**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Métodos de seleção genômica aplicados a sorgo biomassa para produção de
etanol de segunda geração**

Amanda Avelar de Oliveira

Dissertação apresentada para a obtenção do título de
Mestra em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2015**

Amanda Avelar de Oliveira
Engenheira Agrônoma

Métodos de seleção genômica aplicados a sorgo biomassa para produção de etanol de segunda geração

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

Dissertação apresentada para a obtenção do título de
Mestra em Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2015

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

Oliveira, Amanda Avelar de

Métodos de seleção genômica aplicados a sorgo biomassa para produção de etanol de segunda geração / Amanda Avelar de Oliveira. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2015.

129 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Bioenergia 2. Modelos preditivos 3. Acurácia 4. Genotipagem por sequenciamento
5. Enriquecimento funcional I. Título

CDD 633.174
O48m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Aos meus pais, Carlos Alberto e Maria Lilian, pelo amor e dedicação. Especialmente, por nunca medirem esforços para me proporcionar uma educação de qualidade. Dedico.

AGRADECIMENTOS

A Deus por guiar meus caminhos.

Aos meus pais, Carlos Alberto e Maria Lilian, pelo apoio constante, por acreditarem nos meus sonhos e se fazerem diariamente presentes apesar da distância.

À minha irmã, Ana Luiza, pelo companheirismo, amizade e por estar sempre disposta a ajudar

Ao Prof. Dr. Gabriel Rodrigues Alves Margarido, pela excelente orientação e paciência em ensinar, contribuindo para o desenvolvimento deste trabalho, assim como para o meu desenvolvimento profissional e pessoal. Acima de tudo, por ser exemplo de profissionalismo e dedicação.

Ao Prof. Dr. Antonio Augusto Franco Garcia, por ter me recebido tão bem em seu laboratório fazendo eu me sentir como se fosse do seu grupo. Certamente a convivência neste ambiente de trabalho foi inspiradora e valiosa.

Aos demais professores e funcionários do Departamento de Genética que contribuíram para a minha formação..

Aos amigos do Laboratório de Genética Estatística, Marcelo, Guilherme, Adriana, João Ricardo, Carina, Luis Felipe, Marianella, Letícia, Danilo, Fernando, Christiane, Maria Izabel e Rodrigo pela harmoniosa convivência e por contribuírem imensamente pelo meu amadurecimento profissional e pessoal. Em especial ao Guilherme, por estar sempre disposto a ajudar e a transmitir seus conhecimentos. Ao Luis Felipe, por todas as conversas sobre seleção genômica, sempre esclarecendo minhas dúvidas e compartilhando experiências.

Às amigas de piracicaba Evelyn, Fernanda, Letícia, Maria Izabel e Natália pela companhia e por fazerem a saudade de casa ser menos dolorosa.

A todos os meus familiares, pelo carinho.

Às amigas de Lavras e da UFLA, por sempre me apoiarem e por entenderem meus momentos de ausência.

À Embrapa Milho e Sorgo pela disponibilização dos dados genotípicos e fenotípicos, e valiosas discussões.

À Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pela oportunidade e qualidade de ensino.

À CAPES e FAPESP, processo 2013/25132-7, pelas bolsas de estudo concedidas.

"Everybody is a genius. But if you judge a fish by its ability of climb a tree, it will live its whole life believing that it is stupid"

Albert Einstein

SÚMARIO

RESUMO	11
ABSTRACT	13
1 INTRODUÇÃO	15
2 REVISÃO BIBLIOGRÁFICA.....	19
2.1 Sorgo.....	19
2.1.1 Sorgo Biomassa	19
2.2 Melhoramento Genético do Sorgo Biomassa	21
2.3 Recursos Genéticos e Genômicos Disponíveis	22
2.4 Genotipagem por Sequenciamento	23
2.5 Desequilíbrio de Ligação.....	25
2.6 Seleção Genômica	27
2.6.1 Populações de Treinamento, Teste e Seleção	29
2.6.2 Acurácia da Seleção Genômica	31
2.6.3 Modelos Estatísticos.....	31
2.6.3.1 Modelo RR-BLUP/G-BLUP	32
2.6.3.2 Família de Modelos Bayesianos	33
2.6.3.2.1 Modelos BayesA.....	34
2.6.3.2.3 Modelos BayesC π	35
2.6.3.2.4 Modelos BayesRR	35
2.6.3.2.5 Modelos Lasso Bayesiano	36
3 MATERIAL E MÉTODOS.....	39
3.1 Material.....	39
3.1.1 Material Genético	39
3.1.2 Dados Moleculares	39
3.1.2.1 Análise Descritiva dos Dados Genotípicos	40
3.1.3 Dados Fenotípicos	40
3.2 Métodos	41
3.2.1 Análise Fenotípica	41
3.2.2 Ajuste dos Modelos de Seleção Genômica.....	43
3.2.2.1 Análise Conjunta dos Dois Sub-Painéis	44
3.2.2.2 Seleção Genômica entre Sub-painéis	44

