informações residuais do primeiro estágio sejam incorporadas nas análises de segundo estágio para a predição dos componentes de variância e covariância genéticos [4,21,23].

Dessa forma, os objetivos do presente trabalho foram: (i) avaliar a acurácia da seleção genômica (GS) para a predição do desempenho de híbridos de sorgo biomassa para produção de massa verde, utilizando dados de marcadores moleculares do tipo SNP e dados fenotípicos de múltiplos ambientes; (ii) comparar a acurácia preditiva obtida a partir de modelos aditivos (Modelo A) e de modelos aditivo-dominantes (Modelo AD); (iii) avaliar diferentes esquemas para a predição genômica de híbridos ainda não realizados no contexto de programas de melhoramento de sorgo.

Material e Métodos

1. Material vegetal

Foram avaliados dois tipos de ensaios, o de valor de cultivo e uso (VCU) e o de avaliação preliminar de híbridos (APH), que juntos integraram 202 híbridos derivados do cruzamento entre 17 linhagens R (macho-férteis) e 46 linhagens A (macho-estéreis) do programa de melhoramento de sorgo biomassa da Embrapa. Os ensaios de VCU foram conduzidos em dez locais, alocados em sete diferentes estados brasileiros, nas safras 2012/13 a 2017/18, totalizando 40 ensaios. Também foram considerados seis ensaios de APH, sendo cinco conduzidos em Sete Lagoas/MG nos anos agrícolas de 2012/13, 2014/15, 2015/16 e 2017/18, e um ensaio conduzido em Nova Porteirinha/MG no ano agrícola 2014/15. Dessa forma, totalizou-se 46 ensaios avaliados.

2. Design experimental

Os ensaios foram conduzidos em látice triplo, em pelo menos três locais por safra para os experimentos de valor de cultivo e uso (VCU), e em um local, Sete Lagoas-MG, para os ensaios preliminares. Nas avaliações preliminares de híbridos, de 2012 a 2017, foram avaliados 182 híbridos experimentais, e nas avaliações de VCU, de 2012 a 2017, um total de 58 híbridos experimentais. Quanto à composição dos experimentos, é importante ressaltar que tanto entre as safras, quanto entre os ensaios de avaliação de híbridos, APHs e VCUs, houve desbalanceamento dos dados, ou seja, os genótipos avaliados foram diferentes entre os diferentes ensaios e safras.

Na Tabela 1, estão representados os experimentos de APHs, sendo que na diagonal estão os números de genótipos avaliados em cada ensaio APH correspondente e entre parênteses o número de ensaios conduzidos e, acima da diagonal principal, estão os números de genótipos

comuns entre os APHs nos diferentes anos. Nota-se que há elevado desbalanceamento em relação aos genótipos entre as safras avaliadas. No ano de 2014 foram conduzidos dois ensaios de avaliação de híbridos, o primeiro conduzido em Sete Lagoas no qual foram avaliados 25 híbridos experimentais em ensaios preliminares, e o segundo conduzido em Sete Lagoas e Nova Porteirinha, no qual foram avaliados 36 híbridos e 12 linhagens genitoras em cruzamento dialélico parcial 6 × 6 [2014(A)].

Tabela 1: Número de genótipos comuns entre os experimentos de avaliação preliminar de híbridos (APHs) de sorgo biomassa do programa de melhoramento de sorgo da Embrapa.

Experimento	2012	2014	2014(A)°	2015	2017
2012	89(1*)	8	5	12	7
2014		25(1)	1	19	26
2014(A)			49(2)	8	14
2015				59(1)	47
2017					81(1)

* número de ensaios (ambientes) avaliados em cada experimento de APH; (A)° cruzamento dialélico conduzido no ano de 2014, em Sete Lagoas e Nova Porteirinha.

Na Tabela 2 estão representados os sete experimentos de VCUs, e nota-se também presença de alto desbalanceamento entre eles, mas vale ressaltar que entre os locais de avaliação dentro do mesmo ano de experimento de VCU não houve desbalanceamento.

Tabela 2: Número de genótipos comuns entre os experimentos de valor de cultivo e uso (VCU) de sorgo biomassa do programa de melhoramento de sorgo da Embrapa, conduzidos na primeira safra dos respectivos anos, e na segunda safra no ano de 2013.

Experimento	2012	2013	2013(A)°	2014	2015	2016	2017
2012	22(3*)	5	22	9	6	7	7
2013		16(5)	5	12	7	6	6
2013(A)			22(3)	9	6	7	7
2014				36(7)	17	16	14
2015					25(8)	11	11
2016						25(8)	21
2017							23(6)

*número de ensaios (ambientes) avaliados em cada experimento de VCU; (A)° ensaio de VCU conduzido na segunda safra no ano de 2013.

Entre os experimentos de VCUs e APHs, o número de genótipos avaliados em comum está representado na Tabela 3. Apenas um pequeno número de genótipos foram avaliados em

comum entre os dois tipos de experimento, exceto para o primeiro ano de avaliação de híbridos, cuja maior parte dos genótipos foram comuns entre os dois tipos de experimentos. Este fato ocorre, pois, apenas os híbridos que apresentaram superioridade em mais de dois ensaios de APH são posteriormente avaliados nos ensaios finais de VCU.

Tabela 3: Número de genótipos comuns entre os experimentos de avaliação preliminar de híbridos (APHs) e valor de cultivo e uso (VCU) de sorgo biomassa do programa de melhoramento da Embrapa.

Ensaios		VCUs						
		2012	2013	2013(A)*	2014	2015	2016	2017
	2012	26	16	26	26	23	18	17
A	2014	1	1	1	1	1	2	2
P	2014(A)§	1	1	1	1	1	1	1
H	2015	4	6	4	9	8	6	6
	2017	3	3	3	3	3	4	4

(A)* Ensaio de VCU conduzido em safrinha no ano de 2013. (A) § ensaio de APH conduzido no ano de 2014, em Sete Lagoas e Nova Porteirinha.

A característica avaliada em todos os ensaios foi o peso de matéria verde (PMV), dado pela pesagem da parte aérea, da superfície do solo ao ápice da panícula, de todas as plantas da área útil das parcelas (kg), e posteriormente convertido em tonelada por hectare (t ha^{-1}). As parcelas foram constituídas por duas linhas de cinco metros, espaçadas em 0,70 centímetros, totalizando em uma área útil de sete metros quadrados. A característica peso de matéria verde foi utilizado neste estudo dado a alta correlação com o peso de matéria seca (PMS), característica de grande importância para a cogeração de energia e queima direta em caldeiras [24]. Ainda, o PMV é de mais fácil mensuração quando comparado à PMS, dado o menor número de processos, o que também justifica sua utilização.

Todas as práticas agrícolas foram realizadas conforme recomendado para a cultura do sorgo biomassa.

3. Dados genotípicos

Linhagens parentais do programa de melhoramento de sorgo biomassa foram genotipadas usando a técnica *genotyping-by-sequencing* (GBS) com base no protocolo padrão do GBS [25]. As bibliotecas *Illumina* foram criadas pela digestão do DNA com a enzima de restrição ApeKI e pela adição de adaptadores codificados com códigos de barras únicos para cada amostra de DNA. Três milhões de bibliotecas das 96 amostras foram multiplexadas por célula de fluxo *Illumina* para sequenciamento com o equipamento NextSeq500 V2 (1 x 75pb). Foram genotipadas um total de 96 linhagens genitoras do programa de melhoramento de sorgo biomassa da Embrapa, sendo 34 linhagens B macho-férteis, isogênicas as linhagens A macho-estéreis, e 62 linhagens R.

O *pipeline* para análises de GBS foi implementado usando o software TASSEL v.5. [26]. Primeiro, os dados brutos da *Illumina* foram cortados para remover as leituras que não correspondiam a um código de barras e ao sítio de corte da ApeKI. Em sequência, as tags sequenciadas foram alinhados ao genoma de referência "GCF_000003195.3_Sorghum_bicolor_NCBIv3_genomic" usando a ferramenta de alinhamento Bowtie2 [27]. Em seguida, os polimorfismos de nucleotídeo único (SNPs, do inglês *single nucleotide polymorphisms*) foram definidos para cada amostra com base em uma distribuição binomial das tags exclusivas alinhadas [28] e salvas em um formato de chamada variante (arquivo VCF, do inglês *variant call format*).

Os SNPs foram descartados quando: i) a menor frequência alélica (MAF) foi inferior a 5%; ii) a porcentagem de genótipos perdidos (não observados) foi superior a 25%; e iii) a frequência de genótipos em heterozigose foi superior a 5%. Após a filtragem, os dados perdidos foram imputados usando o algoritmo de imputação *Npute* [29].

Assim, o genótipo *in silico* dos híbridos simples foram inferidos, para cada SNP, com base no genótipo dos genitores. Após o controle de qualidade, um total de 8862 SNPs foram obtidos para os dez cromossomos, com uma média de 68 SNPs por Mb (Figura Suplementar 1).

3.1.Matriz de Relacionamento Genético

A matriz de relacionamento genético entre híbridos foi construída com base nas informações dos SNPs obtidos ao longo do genoma. Assim, as matrizes de relacionamento genômico aditivo (matriz A_g) e de dominância (matriz D_g), foram calculadas de acordo com os métodos descritos em VanRaden et al. [30] e Vitezica et al. [31], respectivamente. Vale ressaltar que ambos os métodos consideram dois alelos para cada loco. As matrizes A_g e D_g foram obtidas através do pacote AGHmatrix [32] disponível no *software* R, e caso as matrizes fossem positivas definidas, as inversas foram obtidas pelo método iterativo descrito por Nazarian & Gezan [33]. Este método modifica os autovalores, negativos ou os positivos que tenham reduzida magnitude, da matriz de relacionamento genético, até que esta passe a ser uma matriz positiva definida com todos os autovalores iguais ou maiores que a constante especificada *a priori* nas análises ($\varepsilon = 0.0001$), como proposto por Jorjani et al. [34].

A fim de determinar o grau de relacionamento genético entre os genitores do programa, dado pela Matriz *Kinship* (K), utilizou-se uma aproximação do método IBS (*Identify-by-state*) [35], calculada com o auxílio do *software* TASSEL [26]. A partir da matriz K calculou-se a medida de dissimilaridade entre os genitores, dado pela distância euclidiana, pelo pacote *Stats* do programa R. Também, usando as medidas de dissimilaridade calculou-se a análise de agrupamento *via* a método UPGMA (*Unweighted pair group method with arithmetic mean*, [36]). O mesmo procedimento foi realizado para construir a matriz de similaridade entre os híbridos fenotipados.

4. Predição Genômica

Primeiramente, foram realizadas análises fenotípicas para os vários anos de cada local, com o objetivo de verificar a qualidade dos dados, estimar o coeficiente de herdabilidade, bem como as médias marginais ajustadas dos genótipos (híbridos) para cada local, via BLUE (do inglês, *Best Linear Unbiased Estimator*). Diferentes esquemas de validação cruzada (VC) foram considerados para a predição de híbridos em múltiplos ambientes (locais), e também, para a predição de híbridos ainda não sintetizados, levando em conta o contexto prático de um programa de melhoramento de sorgo biomassa. Para a predição de híbridos em múltiplos ambientes, as médias ajustadas no primeiro estágio das análises fenotípicas foram utilizadas para o ajuste de modelos de seleção genômica, com a incorporação do efeito da interação entre genótipos e ambientes, e dos efeitos genéticos aditivos e de dominância. Já para a predição de híbridos ainda não sintetizados, as médias ajustadas no primeiro estágio foram utilizadas para o ajuste de um segundo modelo fenotípico para múltiplos locais. Assim, após esses dois estágios de análises fenotípicas, as médias globais ajustadas dos híbridos foram utilizadas para o ajuste de modelos de seleção genômica para a predição de híbridos ainda não realizados, ou seja, ainda não sintetizados e não fenotipados em condições de campo.

4.1.Análises individuais para diferentes locais em diferentes anos

Para cada local, as análises foram realizadas com base no seguinte modelo:

$$\boldsymbol{y} = \boldsymbol{\mu 1_n} + \boldsymbol{X_1 f} + \boldsymbol{X_2 r.f} + \boldsymbol{Z_1 t} + \boldsymbol{Z_2 b.r.f} + \boldsymbol{Z_3 t.f} + \boldsymbol{e} \qquad \textbf{Eq.1}$$

onde $\boldsymbol{y}$ é o vetor ($n \times 1$) de observações fenotípicas dos i híbridos (genótipos, em que i = 1, 2, ..., I), em k blocos (k = 1, 2, ..., K), e j repetições (j = 1, 2, ..., J) em m anos de avaliação (m = 1, 2, ..., M), sendo $n = IJKL$; μ é a média geral; f é o vetor ($m \times 1$) dos efeitos fixos de ano;

$r.f$ é o vetor ($jm \times 1$) dos efeitos fixos de repetição dentro de ano; t é o vetor ($i \times 1$) dos efeitos aleatórios de híbridos, com $\boldsymbol{t \sim N(0, I_i \sigma_t^2)}$, dado que σ_t^2 é a variância de híbridos; $b.r.f$ é o vetor ($jkm \times 1$) de efeitos aleatórios de blocos dentro de repetições dentro de anos, com $\boldsymbol{b \sim N(0, I_{jkm} \sigma_b^2)}$, dado que σ_b^2 é a variância de bloco dentro de repetição dentro de ano; $t.f$ é o vetor ($im \times 1$) dos efeitos aleatórios da interação entre híbridos e anos, com $\boldsymbol{t.f \sim N(0, I_{im} \sigma_{tf}^2)}$, dado que σ_{tf}^2 é a variância da interação entre híbridos e anos; e e é o vetor de resíduos, com $\boldsymbol{e \sim N(0, I_n \sigma_e^2)}$. $\boldsymbol{X_1}(n \times m)$, $\boldsymbol{X_2}(n \times jm)$, $\boldsymbol{Z_1}(n \times i)$, $\boldsymbol{Z_2}(n \times \mathrm{k}jm)$ e $\boldsymbol{Z_3}(n \times im)$ representam as matrizes de incidência para seus respectivos efeitos, $1_n(n \times 1)$ é o vetor de uns, e I_i, I_{jkm}, I_{im} e I_n são matrizes identidade para suas respectivas ordens.

Neste estágio, foram estimados os coeficientes de herdabilidades e os componentes de variância, a fim de verificar a qualidade dos dados fenotípicos. Em seguida os efeitos de híbridos e o efeito da interação entre híbridos e anos foram considerados como fixos para estimar as médias marginais ajustadas dos híbridos para cada local *via* BLUE, e também, a matriz de variância-covariância (VCOV) dos médias ajustadas dos híbridos em cada ambiente, a partir da qual serão calculados os pesos $w = diag(VCOV^{-1})$, com dimensão $i \times l$, que serão utilizados para o efeito residual na próxima etapa de análise [37].

4.2.Validação cruzada para a predição de híbridos em múltiplos ambientes

Diferentes modelos de seleção genômica, que incorporam informações de múltiplos ambientes, foram examinados considerando seis esquemas de validação cruzada (VC1, VC2, VC3.1, VC3.2, VC4.1 e VC4.2), com o objetivo de verificar se a inclusão de informações de ambientes correlacionados pode melhorar as acurácias de predição, como proposto por Burgueño et al. [38]. Para cada esquema de VC, foram amostradas aleatoriamente 100 populações de validação (Figura 1).

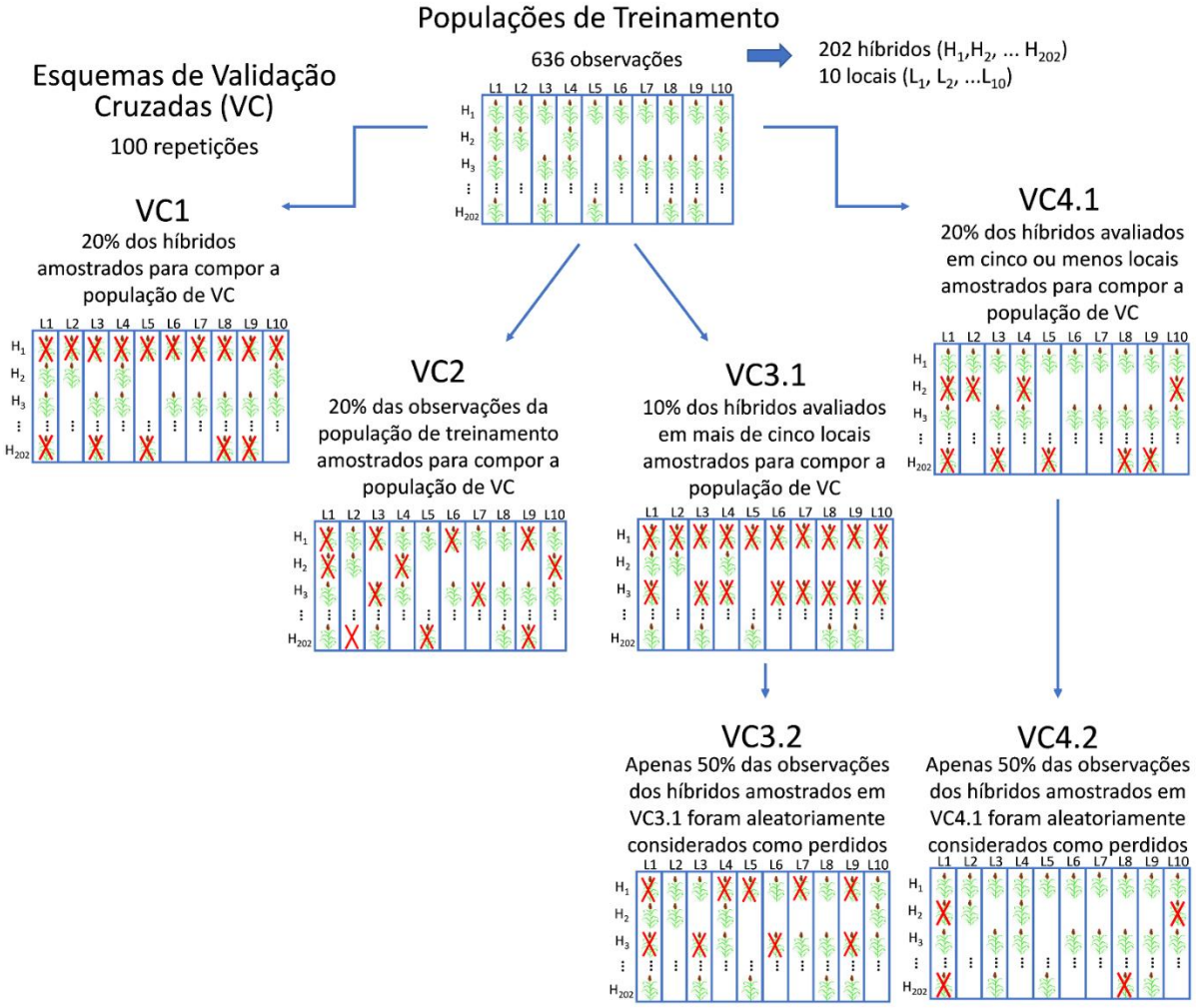


Figura 1. Esquemas de validação cruzada utilizados para avaliar a acurácia de modelos de predição genômica de híbridos de sorgo biomassa em múltiplos ambientes. Indivíduos marcados com "X" em vermelho foram usados para compor a população de validação, e os outros a população de treinamento. Cada esquema de validação foi repetido 100 vezes.

O primeiro esquema de validação (VC1) é o mais tradicional, no qual novos híbridos não foram avaliados em nenhum local. Neste caso, a predição do desempenho de um híbrido ainda não testado é baseada em informações de híbridos relacionados geneticamente, mas sem considerar informações de ambientes correlacionados. Dessa forma, a predição de novos híbridos é baseada apenas em informações fenotípicas e genotípicas de genótipos aparentados. Para esse esquema, 20% dos híbridos (40 genótipos) foram amostrados para compor a população de validação VC1, cujos fenótipos foram considerados como perdidos, e os demais híbridos (80%) foram considerados como população de treinamento. O segundo esquema

(VC2) considera a predição de híbridos que foram avaliados em pelo menos um local. Neste caso, a predição baseia-se no uso de informações de locais correlacionados, além de informações de híbridos aparentados. Assim, a presença de correlação entre os locais avaliados poderia melhorar as acurácias de predição. Neste caso, 20% das observações da população de treinamento (127 observações – Híbrido/Local) foram amostrados para compor a população de validação VC2.

Os esquemas VC3.1 e VC4.1 são decompostos do esquema VC1, dado que os materiais serão preditos para mais de cinco, ou menos de cinco locais, respectivamente. Do mesmo modo que em VC1, a predição do desempenho desses híbridos é baseada apenas em informações de híbridos relacionados, sem levar em conta a informação de ambientes correlacionados, uma vez que os genótipos são considerados como não avaliados em nenhum local. A população de validação VC3.1 foi composta pela amostragem de 10% (5 híbridos) dos híbridos avaliados em mais de cinco locais, 49 dos 202 híbridos fenotipados, e a população de validação VC4.1 pela amostragem de 20% (30 híbridos) dos híbridos avaliados em cinco ou menos locais, 153 do total de 202 híbridos.

Os esquemas VC3.2 e VC4.2 são uma decomposição de VC3.1 e VC4.1, mas a predição do desempenho dos híbridos é baseada tanto em informação de híbridos relacionados quanto em informação de ambientes correlacionados. Para isso, as populações de validação de VC3.2 e VC4.2 foram compostas pela amostragem de 50% das observações das populações de validação em VC3.1 e VC4.1, respectivamente.

A predição dos híbridos em múltiplos ambientes, considerando os esquemas de validação descritos acima, foi realizada com base nas médias marginais ajustadas para cada ambiente, utilizando o vetor de pesos (w) para as variâncias residuais de cada local, obtidos no

primeiro estágio (**Eq. 1**), de acordo com o método GBLUP (do inglês, *Genomic Best Linear Unbiased Prediction*). Para isso, foi considerado o seguinte modelo de seleção genômica:

$$\boldsymbol{y = \mu 1_n + Xs + Z_1 a.t.s + Z_2 d.t.s +} \qquad \textbf{Eq.2}$$

onde y é o vetor ($il \times 1$) de médias marginais ajustadas no primeiro estágio para os i híbridos em cada local l; μ é a média geral; s é o vetor ($l \times 1$) de efeitos fixos de local; $a.t.s$ é o vetor ($il \times 1$) de efeitos genéticos aditivos aleatórios de híbridos dentro de locais, com $\boldsymbol{a.t.s \sim NM(0, A_g \otimes \Sigma_A)}$, dado que $\boldsymbol{A_g}$ é a matriz de relacionamento genético aditivo entre os híbridos; $d.t.s$ é o vetor ($il \times 1$) de efeitos genético de dominância aleatórios de híbridos dentro de locais, com $\boldsymbol{d.t.s \sim NM(0, D_g \otimes \Sigma_D)}$, dado que $\boldsymbol{D_g}$ é a matriz de relacionamento genético de dominância entre os híbridos; e e é o vetor ($il \times 1$) de resíduos, com $\boldsymbol{e \sim NM(0, \Sigma)}$, em que $\boldsymbol{\Sigma}$ corresponde à matriz de VCOV para e, supostamente conhecida da primeira etapa de análises fenotípicas. $\boldsymbol{\Sigma}$ é uma matriz de VCOV diagonal em que cada local tem um componente de variância específico e independente. $\boldsymbol{\Sigma_A}$ e $\boldsymbol{\Sigma_D}$ são as matrizes de VCOV para os efeitos genéticos aditivos e de dominância de híbridos dentro de local com dimensão $il \times il$. $\boldsymbol{X}$ ($il \times l$), $\boldsymbol{Z_1}$ ($il \times il$) e $\boldsymbol{Z_2}$ ($il \times il$) são as matrizes de incidência para os seus respectivos efeitos fixos (X) e aleatórios (Z).

Também foram estimados os componentes de herdabilidades no sentido restrito (h^2) e a proporção da variância explicada pelos efeitos de dominância (d^2), de acordo com as equações:

$$\boldsymbol{h^2 = \sigma_a^2 / \sigma_a^2 + \sigma_d^2 + \sigma_e^2} \qquad \boldsymbol{e} \qquad \boldsymbol{d^2 = \sigma_d^2 / \sigma_a^2 + \sigma_d^2 + \sigma_e^2}$$

onde σ_a^2 é o componente de variância aditiva, σ_d^2 é o componente de variância dominante, e σ_e^2 é o componente de variância do erro. A herdabilidade no sentido amplo (H^2) foi calculada pela somatória entre estes dois parâmetros ($h^2 + d^2$).

O modelo apresentado na equação **Eq.2** corresponde ao modelo AD, que considera tanto efeitos genéticos aditivos quanto de dominância para a predição da performance dos híbridos em múltiplos ambientes. Para fins comparativos, também foi ajustado um modelo alternativo, considerando apenas efeitos aditivos (Modelo A). Além disso, diferentes estruturas de variância foram examinadas para modelar as matrizes $\boldsymbol{\Sigma_A}$ e $\boldsymbol{\Sigma_D}$, e a melhor estrutura foi selecionada com base nos critérios de informação de Akaike (AIC) [39] e Bayesiano (BIC) [40].

As acurácias de predição foram estimadas como a correlação de Pearson entre as médias marginais ajustadas para cada local e as médias preditas a partir dos modelos de seleção genômica A ou AD, considerando os diferentes esquemas de VC.

4.3.Validação cruzada para a predição de híbridos ainda não sintetizados

A síntese de híbridos em programas de melhoramento de sorgo biomassa é baseada em três tipos de linhagens: macho-estéreis (linhagens A), usadas como genitor feminino; as linhagens mantenedoras (B), macho-férteis, usadas para manter as linhagens A; e as linhagens restauradoras (R), macho-férteis, usadas como genitor masculino na obtenção de híbridos. As linhagens A e B são isogênicas, e se diferem apenas pelo citoplasma, dado que nas linhagens A são macho-estéreis e nas linhagens B são macho-férteis (Figura 2).