3.2.2.3 Simulações de Redução do Número de Marcadores.....	45
3.2.3 Relação entre Capacidade Preditiva e Herdabilidade	46
3.2.4 Comparação dos Modelos.....	46
3.2.5 Enriquecimento Funcional	46
4 RESULTADOS.....	49
5 DISCUSSÃO	85
6 CONCLUSÕES	93
REFERÊNCIAS.....	95
ANEXOS	107

RESUMO

Métodos de seleção genômica aplicados a sorgo biomassa para produção de etanol de segunda geração

As crescentes preocupações com questões ambientais têm despertado interesse global pelo uso de combustíveis alternativos, e o uso da biomassa vegetal surge como uma alternativa viável para a geração de biocombustíveis. Diferentes materiais orgânicos têm sido utilizados, e dentre eles destaca-se o sorgo biomassa (*Sorghum bicolor* L. Moench). A seleção genômica apresenta grande potencial e pode, em médio prazo, reestruturar os programas de melhoramento de plantas, promovendo maiores ganhos genéticos quando comparada a outros métodos, além de reduzir significativamente o tempo necessário para o desenvolvimento de novas cultivares, através da seleção precoce. Este trabalho teve como objetivo avaliar modelos de seleção genômica e aplicá-los para a predição dos valores genéticos de indivíduos do painel de sorgo biomassa da Embrapa/Milho e Sorgo. Tal painel inclui materiais do banco de germoplasma e materiais utilizados em programas de melhoramento de sorgo dessa instituição, bem como coleções núcleo do CIRAD e ICRISAT, sendo, portanto, subdividido em dois sub-painéis. As 100 linhagens do sub-painel 1 foram avaliadas fenotipicamente por dois anos (2011 e 2012) e as 100 linhagens do sub-painel 2 por um ano (2011), ambas no município de Sete Lagoas-MG, para as seguintes características fenotípicas: tempo até o florescimento, altura de plantas, produção de massa verde e massa seca, proporções de fibra ácida e neutra, celulose, hemicelulose e lignina. Posteriormente, as 200 linhagens integrantes do painel foram genotipadas através da técnica de genotipagem por sequenciamento. A partir desses dados genotípicos e fenotípicos, os modelos de seleção genômica Bayes A, Bayes B, Bayes C π , Bayes Lasso, Bayes Ridge Regression e Random Regression BLUP (RRBLUP) foram ajustados e comparados. As capacidades preditivas obtidas foram elevadas e pouco variaram entre os diversos modelos, variando de 0,61 para o caráter florescimento a 0,85 para a proporção de fibra ácida, quando o modelo RRBLUP foi empregado na análise conjunta dos dois sub-painéis. Por outro lado, a predição cruzada entre sub-painéis resultou em capacidades preditivas substancialmente menores, nunca superiores a 0,66 e em alguns cenários virtualmente iguais a zero, além de apresentar maiores variações entre os modelos ajustados. Simulações do uso de subconjuntos dos marcadores moleculares são apresentadas e indicam possibilidades de obtenção de capacidades preditivas mais elevadas. Análises de enriquecimento funcional realizadas a partir dos efeitos preditos dos marcadores sugeriram associações interessantes, as quais devem ser investigadas com maiores detalhes em estudos futuros, com potencial de elucidação da arquitetura genética dos caracteres quantitativos.