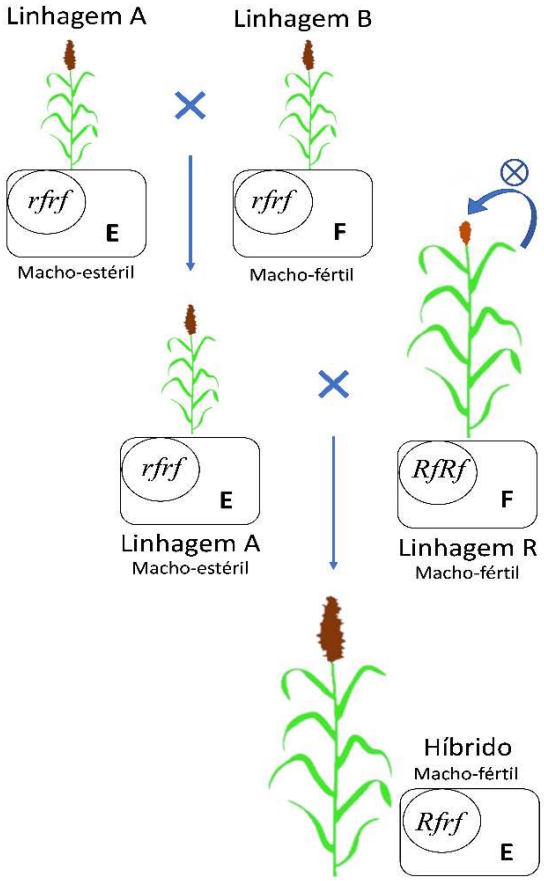


Figura 2. Esquema de manutenção de linhagens e obtenção de híbridos em programas de melhoramento de sorgo biomassa.

Dessa forma, foram propostos três esquemas de VC para predizer híbridos ainda não sintetizados (T0, T1, T2). Para compor as populações de validação (PV), primeiramente foram selecionadas de forma aleatória linhagens parentais (circuladas em vermelho), e posteriormente os híbridos a serem preditos (marcados com um "X" em vermelho). Para cada esquema de validação, foram aleatoriamente amostradas 100 populações de validação (Figura 3).

No esquema T0, a predição de híbridos não avaliados não conta com informação de parentesco direto, ou seja, não se tem informação de híbridos meios-irmãos por parte do genitor masculino e nem por parte do genitor feminino na população de treinamento. Neste caso, a predição de novos híbridos é baseada em informações genotípicas de indivíduos com parentesco indireto. Dessa forma, selecionou-se aleatoriamente genitores masculinos (linhagens R) e femininos (linhagens A), e amostrou-se todos os híbridos derivados destas linhagens para

compor a população de validação T0, cujas médias globais ajustadas foram consideradas como perdidas (não observadas). Com isso, não se tinha híbridos com parentesco direto, ou seja, híbridos meios-irmãos nas populações de treinamento.

No esquema T1, a predição de híbridos ainda não sintetizados conta com a informação de meios-irmãos derivados de uma das duas linhagens genitoras. Assim, na população de treinamento há informação de híbridos meios-irmãos, derivados dos genitores masculinos para a predição da população de validação T1F, ou seja, novos híbridos derivados de fêmeas ainda não avaliadas em outros cruzamentos. Ou então, informação de híbridos meios-irmãos derivados dos genitores femininos para a predição da população de validação T1M, composta por novos híbridos, ou seja, derivados de linhagens machos-férteis ainda não avaliadas em outros cruzamentos.

No esquema T2, a predição de híbridos não sintetizados conta com a informação de híbridos meios-irmãos para ambos os genitores, ou seja, ambos os pais já foram avaliados em outros cruzamentos na população de treinamento. Para isso, selecionou-se aleatoriamente as linhagens genitoras, macho-estéreis e macho férteis, que apresentavam mais de dois híbridos como descendentes, e amostrou-se 10%, 20% e 30% dos híbridos derivados dessas linhagens para compor a população de validação. Vale ressaltar que em todas as populações de treinamento manteve-se pelo menos um híbrido de cada uma das linhagens parentais da população de validação.

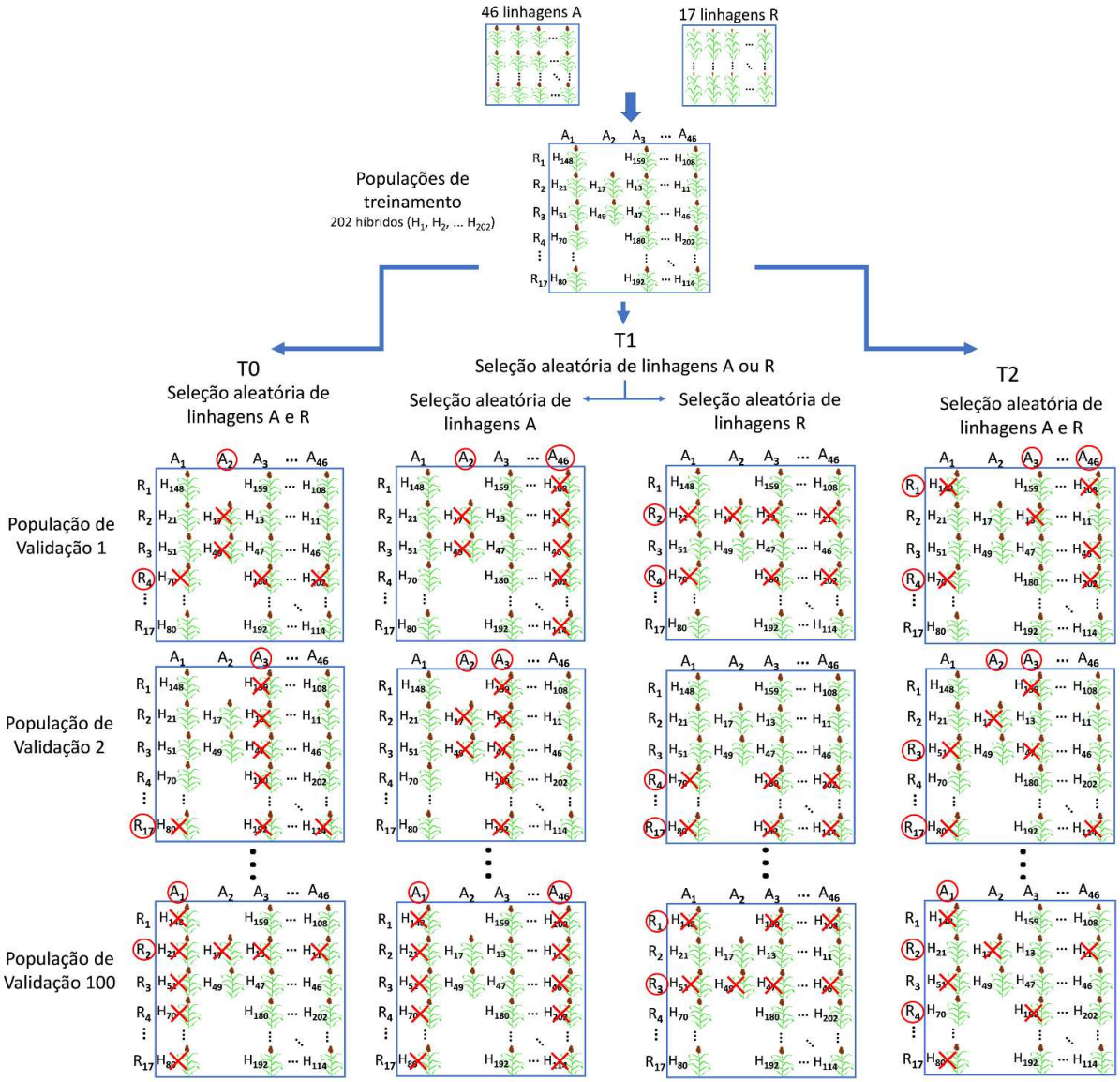


Figura 3. Esquemas de validação cruzada para avaliar a acurácia dos modelos de predição genômica de híbridos de sorgo biomassa ainda não sintetizados. Linhagens parentais circuladas em vermelho foram selecionadas para amostrar os híbridos a serem preditos (população de validação), marcados com "X" em vermelho. Os outros híbridos não marcados foram usados para compor as populações de treinamento. T0: esquema de validação cruzada para a predição de híbridos cujos genitores ainda não foram avaliados em outras combinações híbridas. T1: esquema de validação cruzada em que apenas um dos genitores já foi avaliado em outras combinações híbridas. T2: esquema de validação cruzada em que ambos os genitores já foram avaliados em outras combinações híbridas.

As análises estatísticas para os esquemas de VC para a predição de híbridos ainda não sintetizados foram realizados em três estágios. No primeiro estágio, foram obtidas as médias marginais ajustadas para cada local (**Eq. 1**). No segundo estágio, com base nas médias ajustadas dos híbridos em cada local e no vetor de pesos associados à variância residual de cada local (w) obtidos no primeiro estágio, foi ajustado o seguinte modelo:

$$\boldsymbol{y = \mu 1_n + Xt + Z_1 s + Z_2 t.s + e} \qquad \textbf{Eq.3}$$

onde y é o vetor ($il \times 1$) de médias ajustadas dos i híbridos; μ é a média geral; t é o vetor ($i \times 1$) de efeitos fixo de híbridos; s é o vetor ($l \times 1$) de efeitos aleatórios de local, com $\boldsymbol{s \sim N(0, I_l \sigma_s^2)}$, dado que σ_s^2 é a variância de local; $t.s$ é o vetor ($il \times 1$) de efeitos aleatórios da interação entre híbridos e locais, com $\boldsymbol{t.s \sim NM(0, \sigma_{ts}^2)}$; e e é o vetor ($il \times 1$) de resíduos, com $\boldsymbol{e \sim NM(0, \Sigma)}$. $\boldsymbol{X}$ ($il \times i$), $\boldsymbol{Z_1}$($il \times l$), $\boldsymbol{Z_2}$($il \times il$) são as matrizes de incidência para os seus respectivos efeitos, e $\boldsymbol{I_l}$ são matrizes identidade com suas correspondentes ordens. No modelo apresentado na **Eq.3**, o efeito de híbridos

No terceiro estágio, foram ajustados os modelos de seleção genômica, com a inclusão de efeitos genéticos aditivos e de dominância, utilizando as médias globais ajustadas para cada híbrido e o vetor de pesos associados à variância residual (w) obtidos no segundo estágio (**Eq. 3**). Para isso, foi utilizado o seguinte modelo:

$$\boldsymbol{y = \mu 1_n + Z_1 a.t + Z_2 d.t + e} \qquad \textbf{Eq.4}$$

onde y é o vetor ($il \times 1$) de médias globais ajustadas dos i híbridos; μ é a média geral; $a.t$ é o vetor ($i \times 1$) de efeitos aditivos aleatórios de híbridos, com $\boldsymbol{a.t \sim NM(0, A_g)}$; $d.t$ é o vetor ($i \times 1$) de efeitos de dominância aleatórios de híbridos, com $\boldsymbol{d.t \sim NM(0, D_g)}$; e e é o vetor ($il \times 1$) de resíduos, com $\boldsymbol{e \sim NM(0, \Sigma)}$. $\boldsymbol{A_g}$ e $\boldsymbol{D_g}$ são as matrizes relacionamento genético aditivos [30] e de dominância [31], respectivamente, com dimensão $i \times i$, que foram calculadas utilizando o pacote AGHmatrix [32] disponível no software R [41]. $\boldsymbol{Z_1}$($i \times i$) e $\boldsymbol{Z_2}$ ($i \times i$) são as matrizes de incidência para os seus respectivos efeitos.

O modelo apresentado na equação **Eq.4** corresponde ao modelo AD, que considera tanto efeitos genéticos aditivos quanto de dominantes para a predição da performance de híbridos ainda não realizados. Para fins comparativos, também foi ajustado um modelo alternativo,

considerando apenas efeitos aditivos (Modelo A). As acurácias de predição foram estimadas como a correlação de Pearson entre as médias globais ajustadas e as médias preditas a partir dos modelos de seleção genômica A ou AD, considerando os diferentes esquemas de VC (T0, T1 e T2).

Todas as análises estatísticas foram realizadas usando o pacote ASReml-R [42], disponível no software R [43].

Resultados

1. Matriz de relacionamento genético

Os dados genômicos das linhagens genitoras, após imputação e filtragem, foram usados para inferir o genótipo dos híbridos fenotipados nos ensaios de valor de cultivo e uso (VCU) e de avaliação preliminar de híbridos (APH). Na Figura 4 (A), está representada a matriz de relacionamento genético das linhagens genitoras do programa de melhoramento de sorgo da Embrapa Milho e Sorgo (Matriz de *Kinship*, K). O coeficiente de relacionamento genético entre as linhagens de sorgo biomassa do programa de melhoramento da Embrapa variou de 0,40 a 0,60. Com base no padrão de agrupamento das linhagens, foi possível a identificação de dois grupos de genitores, indicados pelas barras (azul e roxa) abaixo do dendograma, que correspondem aos grupos das 17 linhagens R e das 46 linhagens A, respectivamente. Além disso, quatro linhagens A (LinA_24, LinA_1, LinA_34 e LinA_45) foram alocadas no grupo das linhagens R, destacadas com a cor roxa na barra horizontal azul do grupo de linhagens R.

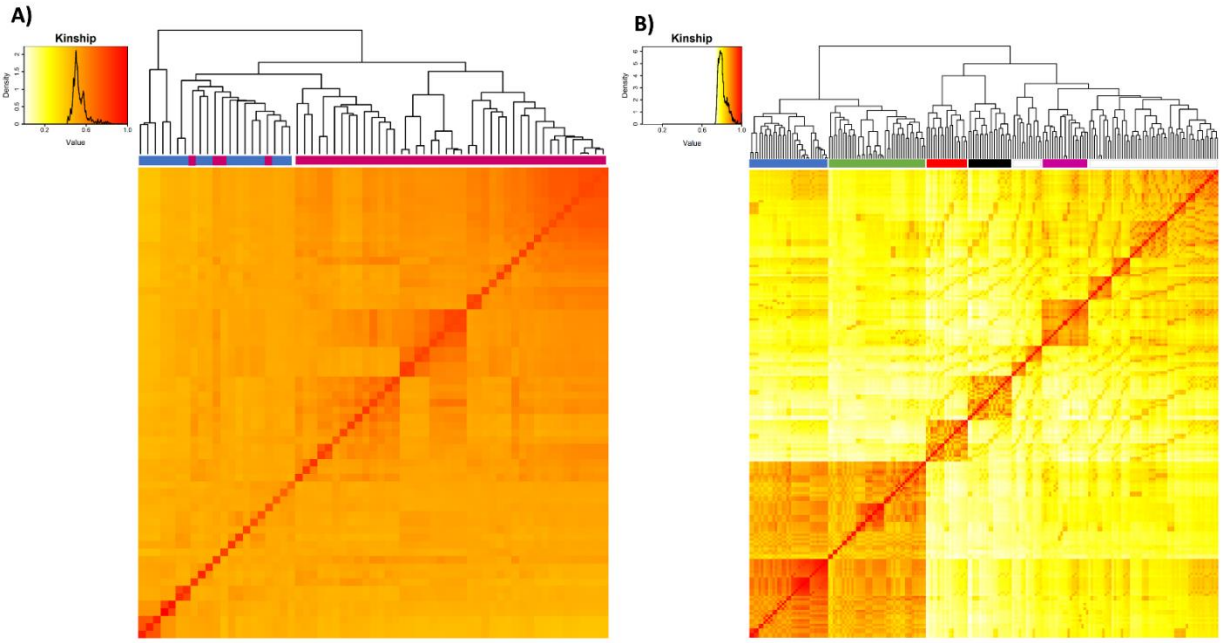


Figura 4. A) Matriz de relacionamento genético entre 67 linhagens genitoras de sorgo biomassa. B) Matriz de relacionamento genético entre 202 híbridos de sorgo biomassa. *Heatmap* indicando a magnitude dos coeficientes de relacionamento estimados baseado em uma aproximação do método IBS (*Identify-by-state*). Acima do *heatmap*: dendograma resultante de uma análise de agrupamento (UPGMA) com base em distâncias euclidianas, e uma atribuição de cores para os grupos formados em cada matriz.

Os coeficientes de similaridade entre os híbridos (Figura 4 B) variaram de 0,70 a 0,90. E assim como nos genitores há a formação de grupos de híbridos mais semelhantes entre si, com a identificação de cinco possíveis grupos de diversidade, que correspondem as barras horizontais com cores diferentes. O grupo I (barra azul) foi formado por 34 híbridos, o grupo II (barra verde) por 42, o grupo III (barra vermelha) por 18, o grupo IV (barra preta) por 19, e o grupo V (barra roxa) por 20 híbridos. Observa-se que a formação dos grupos a partir da matriz de similaridade entre os híbridos ocorreu predominantemente de acordo com as similaridades genéticas entre seus genitores masculinos (linhagens R). Por exemplo, o grupo representado pela barra vermelha, agrupou todos os híbridos descendentes das LinR_14 e LinR_15, o grupo preto agrupou todos os híbridos descendentes das linhagens LinR_5, LinR_6 e LinR_17, o grupo roxo agrupou todos os híbridos descendentes da linhagem LinR_4, e os grupos verde e azul, agruparam os híbridos descendentes tanto da linhagem LinR_2 quanto da linhagem LinR_3. Nota-se uma maior similaridade genética entre os pares de genitores dos híbridos que

foram agrupados em um mesmo grupo de diversidade, como são os casos das linhagens LinR_2 e LinR_3, das linhagens LinR_5, LinR_6 e LinR_17, e das linhagens LinR_14 e LinR_15 (Figura suplementar 2).

2. Validação cruzada para a predição de híbridos em múltiplos ambientes

Diferentes estruturas de variância-covariância (VCOV) para os efeitos genéticos de híbridos foram examinadas para a predição da performance de híbridos em múltiplos ambientes, como: identidade (*id*), diagonal (*diag*), correlação uniforme (*corv*) e fator analítico de primeira ordem (*FA1*). Tais estruturas foram comparadas com base nos valores de AIC, BIC e logREML (Tabela 4). O modelo com matriz *corv* apresentou o menor valor de BIC, e também, não apresentou problemas de convergência para a estimativa dos parâmetros genético-estatísticos (por exemplo, variâncias fixadas em zero), quando comparado ao modelo FA1, cujos valores de AIC e logREML foram menores que os do modelo *corv*. Dessa forma, o modelo com a estrutura *corv* foi selecionado para dar continuidade às análises de seleção genômica.

Tabela 4. Medidas de ajuste, AIC, BIC e logREML, dos modelos testados nas análises em multi-ambientes (Eq. 2) incluindo efeitos Aditivo-Dominância (Modelo AD) e Aditivos (Modelo A), e as diferentes estruturas de variância *id*, *diag*, *corv* e *FA1*. E estimativa de parâmetros genéticos usando a matriz *corv*.

Estruturas de Variância*	Modelo A			Modelo AD		
	AIC	BIC	logREML	AIC	BIC	logREML
Id	3431,19	3435,63	-1714,59	3401,28	3410,15	-1698,64
Diag	3403,40	3447,79	-1691,70	3379,50	3468,28	-1669,75
Corv	**3356,23**	**3365,11**	**-1676,12**	**3371,52**	**3389,28**	**-1681,76**
FA1	**3326,35**	3415,13	**-1643,17**	**3308,37**	3485,94	**-1614,18**
Parâmetros	h^2	H^2	d^2	h^2	H^2	d^2
	0.64 ± 0.05	-	-	0.58 ± 0.06	0.71 ± 0.12	0.13 ± 0.06

Em negrito, estão destacados os menores valores de AIC, BIC e logREML. Em negrito e sublinhado, estão destacados os modelos selecionados para a efeito genético de híbridos nos modelos A e AD.* Id: identidade; Diag: diagonal; Corv: correlação uniforme; FA1: fator analítico de ordem 1. h^2 é a herdabilidade no sentido restrito, H^2 é a herdabilidade no sentido amplo, e d^2 é a proporção de variância genética explicada pelos efeitos de dominância.

Dado o melhor modelo de predição, foram estimados os parâmetros genéticos de herdabilidade no sentido restrito (h^2) e amplo (H^2), e a proporção de variância genética explicada pelos efeitos de dominância (d^2) (Tabela 4). As herdabilidades foram consideradas moderadas, tanto no sentido restrito quanto no sentido amplo, dado que a produção de massa verde em sorgo é uma variável quantitativa.

As acurácias de predição dos híbridos não testados variaram de acordo com o esquema de VC adotado e com o tipo de efeitos genéticos incorporados nos modelos de seleção genômica, ou seja, incluindo apenas efeitos genéticos aditivos (modelo A) ou efeitos genéticos aditivos e de dominância (modelo AD) (Figura 5). Nota-se que a inclusão dos efeitos de dominância (modelo AD) proporcionou um pequeno aumento da média das acurácias predição (Tabela Suplementar 1), quando comparado aos modelos que utilizam apenas informações dos efeitos genéticos matriz aditivos. Na Figura 5, as letras minúsculas comparam estatisticamente os modelos A e AD, dentro de cada um dos esquemas, podendo-se observar que os modelos AD foram estatisticamente superiores aos modelos A para quase todos os esquemas de VC examinados, exceto para VC2 e VC3.1.

Com relação aos diferentes esquemas de VC examinados para a predição de híbridos em múltiplos ambientes, VC3.1 (A = 0,26 e AD = 0,28) e VC3.2 (A = 0,35 e AD = 0,38) foram os que apresentaram as menores médias de acurácia de predição, provavelmente o menor tamanho efetivo da população de validação, possa ter gerado baixos valores de acurácia de predição nestes esquemas e também devido à menor relação entre a população de treinamento e a população de validação efetiva, uma vez que a população de validação foi restringida para a inclusão de apenas híbridos avaliados em pelo menor cinco locais (Figura 5). Em contrapartida, os esquemas VC4.1 (A = 0,72 e AD = 0,75) e VC4.2 (A = 0,73 e AD = 0,77), que consideraram a predição de híbridos em cinco ou menos locais, foram os que apresentaram maiores acurácia médias de predição, provavelmente devido ao maior relacionamento entre a

população de treinamento e a população de validação, uma vez que foram incluídos um maior número de híbridos na população de validação, o que consequentemente aumentou o número de indivíduos aparentados e locais correlacionados entre essas populações. Nota-se que o tamanho da população de treinamento dos esquemas VC3.1 e VC4.1, e VC3.2 e VC4.2 foram similares, o que reforça a importância de manter informações de relacionamento tanto genético quanto entre ambientes que permitam a troca de informações entre as populações de treinamento e validação, resultando em maiores acurácias de predição.

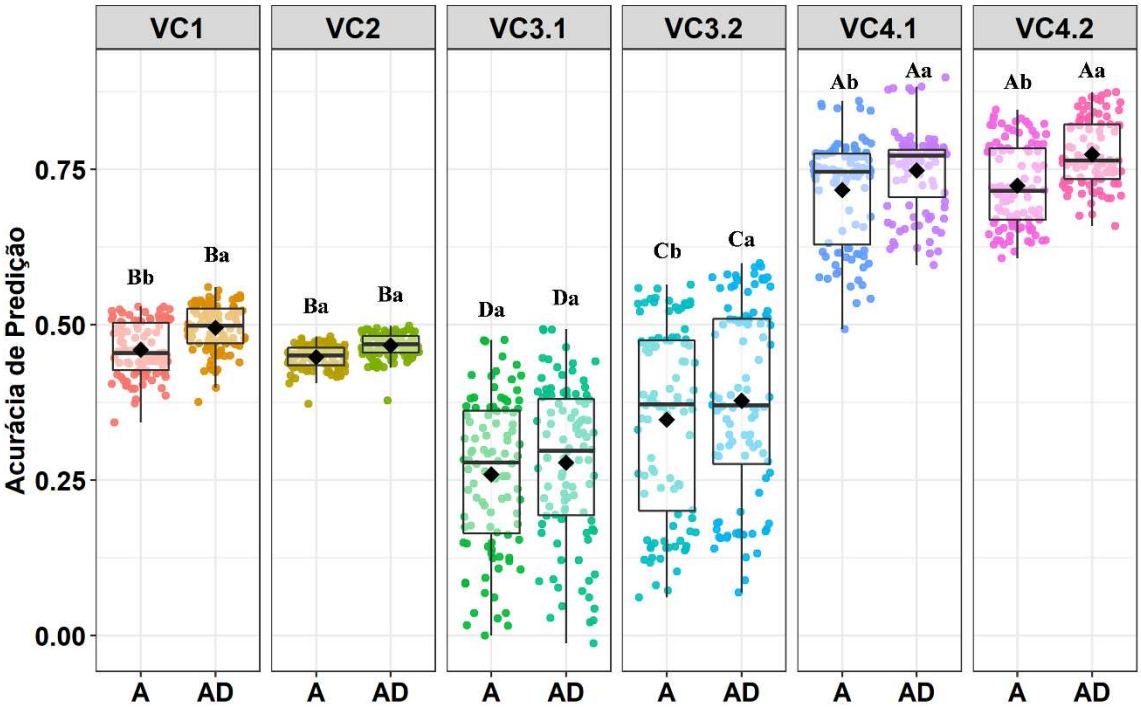


Figura 5. Acurácias de predição para os modelos aditivo (A) e aditivo-dominante (AD) para a característica produção de massa verde em híbridos de sorgo biomassa, considerando seis tipos de esquema de validação cruzada. VC1: 20% dos híbridos amostrados para compor a população de VC; VC2: 20% das observações da população de treinamento amostrados para compor a população de VC; VC3.1: 10% dos híbridos avaliados em mais de cinco locais para compor a população de VC; VC3.2: apenas 50% das observações dos híbridos amostrados em VC3.1 foram aleatoriamente considerados como perdidos. VC4.1: 20 % dos híbridos avaliados em cinco ou menos locais amostrados para compor a população de VC; VC4.2: Apenas 50% das observações dos híbridos amostrados em VC4.1 foram aleatoriamente considerados como perdidos. Letras acima dos boxplots correspondem aos resultados do teste de comparação de médias *via* Skott Knott, considerando o nível de significância de 5%. Médias seguidas pelas mesmas letras maiúsculas ou minúsculas não diferem estatisticamente entre si levando em conta os diferentes esquemas ou modelos (A/AD) examinados, respectivamente.