Palavras-chave: Bioenergia; Modelos preditivos; Acurácia; Genotipagem por sequenciamento; Enriquecimento funcional

ABSTRACT

Genome wide selection methods applied to high biomass sorghum for the production of second generation ethanol

Increased concerns about environmental issues have aroused global interest in the use of alternative fuels, and the use of plant biomass emerges as a viable alternative for the generation of biofuels. Different organic materials have been used, including high biomass sorghum (*Sorghum bicolor* L. Moench). Genomic selection has great potential and could, in the medium term, restructure plant breeding programs, promoting greater genetic gains when compared to other methods and significantly reducing the time required for the development of new cultivars through early selection. This work aimed at evaluating models of genomic selection and applying them to the prediction of breeding values for a panel of high biomass sorghum genotypes of Embrapa / Milho e Sorgo. This panel includes materials from the gene bank and materials used in sorghum breeding programs of this institution, as well as core collections from CIRAD and ICRISAT, and is therefore divided into two sub-panels. The 100 lines of sub-panel 1 were evaluated phenotypically for two years (2011 and 2012) and the 100 lines of sub-panel 2 for one year (2011), both in the city of Sete Lagoas, Minas Gerais, for the following phenotypic traits: days to flowering, plant height, fresh and dry matter yield and fiber, cellulose, hemicellulose and lignin proportions. Subsequently, the 200 lines were genotyped by via the genotyping by sequencing technique. From these genotypic and phenotypic data, genomic selection models Bayes A, Bayes B, Bayes C π , Bayes Lasso, Bayes Ridge Regression and Random Regression BLUP (RRBLUP) were fitted and compared. The predictive capabilities obtained were high and varied little between the different models, ranging from 0.61 for days to flowering to 0.85 for acid fiber, when the RRBLUP model was used on the combined analysis of the two sub-panels. On the other hand, cross prediction between sub-panels resulted in substantially lower predictive capability, never above 0.66 and in some scenarios virtually equal to zero, with greater variations between the fitted models. Simulations of using subsets of molecular markers are presented and indicate possibilities of achieving higher predictive capabilities. Functional enrichment analyses performed with the marker predicted effects suggested interesting associations, which should be investigated in more detail in future studies, with potential for elucidating the genetic architecture of quantitative traits.

Keywords: Bioenergy; Predictive models; Accuracy; Genotyping by sequencing; Functional enrichment

1 INTRODUÇÃO

As crescentes preocupações com poluição ambiental, segurança energética e reservas de combustíveis têm despertado o interesse global pelo uso de fontes alternativas para a produção de energia, especificamente, os biocombustíveis gerados a partir de biomassa (ROONEY et al., 2007; MORRIS et al., 2013). O uso da biomassa vegetal surge como uma alternativa viável para a geração de biocombustíveis. Diferentes materiais orgânicos têm sido utilizados, dentre eles o sorgo biomassa (*Sorghum bicolor* L. Moench), um tipo de sorgo com rápido crescimento e alto potencial produtivo, o qual tem grande potencial para se tornar uma cultura importante na produção de etanol. Isto é devido à sua alta produção de biomassa, baixos requerimentos de água e fertilizantes em termos relativos, sistemas de produção já bem estabelecidos, tolerância à seca e ao calor e grande diversidade genética (MURRAY et al., 2008).

O sorgo biomassa pode ser utilizado para produção de etanol de segunda geração ou para a co-geração de energia (bioeletricidade). O etanol de segunda geração, ou etanol celulósico, é produzido a partir da hidrólise ou quebra da biomassa vegetal em açúcares simples, os quais podem ser utilizados para produção de etanol a partir da fermentação, como é feito com o caldo da cana-de-açúcar ou sorgo sacarino (etanol de primeira geração) (EMBRAPA, 2011). Atualmente, estão entre os principais focos dos programas de melhoramento voltados à produção bioenergética a maximização da produção de biomassa por unidade de área plantada, visando minimizar o uso de área cultivável, e alterações nas proporções e estruturas dos vários polímeros que constituem a parede celular, uma vez que, para a conversão da biomassa em etanol de segunda geração, quanto menor o conteúdo de lignina ou diferenciação na sua composição, mais eficiente o processo de conversão (VERMERRIS et al., 2007). O melhoramento genético é a ciência ou arte que procura alterar geneticamente as plantas em benefício humano. Graças ao melhoramento genético, aumentos de produtividade e a melhoria de várias outras características de interesse para a agricultura têm ocorrido consistentemente. O sucesso de um programa de melhoramento está relacionado à sua capacidade de seleção de indivíduos superiores, ou seja, depende da habilidade em prever os valores genéticos a partir dos valores fenotípicos e/ou dados genotípicos (BUENO; MENDES; CARVALHO, 2006; BERNARDO, 2014a). Os programas de melhoramento genético de sorgo biomassa podem ser muito beneficiados pelo desenvolvimento de métodos que reduzam o tempo gasto para completar um ciclo de

melhoramento e possibilitem uma seleção mais precoce e eficiente de genótipos superiores, entre os quais a seleção genômica tem grande potencial de utilização.