Observa-se também que os esquemas VC1 e VC2, realizados para a predição de 40 novos híbridos e 127 novas observações de híbridos dentro de locais, respectivamente, apresentaram uma acurácia média de predição estatisticamente mais baixa que os esquemas VC4.1 e VC4.2, variando entre 0,45 e 0,50 (Figura 5). Porém, VC1 e VC2 foram os esquemas que apresentaram menor amplitude de variação das acurácias de predição, ou seja, geraram modelos de predição com maior repetibilidade de resultados, independente das amostras utilizadas para a população de validação. No entanto, apesar de apresentarem menores desvios, os seus maiores valores não superam as predições em VC4.1 e VC4.2.

3. Validação cruzada para a predição de híbridos ainda não sintetizados

Nos esquemas de VC para a predição de híbridos ainda não sintetizados, observou-se diferença nas médias das acurácias de predição tanto entre os esquemas propostos (T0, T1 e T2), quanto entre os critérios usados para a amostragem das populações de validação (Figura 6). Também observou diferença nas médias das acurácias de predição para os modelos com ou sem informação das matrizes de dominância, modelos AD e A, respectivamente. Assim, como nos esquemas de VC para múltiplos ambientes, as acurácias de predição de novos híbridos foram maiores para os modelos AD quando comparados aos modelos A (Figura 6), exceto nos esquemas T1F amostrando duas e três fêmeas, que não apresentaram diferença estatística entre os modelos A e AD. Nota-se também que, em alguns casos, o modelo AD apresentou um aumento de até 10% na média da acurácia de predição, como nos esquemas T0-4F4M, T1M-3Machos e T1M-4Machos (Tabela suplementar 2).

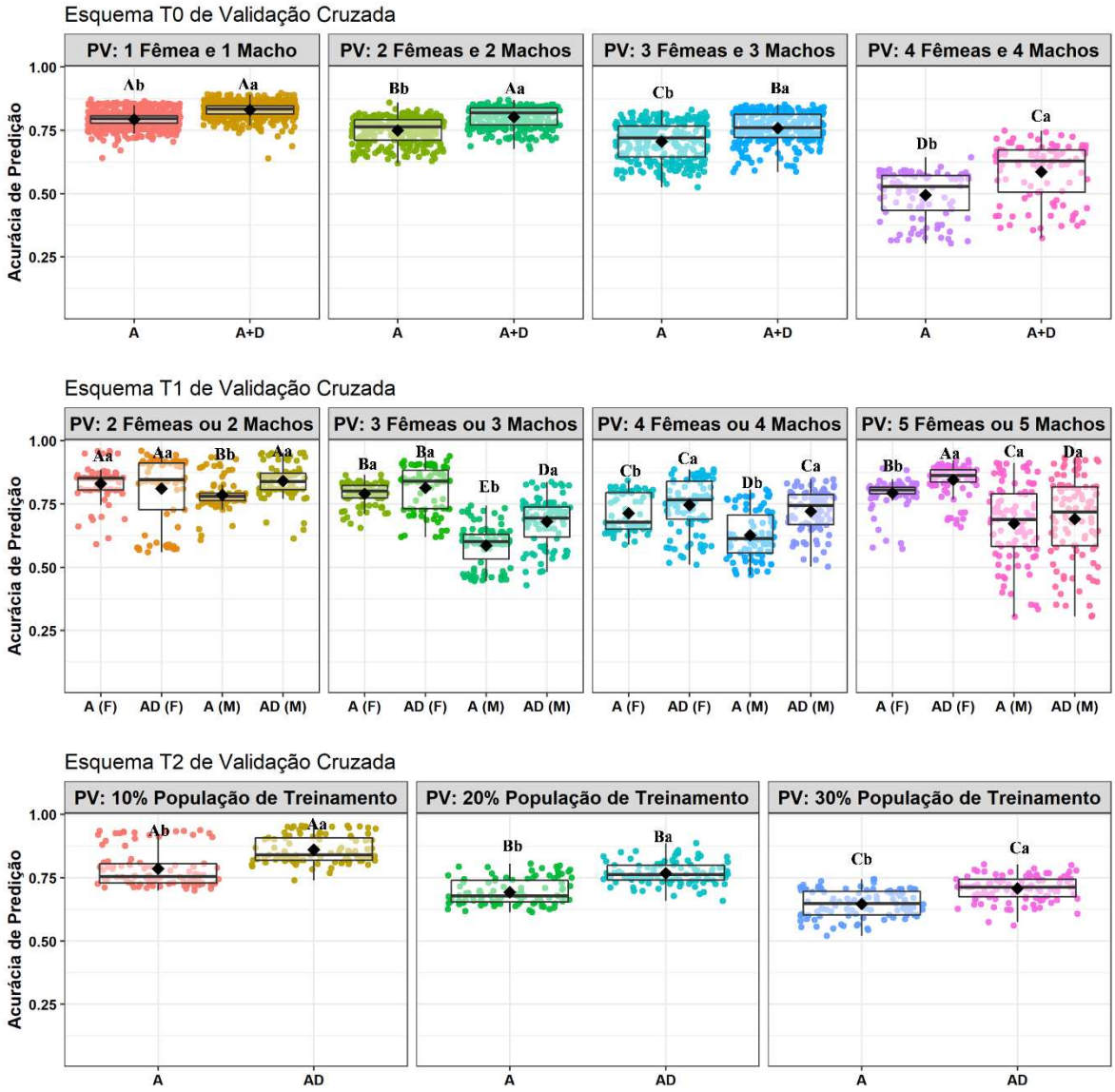


Figura 6. Acurácias de predição estimadas para modelos com efeitos aditivos (A) e efeitos aditivo-dominante (AD) para produção de massa verde em sorgo biomassa (tha^{-1}), considerando três esquemas distintos de validação cruzada (T0, T1 e T2). T0: nenhum dos genitores dos híbridos da população de validação foram avaliados em outros cruzamentos na população de treinamento. T1: população de treinamento composta por híbridos com um dos genitores em comum com os híbridos da população de validação. T2: genitores dos híbridos da população de validação já foram avaliados em pelo menos um cruzamento da população de treinamento. Letras acima dos boxplots correspondem aos resultados do teste de comparação de médias via Skott Knott, considerando nível de significância igual a 5%. Médias seguidas pelas mesmas letras maiúsculas ou minúsculas não diferem estatisticamente entre si levando em conta os diferentes esquemas ou modelos (A/AD) examinados, respectivamente.

No primeiro esquema T0, no qual são realizadas predições de novos híbridos cujos genitores ainda não foram avaliados em nenhum cruzamento da população de treinamento, as médias das acurácias de predição foram maiores e apresentaram menor variabilidade (menor desvio padrão) nos casos de esquemas com amostragem de um menor número de genitores

(Tabela Suplementar 2). Além disso, observa-se que no esquema T0 que a predição dos híbridos de dois genitores (PV: 1 Fêmea e 1 Macho), foram os que apresentaram a maior média das acurácias de predição (Figura 6), devido ao menor tamanho da população de validação quando comparado aos outros esquemas de validação cruzada. O número médio de híbridos por genitor no esquema T0:1Fêmea e 1Macho foi de 1,3 para fêmeas e 3,1 para machos, já para o esquema T0:4Fêmeas e 4Machos foi 2,5 para fêmeas e 5,8 para machos. Ou seja, à medida que houve um aumento no número de genitores cujos híbridos foram amostrados para compor a população de validação, aumentou-se o número de híbridos não sintetizados a serem preditos, o que consequentemente diminuiu o número de informações com maior relacionamento genético entre a população de treinamento e a população de validação, levando a uma menor acurácia de predição.

No esquema T1, realizou-se a predição para os dois grupos de novos híbridos, híbridos sem informação de meios-irmãos maternos (T1F) e híbridos sem informação de meios-irmãos paternos (T1M), ou seja, na população de treinamento não há informação de descendentes que tenham genitores A ou R em comum com híbridos amostrados para compor a população de validação, respectivamente. Dessa forma, neste esquema, fez-se a predição de novos híbridos considerando que apenas um dos genitores participou de cruzamentos para a obtenção de híbridos presentes na população de treinamento. Observou-se que, para esse esquema, a predição de novos híbridos foi mais acurada quando não se tem híbridos na população de treinamento que tenham linhagem materna em comum com os híbridos da população de validação. Por exemplo, quando foram preditos novos híbridos para três linhagens fêmeas distintas (T1F - PV: 3 Fêmeas), a acurácia foi em média 20% maior do que quando foram preditos novos híbridos derivados de três machos (T1M - PV: 3 Machos), dado o modelo usando apenas informação de efeitos aditivos (Modelo A, Figura 6), respectivamente. Vale ressaltar que o número médio de híbridos por genitor na população de validação não alterou entre os

esquemas, o que mostra que, em geral o relacionamento genético entre as populações de treinamento e validação é a principal causa de variações na acurácia de predição dos diferentes esquemas de VC.

Para o esquema T2, que é baseado na predição de novos híbridos, considerando dois genitores já avaliados em outros cruzamentos que compõem a população de treinamento, observou-se um ligeiro aumento na acurácia de predição quando se diminuiu a população de validação, ou seja, diminuiu o número de novos híbridos a serem preditos. Pode-se observar na Figura 6 que em média a acurácia de predição foi maior estatisticamente na população de validação que amostrou-se 10% dos híbridos da população de treinamento (letra "a" maiúscula), do que a que amostrou-se 20% (letra "b" maiúscula) e do que a que mostrou-se 30% (letra "c" maiúscula). Na Tabela Suplementar 2, nota-se que para a população de validação de 10% a acurácia média de predição foi 15% maior do que para a população de validação de 30%. O número de médio de híbridos por genitor na população de validação entre os esquemas variou de 1,6 (PV:10%) a 2,9 (PV:30%), mostrando que quanto maior o número de híbridos, menor o número de informações relacionadas com a população de treinamento, refletindo em uma menor acurácia de predição.

Discussão

Neste estudo, foram realizadas análises em dois estágios para a predição genômica de híbridos de sorgo biomassa em múltiplos ambientes, e em três estágios para a predição de híbridos ainda não sintetizados. A acurácia de predição variou de acordo com os esquemas de validação propostos, sendo mais acurado em populações de treinamento com maior relacionamento com as populações de validação, com observações de híbridos aparentados, e

com a inserção da matriz de dominância no modelo de predição. Além disso, a seleção genômica possibilitou identificar genitores do programa com potencial em cruzamentos futuros.

1. Inclusão de efeitos genômicos

Estudos de seleção genômica são recentes e escassos quando se trata da cultura do sorgo biomassa, e a busca por altas acurácias de predição de híbridos não testados são de extrema valia para os programas de melhoramento. Uma das grandes dificuldades encontradas pelos melhoristas de sorgo em estágios iniciais do programa, é a ampla gama de híbridos possíveis de serem sintetizados dado um pequeno grupo de genitores. Uma vez que o número de híbridos é o produto entre o número de genitores macho-estéreis (linhagens A) e o número de genitores macho-férteis (linhagens R), disponíveis no programa de melhoramento. Portanto, o desafio enfrentado pelos melhoristas é encontrar combinações promissoras entre muitos possíveis híbridos [45,46]. Dessa forma, a seleção genômica possibilitará a redução do número de genótipos inferiores avaliados em campo e o tempo para seleção de materiais superiores, o que consequentemente, pode gerar aumento dos ganhos genético por unidade de tempo [47–49].

Outro entrave encontrado pelos melhoristas nos estágios iniciais do programa é o planejamento de síntese e seleção de novas linhagens como genitoras elites, que segundo Hunt [4] pode levar até 12 anos, o que torna os processos morosos e de alto custo. Neste caso, a seleção genômica também pode auxiliar através da identificação de possíveis linhagens elite e determinação dos melhores cruzamentos a serem avaliados em campo.

Vários autores têm mostrado o aumento da acurácia de predição com a inserção da matriz de dominância aos modelos de seleção genômica, como nas culturas do milho [8,50], feijão [51], eucalipto [52], e batata [53]. Nossos resultados também mostraram superioridade dos modelos com adição da matriz de dominância para a maioria dos esquemas de predição

testados. Santos et al. [50], em trabalhos com milho, relataram que a inclusão da dominância aumentou a acurácia de predição em cerca de 20%, indicando sua importância para prever genótipos não testados. Dessa forma, a não distinção dos componentes de variância em trabalhos de seleção genômica para produção de massa verde em sorgo biomassa podem levar a decisões errôneas na seleção materiais mais promissores pelos melhoristas.

Outros estudos relatam o aumento da eficiência preditiva dos modelos com a adição da matriz de dominância, principalmente para características de menor herdabilidade [50,54], como é o caso da produção de matéria verde em sorgo biomassa, permitindo a exploração da heterose sem alterar drasticamente o tamanho do modelo a ser predito [51].

2. Modelos de seleção genômica usando dados de múltiplos ambientes

Os programas de melhoramento necessitam de avaliações em vários anos e locais, o que torna as análises de dados mais complexas devido a possibilidade de interação dos genótipos com os diferentes ambientes em avaliação. A seleção genômica com a incorporação da interação G × A têm se mostrado vantajosa em estudos com características quantitativas, uma vez que são altamente influenciadas por fatores ambientais [8,18,23].

Como mostrado na Figura 8, as observações com as maiores médias ajustadas *via* BLUE no primeiro estágio, sem a interpretação das interações G × A, mostram-se altamente produtivos. Quando comparados as médias preditas *via* BLUP no segundo estágio, com a interpretação das interações, os genótipos se comportaram de maneira totalmente distinta, podendo em alguns casos ficarem abaixo da média, como é o caso das observações H_016Dou, H_017Dou, H_047Dou, H_049Dou, H_055Dou, H_063SV, H_066Dou, H_067Dou, H_210NP e H_212Dou. Dessa forma, modelos que não consideram a presença e estudo das interações

genótipo por ambiente, podem levar a interpretações errôneas e tornar o modelo de seleção genômica menos preciso [55].

Nos esquemas de validação também foi possível verificar a importância do estudo das interações. Como por exemplo, nos esquemas VC3.1 e VC3.2, onde se realizou a predição de cinco novos híbridos para mais de cinco locais, o número de informação por local presentes nas populações de treinamento foi reduzido, o que interferiu diretamente no poder de predição dos modelos, devido ao menor relacionamento entre a população de treinamento e a população de validação.

Outro fator que influência a redução da acurácia de predição é a herdabilidade das características em estudo. Segundo Combs e Bernardo [56] em estudos de GS é esperado que a acurácia de predição seja menor em características com menor herdabilidade. Em nosso estudo as herdabilidades variaram entre 0,58 e 0,64, obtendo acurácias médias de predição igual a 0,50, variando de 0,26 a 0,77, para ensaios em múltiplos ambientes. Em estudos com sorgo biomassa para produção de etanol de segunda geração, Fernandes [17] obtiveram herdabilidades iguais a 0,26 e acurácia média de predição igual a 0,40. Dado que neste estudo os autores não consideraram a presença de interação genótipo por ambiente, podendo-se notar a inferioridade na predição de híbridos não avaliados, e a importância das interações G × A para a cultura do sorgo biomassa.

3. Acurácia de predição nos diferentes esquemas de validação cruzada

A VC tem se mostrado uma ferramenta útil para avaliar os modelos preditivos [23], e a adição de efeitos de dominância, em alguns casos tem levado ao aumento da capacidade preditiva dos modelos e da qualidade da decomposição dos componentes de variância [50]. No entanto, alguns trabalhos consideram o tamanho e representatividade da população de

treinamento, a herdabilidade da variável em análise, o grau de parentesco entre os indivíduos preditos e a população de treinamento, assim como a consanguinidade entre as linhagens usadas como genitoras, como alguns dos fatores que mais influenciam a acurácia de predição nos modelos de seleção genômica [4,23].

Observou-se em nosso estudo que o principal fator causador da amplitude de variação das acurácias de predição entre os esquemas avaliados foi o tamanho da população de treinamento, e o grau de parentesco e de informatividade entre as populações de treinamento e a população de validação. O que mostra a importância da estruturação bem realizada das populações de treinamento, focando na preservação de informação de genótipos relacionados geneticamente aos novos híbridos a serem preditos. Observa-se também que a manutenção de informações ambientais, ou seja, observações de genótipos nos diferentes ambientes, proporciona a captura das interações dos genótipos com os diferentes locais, e interpretação com maior acurácia de híbridos não avaliados, como observado em VC4.1 e VC4.2.

Nota-se também, que os esquemas de validação que apresentaram menor correlação entre os BLUPs da população de validação e os BLUEs foram os que apresentavam menor população de treinamento, dado o menor número de indivíduos relacionados e o menor número de informação de local, presentes na mesma. Observa-se no VC3.1 e VC3.2 a grande amplitude de variação das acurácias de predição, dado que as piores predições (menores acurácias), foram obtivas para grandes populações de validação uma vez que se tinha pouca informação de genótipos relacionados e de locais para a predição de novos materiais nas populações de treinamento.

Como visto, os esquemas VC4.1 e VC4.2 foram os que apresentaram melhor acurácia de predição. Neste esquema o número de novos indivíduos preditos foi menor do que nos outros esquemas, ou seja, menor população de validação, dessa forma, o número de informações de indivíduos aparentados, e informações por local de avaliação, foi muito maior, tornando o

modelo mais informativo, e de maior precisão nas predições. Em estudo anterior realizado para rendimento de biomassa em sorgo, Yu et al. (2016), obtiveram acurácias de predição variando de 0,56 a 0,76. Os autores também relataram que a população de treinamento e o modelo de predição usado, foram os fatores mais influenciaram na acurácia de predição dos modelos.

Segundo Hayes et al. [57], a acurácia de predição (r) geralmente mostra uma correlação positiva com herdabilidade (h^2). Em estudos com produção de massa verde em sorgo biomassa, Fernandes et al. [17] concordaram com esses resultados, já Oliveira [13] não capturaram conexão entre estes dois parâmetros. Nos nossos estudos os resultados corroboraram com estas pressuposições, sendo os valores médios de acurácia (0,50) próximos dos valores de herdabilidade, que variaram entre 0,58 e 0,71.

4. Acurácia de predição para híbridos não sintetizados

A acurácia da predição de híbridos não sintetizados, usando informações de relacionamento baseadas em dados de marcadores SNPs, foi fortemente influenciada pelo grau em que as linhagens genitoras era representadas na população de treinamento. Com isso, estudos mostram que a acurácia de predição da seleção genômica aumenta à medida que o grau de parentesco entre as linhagens genitoras aumenta [4,13,58]. Segundo Zhao et al. [59], o conjunto de treinamento deve ser uma representação do espaço genético do programa de melhoramento para se conseguir predições acuradas. Technow et al. [60] também mostraram que número de híbridos por genitor e o número total de genitores e híbridos no conjunto de treinamento podem afetar a acurácia de predição dos modelos estudados.

Nos esquemas T0 e T2 o tamanho da população de treinamento, e o grau de relacionamento entre as populações de validação e de treinamento foram os principais agentes causadores de diminuição da acurácia de predição. Mostrando que quanto maior a população

de validação, menor o número de informações por genitor disponíveis no modelo de predição, menor o número de informação relacionadas com a população de relacionamento e menor a acurácia de predição do modelo. Assim, a utilização de genitores com maior consanguinidade gera um aumento na eficiência de predição dos modelos de seleção genômica, e o conhecimento da matriz de parentesco através da matriz de marcadores SNPs, favorece a seleção de genitores em estágios iniciais de avaliação. Nota-se na matriz *Kinship* (Figura 4) uma leve tendência das linhagens fêmeas (Linhagens A) serem mais próximas geneticamente entre si, ou seja, apresentarem maior consanguinidade, do que as linhagens macho, o que pode ser observado pela maior coloração avermelhada dentro do grupo de linhagens fêmeas (tarja roxa). O que retratou em uma leve superioridade na predição de novos híbridos em modelos sem informação das linhagens genitoras femininas (T1F) em nosso estudo, dado o maior número de informação genética disponível nas populações de treinamento.

Como observado nos resultados, 23 híbridos foram os genótipos mais produtivos em todos os esquemas de validação, no qual foram sintetizados através de sete linhagens A (macho-estéreis) e 11 linhagens R (macho-férteis). Entre as fêmeas se destacam as linhagens LinA_17 e LinA_26, que foram genitoras de 10 e 5 destes híbridos, respectivamente. Estas linhagens, como mostrado na Figura Suplementar 2, apresentam alta proximidade genética, observado pela proximidade no dendograma. Já quando se avalia os machos se destacam as linhagens LinR_1 e LinR_7, genitoras de 5 e 4 híbridos, respectivamente, que no dendograma ficaram bem próximas, mostrando também uma maior proximidade genética. Estes resultados mostram a estreita gama de genitores superiores do programa de melhoramento de sorgo da Embrapa, sendo necessário a busca por novas linhagens a fim de aumentar o potencial de síntese de híbridos com maior produção de massa verde.

Technow et al. [60] afirmam que a seleção genômica deve ser empregada como um estágio inicial dos programas de melhoramento, envolvendo posteriormente testes de campo

dos híbridos promissores selecionados. Araus e Cairns [61] recomendam a adoção de técnicas de fenotipagem de alto rendimento como medidas integradas à seleção genômica. Também, Podlich et al. [62] mostram que os modelos de seleção genômica, devem ser constantemente atualizadas, assim, a inserção de novos indivíduos genotipados e fenotipados à população de treinamento, além da inserção de novas linhagens parentais, devem ser realizadas para manter a eficiência de predição ao longo dos anos.

Os programas de melhoramento sorgo biomassa, visam genótipos com alto poder de queima, dado pelos maiores pesos de massa seca dos materiais. Mas devido ao grande número de processos para a mensuração dessa variável, o baixo número de ensaios avaliados para esta característica, e a alta correlação com a produção de massa verde ($\rho = 0,93$), como relatado por Almeida [24], optamos por realizar os estudos de seleção genômica em sorgo biomassa para a variável produção de massa verde.

Os resultados do nosso estudo mostraram que é possível obter altos níveis de acurácia de predição para produção de massa verde através da inclusão de efeitos de interação genótipo por ambiente e da matriz de relacionamento genômico, dados os efeitos aditivo e de dominância, em modelos em múltiplos ambientes de seleção genômica. Dessa forma, concluímos que o uso da seleção genômica pode aumentar significativamente o ganho genético por ano e que os resultados deste estudo justificam mais pesquisas sobre a integração da seleção genômica no melhoramento de plantas de sorgo da Embrapa. O avanço contínuo da genotipagem de alto rendimento, dos modelos estatísticos para o cálculo das médias preditas *via* GBLUP, e das metodologias de fenotipagem só fortalecerão para os avanços da seleção genômica no melhoramento de plantas.

Conclusões

A seleção genômica em sorgo biomassa usando dados históricos pode ser realizada com altas acurácias preditivas para a característica produção de massa verde, contribuindo para uma seleção precoce e eficiente de híbridos superiores, para diferentes locais, do programa de melhoramento de sorgo da EMBRAPA. Além disso, os melhoristas podem ter maior eficiência na seleção de híbridos ainda não testados e maior otimização de recursos, levando para o campo apenas combinações híbridas mais promissoras. Os procedimentos avaliados no presente trabalho sugerem que a predição genômica pode ser diretamente aplicada para reduzir o tempo necessário para o desenvolvimento de novas cultivares.

Agradecimentos

Os autores agradecem todos os pesquisadores, estudantes, e técnicos da Embrapa Milho e Sorgo, e aos mestres da Universidade Federal de Viçosa que colaboraram para a elaboração e condução deste trabalho.

Financiamento

Este trabalho foi apoiado pelo BNDES (Banco Nacional de Desenvolvimento Econômico e Social), CNPq (Concelho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) and Embrapa (Empresa Brasileira de Pesquisa Agropecuária).

Referências

[1] J. Dahlberg, The role of sorghum in renewables and biofuels, in: Sorghum, Springer, 2019: pp. 269–277.

[2] I.C.M. Oliveira, T.D.S. Marçal, C. Bernardino, P.C. De Oliveira, R. Augusto, P. Crescêncio, S. Carneiro, R.E. Schaffert, Combining Ability of Biomass Sorghum Lines for Agroindustrial Characters and Multitrait Selection of Photosensitive Hybrids for Energy Cogeneration, 1566 (2019) 1554–1566.

[3] I.C.M. Oliveira, C.M.B. Damasceno, M.M. Pastina, R.A. da C. Parrella, Desempenho Produtivo de Híbridos de Sorgo Biomassa do Programa de Melhoramento da Embrapa, Bol. Pesqui. E Desenvolv. 187 (2019).

[4] C.H. Hunt, F.A. van Eeuwijk, E.S. Mace, B.J. Hayes, D.R. Jordan, Development of genomic prediction in sorghum, Crop Sci. 58 (2018) 690–700.

[5] D. Udoh, S.K. Rasmussen, S. Jacobsen, G.A. Iwo, W. De Milliano, Yield Stability of Sweet Sorghum Genotypes for Bioenergy Production Under Contrasting Temperate and Tropical Environments, 10 (2019) 42–53.