O sorgo é uma espécie diploide ($2n = 20$), para a qual já está disponível uma montagem de referência do genoma, cujo tamanho é relativamente pequeno, em torno de 700 Mpb (PATERSON et al., 2009). Esta disponibilidade de recursos e informações facilita os esforços de melhoramento da espécie, visto que é possível corroborar resultados prévios e fazer uso de anotações do genoma para, por exemplo, fazer inferências a respeito de potenciais genes candidatos.

A estratégia de genotipagem por sequenciamento, ou GBS (*genotyping-by-sequencing*), é um sistema *multiplex* para genotipagem do genoma, o qual permite a identificação de milhares de polimorfismos de nucleotídeo único (*single nucleotide polymorphisms*, ou SNPs), bem como pequenas inserções e deleções (*indels*), a custo reduzido. O objetivo principal da técnica GBS é simultaneamente descobrir polimorfismos e obter informações genotípicas de toda a população de interesse. Esta combinação faz do GBS uma técnica rápida e flexível para uma variedade de espécies, como, por exemplo, cevada, trigo, milho, arroz, uva, cacau, sorgo, feijão e soja, sendo que pesquisas em muitas outras espécies estão em progresso (ELSHIRE; GLAUBITZ; SUN, 2011; HANSEY et al., 2012; POLAND et al., 2012; MORRIS et al., 2013; LU et al., 2013; SONAH et al., 2013). Devido ao grande número de marcadores detectados, essa técnica é também bastante apropriada para seleção genômica (POLAND; RIFE, 2012).

A seleção genômica, também conhecida como *Genome-Wide Selection* (GWS), foi inicialmente proposta em 2001 por Meuwissen e colaboradores, como uma forma de aumentar a eficiência da seleção assistida por marcadores e acelerar o ciclo do melhoramento genético, através do uso simultâneo de marcadores genéticos cobrindo todo o genoma. A seleção genômica tem como objetivo prever valores genéticos a partir de dados genotípicos, os quais podem presentemente ser obtidos em escala genômica e a custos acessíveis. Tal técnica apresenta grande potencial e pode, em médio prazo, reestruturar os programas de melhoramento de plantas, promovendo maiores ganhos genéticos quando comparada a outros métodos e reduzir significativamente o tempo necessário para o desenvolvimento de novas cultivares, através da seleção precoce (HEFFNER; SORRELLS; JANNINK, 2009).

A seleção genômica foi primeiramente empregada no melhoramento animal, em que o alto custo de fenotipagem e a impossibilidade de fazer uso de repetições de indivíduos tornaram-na atrativa (DEKKERS, 2007; LONG et al., 2008; LE et al., 2008; LEGARRA et al., 2008; VANRADEN et al., 2008). Posteriormente, atraiu também a atenção de melhoristas

de plantas (BERNADOR; YU, 2007; HEFFNER; SORRELLS; JANNINK, 2009; JANNINK; LORENZ; IWATA, 2010). Atualmente, estudos de simulação e empíricos em diversas culturas vêm sendo realizados, os quais têm comprovado a superioridade em termos de acurácia da seleção genômica sobre a seleção assistida por marcadores tradicional e a seleção baseada apenas em informações de parentesco (BERNARDO; YU, 2007; BERNARDO, 2009, 2014; ZHONG et al., 2009; LORENZANA; BERNARDO, 2009; MAYOR; BERNARDO, 2009; GRATTAPAGLIA; RESENDE, 2010; CROSSA et al., 2010, 2013, 2014; ZHAO et al., 2013).

Neste contexto, o presente trabalho teve como objetivo avaliar métodos de seleção genômica aplicados ao painel de sorgo biomassa da Embrapa/Milho e Sorgo para produção de etanol de segunda geração, visando auxiliar o programa de melhoramento da espécie por meio da seleção precoce de genótipos superiores. Além disso, objetivou investigar o potencial uso da seleção genômica para elucidação da arquitetura genética de caracteres quantitativos, por meio de análises de enriquecimento funcional.