[6] K.O.G. Dias, H.P. Piepho, L.J.M. Guimarães, P.E.O. Guimarães, S.N. Parentoni, M.O. Pinto, R.W. Noda, C.T. Guimarães, A.A.F. Garcia, M.M. Pastina, Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data, Theor. Appl. Genet. (2019) 1–22.

[7] B.J. Hayes, M.E. Goddard, others, Prediction of total genetic value using genome-wide dense marker maps, Genetics. 157 (2001) 1819–1829.

[8] K.O.D.G. Dias, S.A. Gezan, C.T. Guimarães, A. Nazarian, L. Da Costa E Silva, S.N. Parentoni, P.E. De Oliveira Guimarães, C. De Oliveira Anoni, J.M.V. Pádua, M. De

Oliveira Pinto, R.W. Noda, C.A.G. Ribeiro, J.V. De Magalhães, A.A.F. Garcia, J.C. De Souza, L.J.M. Guimarães, M.M. Pastina, Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials, Heredity (Edinb). 121 (2018) 24–37.

[9] P.K. Voss-Fels, M. Cooper, B.J. Hayes, Accelerating crop genetic gains with genomic selection, Theor. Appl. Genet. 132 (2019) 669–686.

[10] B. Rice, A.E. Lipka, Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum, Plant Genome. 12 (2019).

[11] L.W. Pembleton, C. Inch, R.C. Baillie, M.C. Drayton, P. Thakur, Y.O. Ogaji, G.C. Spangenberg, J.W. Forster, H.D. Daetwyler, N.O.I. Cogan, Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass, Theor. Appl. Genet. 131 (2018) 1891–1902.

[12] J. Crossa, P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, C. Magorokosho, Genomic selection and prediction in plant breeding, J. Crop Improv. 25 (2011) 239–261.

[13] A.A. de Oliveira, M.M. Pastina, V.F. de Souza, R.A. da Costa Parrella, R.W. Noda, M.L.F. Simeone, R.E. Schaffert, J.V. de Magalhães, C.M.B. Damasceno, G.R.A. Margarido, Genomic prediction applied to high-biomass sorghum for bioenergy production, Mol. Breed. 38 (2018).

[14] R.A. da C. Parrella, J.A.S. Rodrigues, F.D. Tardin, C.M.B. Damasceno, R.E. Schaffert, Desenvolvimento de híbridos de sorgo sensíveis ao fotoperíodo visando alta produtividade de biomassa., Embrapa Milho e Sorgo-Boletim Pesqui. e Desenvolv. (2010).

[15] D.J. Packer, W.L. Rooney, High-parent heterosis for biomass yield in photoperiod-sensitive sorghum hybrids, F. Crop. Res. 167 (2014) 153–158.

[16] X. Yu, X. Li, T. Guo, C. Zhu, Y. Wu, S.E. Mitchell, K.L. Roozeboom, D. Wang, M.L. Wang, G.A. Pederson, others, Genomic prediction contributing to a promising global strategy to turbocharge gene banks, Nat. Plants. 2 (2016) 16150.

[17] S.B. Fernandes, K.O.G. Dias, D.F. Ferreira, P.J. Brown, Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum, Theor. Appl. Genet. 131 (2018) 747–755.

[18] H. Oakey, B. Cullis, R. Thompson, J. Comadran, C. Halpin, R. Waugh, Genomic selection in multi-environment crop trials, G3 Genes, Genomes, Genet. 6 (2016) 1313–1326.

[19] A. Mocoeur, Y.M. Zhang, Z. Quan, L. Xin, L.M. Zhang, Stability and genetic control of morphological , biomass and biofuel traits under temperate maritime and continental conditions in sweet sorghum (Sorghum bicolour), Theor. Appl. Genet. 128 (2015) 1685–1701.

[20] F.M. Bassi, A.R. Bentley, G. Charmet, R. Ortiz, J. Crossa, Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.), Plant Sci. 242 (2016) 23–36.

[21] J. Möhring, H.-P. Piepho, Comparison of one-stage and two-stage analysis in series of experiments, Lifestat 2008. (2008) 108.

[22] J. Möhring, H.-P. Piepho, Comparison of weighting in two-stage analysis of plant breeding trials, Crop Sci. 49 (2009) 1977–1988.

[23] J. Burgueño, G. de los Campos, K. Weigel, J. Crossa, Genomic prediction of breeding

values when modeling genotype$\times$ environment interaction using pedigree and dense molecular markers, Crop Sci. 52 (2012) 707–719.

[24] L.G.F. de Almeida, R.A. da C. Parrella, M.L.F. Simeone, P.C.D.O. Ribeiro, A.S. dos Santos, A.S.V. da Costa, A.G. Guimarães, R.E. Schaffert, Biomass and Bioenergy Composition and growth of sorghum biomass genotypes for ethanol production, 122 (2019) 343–348.

[25] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS One. 6 (2011) e19379.

[26] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, E.S. Buckler, TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, PLoS One. 9 (2014) e90346.

[27] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods. 9 (2012) 357.

[28] J. Crossa, Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueño, V.S. Windhausen, E. Buckler, others, Genomic prediction in maize breeding populations with genotyping-by-sequencing, G3 Genes, Genomes, Genet. 3 (2013) 1903–1926.

[29] A. Roberts, L. McMillan, W. Wang, J. Parker, I. Rusyn, D. Threadgill, Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows, Bioinformatics. 23 (2007) i401--i407.

[30] P.M. VanRaden, Efficient methods to compute genomic predictions, J. Dairy Sci. 91 (2008) 4414–4423.

[31] Z.G. Vitezica, L. Varona, A. Legarra, On the additive and dominant variance and covariance of individuals within the genomic selection scope, Genetics. 195 (2013) 1223–1230.

[32] R.R. Amadeu, C. Cellon, J.W. Olmstead, A.A.F. Garcia, M.F.R. Resende, P.R. Muñoz, AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example, Plant Genome. 9 (2016).

[33] A. Nazarian, S.A. Gezan, GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits, J. Hered. 107 (2016) 372–379.

[34] H. Jorjani, L. Klei, U. Emanuelson, A simple method for weighted bending of genetic (co) variance matrices, J. Dairy Sci. 86 (2003) 677–679.

[35] J.B. Endelman, J.-L. Jannink, Shrinkage estimation of the realized relationship matrix, G3 Genes, Genomes, Genet. 2 (2012) 1405–1413.

[36] P.H.A. Sneath, R.R. Sokal, others, Numerical taxonomy. The principles and practice of numerical classification., 1973.

[37] A.B. Smith, B.R. Cullis, R. Thompson, Analysing variety by environment data using multiplicative mixed models, Biometrics. 57 (2001) 1138–1147.

[38] J. Burgueño, G. de los Campos, K. Weigel, J. Crossa, Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers, Crop Sci. 52 (2012) 707–719.

[39] H. Bozdogan, Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, Psychometrika. 52 (1987) 345–370.

[40] G. Schwarz, others, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.

[41] R Core Team, R: A language and environment for statistical computing, (2017).

[42] D.G. Butler, B.R. Cullis, A.R. Gilmour, B.J. Gogel, ASReml-R reference manual, State Queensland, Dep. Prim. Ind. Fish. Brisbane. (2009).

[43] R.C. Team, others, R: A language and environment for statistical computing, (2015).

[44] J. Hamblin, M.J. de O. Zimmermann, Breeding common bean for yield in mixtures, Plant Breed. Rev. 4 (1986) 245–272.

[45] R. Bernardo, Prediction of maize single-cross performance using RFLPs and information from related hybrids, Crop Sci. 34 (1994) 20–25.

[46] T.A. Schrag, J. Möhring, A.E. Melchinger, B. Kusterer, B.S. Dhillon, H.-P. Piepho, M. Frisch, Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds, Theor. Appl. Genet. 120 (2010) 451–461.

[47] A.K. Sonesson, T.H.E. Meuwissen, Testing strategies for genomic selection in aquaculture breeding programs, Genet. Sel. Evol. 41 (2009) 37.

[48] E.L. Heffner, A.J. Lorenz, J.-L. Jannink, M.E. Sorrells, Plant breeding with genomic selection: gain per unit time and cost, Crop Sci. 50 (2010) 1681–1690.

[49] C. Riedelsheimer, J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink, A.E. Melchinger, Genomic predictability of interconnected biparental maize populations, Genetics. 194 (2013) 493–503.

[50] J.P.R. dos Santos, R.C. de Castro Vasconcellos, L.P.M. Pires, M. Balestre, R.G. Von Pinho, Inclusion of dominance effects in the multivariate GBLUP model, PLoS One. 11 (2016) e0152045.

[51] M. Balestre, P.P. Torga, R.G. Von Pinho, J.B. Dos Santos, Applications of multi-trait

selection in common bean using real and simulated experiments, Euphytica. 189 (2013) 225–238.

[52] B. Tan, D. Grattapaglia, H.X. Wu, P.K. Ingvarsson, Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus, Plant Sci. 267 (2018) 84–93.

[53] F. Enciso-Rodriguez, D. Douches, M. Lopez-Cruz, J. Coombs, G. de los Campos, Genomic Selection for Late Blight and Common Scab Resistance in Tetraploid Potato (Solanum tuberosum), 8 (2018).

[54] Y. Da, C. Wang, S. Wang, G. Hu, Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers, PLoS One. 9 (2014) e87666.

[55] N. Heslot, J.-L. Jannink, M.E. Sorrells, Perspectives for genomic selection applications and research in plants, Crop Sci. 55 (2015) 1–12.

[56] E. Combs, R. Bernardo, Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers, Plant Genome. 6 (2013).

[57] B.J. Hayes, H.D. Daetwyler, P. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H.W. Raadsma, M.E. Goddard, others, Accuracy of genomic selection: comparing theory and results, in: Proc Assoc Advmt Anim Breed Genet, 2009: pp. 34–37.

[58] F. Technow, C. Riedelsheimer, T.A. Schrag, A.E. Melchinger, Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects, Theor. Appl. Genet. 125 (2012) 1181–1194.

[59] Y. Zhao, M. Gowda, F.H. Longin, T. Würschum, N. Ranc, J.C. Reif, Impact of selective genotyping in the training population on accuracy and bias of genomic selection, Theor.

Appl. Genet. 125 (2012) 707–713.

[60] F. Technow, T.A. Schrag, W. Schipprack, E. Bauer, H. Simianer, A.E. Melchinger, Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize, Genetics. 197 (2014) 1343–1355.

[61] J.L. Araus, S.C. Kefauver, M. Zaman-Allah, M.S. Olsen, J.E. Cairns, Translating high-throughput phenotyping into genetic gain, Trends Plant Sci. 23 (2018) 451–466.

[62] D.W. Podlich, C.R. Winkler, M. Cooper, Mapping as you go, Crop Sci. 44 (2004) 1560–1571.

Material suplementar

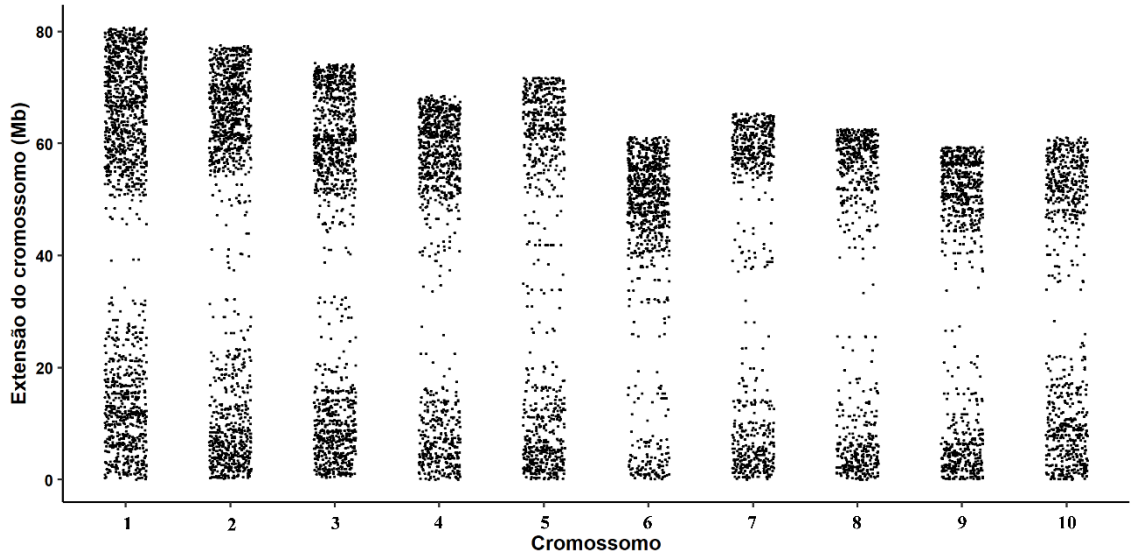


Figura Suplementar 1. Distribuição dos SNPs nos dez cromossomos de sorgo biomassa, após imputação e filtragem dos dados.

Tabela Suplementar 1: Acurácia média baseada nos seis esquemas de validação cruzada para a característica produção de massa verde (PMV t ha^{-1}) usando modelos aditivos (A) e modelos aditivos + dominantes (A + D).

Esquemas de validação	Modelo A	Modelo A + D
VC1	0,46 ±0,04	0,49 ±0,04
VC2	0,45 ±0,02	0,47 ±0,02
VC3	0,26 ±0,12	0,28 ±0,12
VC3.1	0,35 ±0,15	0,38 ±0,15
VC4	0,72 ±0,08	0,75 ±0,07
VC4.1	0,73 ±0,13	0,77 ±0,06

± valores de Desvio Padrão.

Tabela Suplementar 2: Acurácia média baseada em três esquemas de teste para a característica produção de massa verde (PMV t ha^{-1}) usando modelos aditivos (A) e modelos aditivos + dominantes (A + D).

Esquemas de teste		Modelo A	Modelo A + D
T0	1F1M	0,79 ± 0,03	0,83 ± 0,03
	2F2M	0,75 ± 0,05	0,80 ± 0,04
	3F3M	0,71 ± 0,07	0,76 ± 0,06
	4F4M	0,49 ± 0,09	0,59 ± 0,12
T1	2Fem	0,84 ± 0,08	0,84 ± 0,14
	2Mac	0,79 ± 0,06	0,84 ± 0,06
	3Fem	0,79 ± 0,05	0,81 ± 0,09
	3Mac	0,59 ± 0,08	0,69 ± 0,09
	4Fem	0,71 ± 0,07	0,75 ± 0,10
	4Mac	0,63 ± 0,09	0,73 ± 0,08
	5Fem	0,79 ± 0,05	0,85 ± 0,06
	5Mac	0,67 ± 0,15	0,69 ± 0,17
T2	10%	0,79 ± 0,07	0,86 ± 0,06
	20%	0,69 ± 0,05	0,77 ± 0,04
	30%	0,64 ± 0,05	0,71 ± 0,05

± valores de Desvio Padrão.

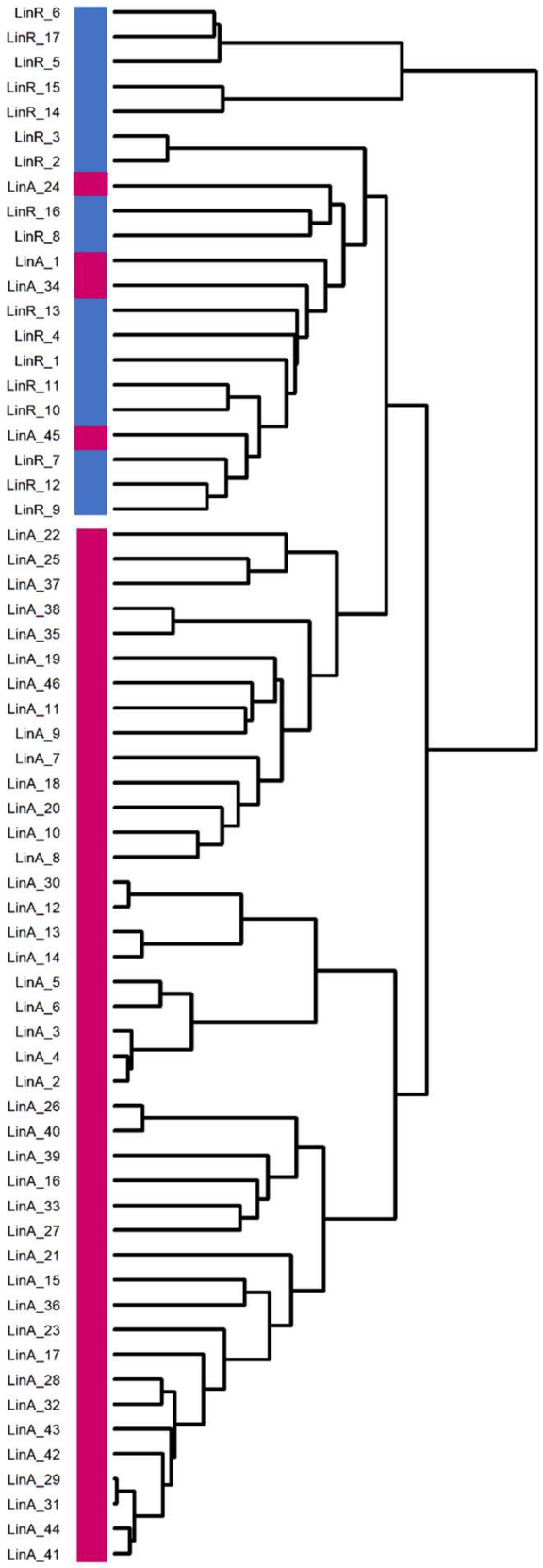


Figura Suplementar 2. Formação de grupos de parentesco entre linhagens genitoras (A) e híbridos (B) avaliados no programa de melhoramento.

APÊNDICES

Tabela 1: Médias fenotípicas dos 202 híbridos experimentais e das quatro testemunhas comerciais avaliadas em ensaios de VCU de sorgo biomassa, entre os anos de 2012 a 2017, em dez locais distribuídos no território brasileiro.

Trat	Código	Média	Trat	Código	Média	Trat	Código	Média
H_001	201209B01	61,83	H_039	201209B45	71,92	H_080	201342B04	54,94
H_002	201209B02	50,56	H_040	201209B46	59,40	H_081	201342B05	83,89
H_003	201209B03	61,58	H_041	201209B47	56,66	H_084	201342B08	50,16
H_004	201209B04	44,93	H_042	201209B48	65,57	H_085	201342B09	71,91
H_005	201209B05	59,49	H_043	201209B49	60,38	H_086	201342B10	73,53
H_006	201209B06	70,23	H_044	CMSXS7025	54,23	H_087	201342B11	87,62
H_007	201209B07	80,58	H_045	CMSXS7012	67,38	H_088	201342B12	59,21
H_008	201209B08	62,18	H_046	CMSXS7031	60,24	H_090	201342B14	45,37
H_009	201209B09	70,78	H_047	CMSXS7027	65,75	H_091	201342B15	59,14
H_010	201209B10	79,09	H_048	201209B59	73,16	H_092	201342B16	72,45
H_011	201209B11	68,59	H_049	CMSXS7026	59,21	H_094	201342B18	73,35
H_012	201209B12	53,39	H_050	CMSXS7023	62,46	H_095	201342B19	52,90
H_013	201209B13	67,76	H_051	CMSXS716	81,58	H_096	201342B20	61,20
H_014	201209B14	59,59	H_052	201209B63	59,30	H_097	201342B21	75,45
H_015	201209B15	61,72	H_053	201209B64	54,73	H_099	201342B23	63,24
H_016	201209B21	65,10	H_054	CMSXS7016	76,99	H_100	201342B24	79,12
H_017	201209B22	59,37	H_055	CMSXS7028	66,11	H_103	201342B27	67,11
H_018	201209B23	66,86	H_056	201209B67	53,39	H_105	201342B29	72,03
H_019	201209B24	66,17	H_057	201209B68	67,54	H_107	201342B31	74,59
H_020	201209B25	77,14	H_058	CMSXS7029	60,25	H_108	201342B32	53,63
H_021	201209B26	72,72	H_059	201209B70	66,09	H_109	201342B33	96,35
H_022	201209B27	55,21	H_060	201209B72	62,40	H_110	201342B34	70,85
H_023	201209B28	57,94	H_061	201209B73	46,83	H_111	201342B35	67,58
H_024	201209B29	60,44	H_062	201209B74	54,30	H_112	201342B36	39,80
H_025	201209B30	64,06	H_063	201209B75	70,40	H_114	201342B38	49,52
H_026	201209B31	44,90	H_064	201209B77	63,57	H_115	201342B39	52,18
H_027	201209B32	43,70	H_065	201209B78	59,59	H_116	201342B40	62,19
H_028	201209B33	65,20	H_066	201209B79	61,04	H_119	201427B01	66,29
H_029	201209B34	71,72	H_067	201209B82	61,74	H_120	201427B02	76,10
H_030	201209B35	75,08	H_068	201209B83	78,56	H_121	201427B03	84,57
H_031	201209B36	61,14	H_069	201209B84	57,39	H_122	201427B04	54,76
H_032	201209B37	54,51	H_070	201209B85	73,58	H_123	201427B05	89,71
H_033	201209B38	50,97	H_071	201209B86	58,48	H_124	201427B06	90,70
H_034	201209B39	74,07	H_072	201209B87	62,97	H_125	201427B07	79,11
H_035	201209B40	52,63	H_073	201209B88	81,37	H_126	201427B08	78,86
H_036	201209B41	70,70	H_074	201209B89	74,37	H_127	201427B09	72,18
H_037	201209B42	60,66	H_075	201209B90	69,61	H_128	201427B10	72,96
H_038	201209B43	71,85	H_076	201209B91	54,64	H_129	201427B11	77,34

Continua...

Trat	Código	Média
H_130	201427B12	52,42
H_131	201427B13	65,09
H_132	201427B14	55,14
H_133	201427B15	60,95
H_134	CMSXS7101	80,71
H_135	201427B17	96,19
H_136	201427B18	74,19
H_137	201427B19	58,58
H_138	201427B20	55,08
H_139	201427B21	85,09
H_140	201427B22	80,46
H_141	201428B13	79,38
H_142	201428B14	75,44
H_143	201428B15	68,58
H_144	201428B16	62,51
H_145	201428B17	91,03
H_146	201428B18	68,31
H_147	201428B19	90,46
H_148	201428B20	81,78
H_149	201428B21	65,82
H_150	201428B22	73,84
H_151	201428B23	71,89
H_152	201428B24	88,37
H_153	201428B26	76,34
H_154	201428B27	57,07
H_155	201428B30	88,07
H_156	201428B32	90,27
H_157	201428B33	56,92
H_158	201428B37	83,02
H_159	201428B38	69,20
H_160	201428B39	62,40

Trat	Código	Média
H_161	201428B43	81,39
H_162	201428B44	89,44
H_163	201428B45	63,32
H_164	201428B47	67,09
H_165	201551B03	98,65
H_166	201551B04	87,40
H_167	201551B05	76,56
H_168	201551B11	74,35
H_169	201551B13	76,09
H_170	201551B15	65,92
H_171	201551B23	64,09
H_172	201551B24	48,67
H_173	201551B26	50,73
H_174	201551B28	73,92
H_175	201551B29	56,22
H_176	201551B30	39,98
H_177	201551B36	62,49
H_178	201551B41	82,26
H_179	201551B43	50,48
H_180	201551B57	63,99
H_181	201735B03	103,34
H_182	201735B05	77,98
H_183	201735B06	91,87
H_184	201735B27	77,19
H_185	201735B28	73,17
H_186	201735B35	88,31
H_187	201735B38	89,27
H_188	201735B47	87,02
H_189	201735B48	85,20
H_190	201735B50	75,66
H_191	201735B51	61,86

Trat	Código	Média
H_192	201735B53	65,74
H_193	201735B58	75,94
H_194	201735B59	74,79
H_195	201735B60	75,50
H_196	201735B70	72,50
H_197	201735B73	63,40
H_198	201334B01	59,98
H_199	201334B02	71,57
H_200	201334B04	72,99
H_201	201334B10	60,24
H_202	201429B02	66,21
H_203	201429B03	68,38
H_204	201429B05	74,55
H_205	201429B06	66,59
H_206	201429B09	64,65
H_207	201429B14	60,76
H_208	201429B15	62,08
H_209	201429B23	61,98
H_210	201429B24	63,40
H_211	201429B25	58,92
H_212	201429B28	60,09
H_213	201545B012	71,04
H_214	201636B07	74,88
H_215	201636B09	77,76
H_216	201636B18	76,57
H_217	201636B21	78,17
T_01	N52K1009	62,46
T_02	Volumax	44,82
T_03	BRS610	39,38
T_04	BRS655	34,44

CONCLUSÕES GERAIS

Modelos de fator analítico (FA) possibilitaram o estudo e entendimento das interações G × A em ensaios de valor de cultivo e uso do programa de melhoramento de sorgo biomassa da Embrapa Milho e Sorgo, sendo possível capturar as correlações genéticas existentes entre os ambientes e identificar genótipos adaptados e/ou estáveis aos ambientes avaliados, ou à ambientes específicos. Além disso, os melhoristas de sorgo podem adotar a seleção genômica para a característica produção de massa verde, com altas acurácias, na predição de híbridos em múltiplos ambientes e híbridos ainda não sintetizados, possibilitando a redução do número de genótipos inferiores avaliados em campo e do tempo para lançamento de novas cultivares. Dessa forma, a implementação destes dois métodos às análises de dados dos programas de melhoramento de sorgo biomassa podem promover um aumento do ganho genético anual e favorecer a seleção de genótipos geneticamente superiores para a característica produção de massa verde.