2 REVISÃO BIBLIOGRÁFICA

2.1 Sorgo

O sorgo (*Sorghum bicolor* (L.) Moench) é uma gramínea nativa da África (WET; HUCKABAY, 1967), que se destaca como o quinto cereal mais produzido no mundo, precedido apenas do milho, arroz, trigo e cevada. Os principais produtores mundiais são os Estados Unidos, o México e a Nigéria (FAO, 2015). No Brasil, as previsões para a cultura do sorgo na safra 2014/2015 são de 751 mil ha de área plantada, com produção de 2,04 milhões de toneladas (COMPANHIA NACIONAL DE ABASTECIMENTO – CONAB, 2015).

Apresenta-se como uma espécie facilmente adaptada a ambientes extremos de estresses abióticos, especialmente de temperatura e umidade. Esta rusticidade confere ao sorgo condições favoráveis à sua adaptação em relação a outras espécies comerciais, sendo, portanto, uma alternativa viável para plantio em áreas de cultivo marginais (ROONEY et al., 2007).

Botanicamente, o sorgo é uma monocotiledônea pertencente à família *Poaceae*, tribo Andropogoneae, gênero *Sorghum*. É designado como uma planta com metabolismo C4, autógama, com propagação via sementes e ciclo de vida de curta duração, de aproximadamente quatro meses (DOGETT, 1988).

Agronomicamente, o sorgo é classificado em diferentes tipos de acordo com seus usos e finalidades. Os tipos mais conhecidos são: o granífero, o sacarino, o forrageiro e o vassoura. O sorgo granífero é utilizado na alimentação humana em países asiáticos e africanos; já em países desenvolvidos é utilizado como substituto do milho na formulação de rações para animais. O sorgo sacarino apresenta colmos com caldo semelhante ao da cana-de-açúcar, rico em açúcares fermentescíveis, podendo ser utilizado para a produção de etanol com a utilização da mesma estrutura de colheita e fermentação utilizada em cana-de-açúcar. O sorgo forrageiro apresenta alta qualidade e versatilidade para a produção de ração animal, podendo ser utilizado na forma de feno, silagem e pastejo direto. O sorgo vassoura é utilizado para a produção de vassouras (SCHIMIDT, 1987; EMBRAPA, 2008).

2.1.1 Sorgo Biomassa

A crescente preocupação mundial com a disponibilidade de recursos energéticos fez surgir a necessidade de se pensar em fontes alternativas para a produção de energia,

especificamente, os biocombustíveis gerados a partir de biomassa. Algumas espécies são listadas como potenciais candidatas a se tornarem culturas bioenergéticas, como é o caso do milho (*Zea mays*), *switchgrass* (*Panicum virgatum*), miscanto (*Miscanthus giganteus*), cana-de-açúcar (*Saccharum* spp.), a espécie arbórea álamo (*Populus trichocarpa*) e o sorgo (*Sorghum bicolor* L. Moench) (ROONEY et al., 2007; MORRIS et al., 2013). Acredita-se que o próximo aumento substancial na produção de biocombustíveis seja baseado na conversão de biomassa celulósica e na produção de biocombustíveis a partir de algas (OLSON et al., 2012).

A utilização de culturas, até então dedicadas à alimentação humana e animal, para a geração de biocombustíveis, como é o caso do milho e da cana-de-açúcar, gera insatisfação. Isto ocorre devido ao fato de o aumento na demanda da cultura refletir em aumento dos preços dos alimentos e das rações animais. Neste contexto, a produção de biomassa para fins bioenergéticos não deve competir com a produção de alimentos. Além disso, culturas dedicadas à produção de biocombustíveis podem potencialmente ser produzidas em regiões marginais de cultivo, onde a produção de alimentos não pode ser realizada de maneira comercialmente viável, minimizando assim a competição entre produção de alimentos e energia (ANTONOPOULOU et al., 2007).

Há cerca de 30 anos o sorgo foi proposto como uma cultura dedicada à bioenergia, quando, após a interrupção do fornecimento de petróleo na década de 1970, foi desenvolvida a primeira geração de sorgo de alta energia ou sorgo biomassa, uma promissora cultura especificamente desenvolvida para produção bioenergética, a qual combina características do sorgo granífero e sacarino (MONK et al., 1984; HONS et al., 1986). No entanto, apenas recentemente avanços satisfatórios no melhoramento do sorgo biomassa ocorreram, os quais foram impulsionados pela caracterização genética de um sistema regulatório responsável pelo controle do fotoperiodismo e do tempo de florescimento. Usando esse sistema, híbridos de sorgo biomassa altamente sensíveis ao fotoperíodo e com florescimento tardio têm sido desenvolvidos, exibindo prolongado crescimento vegetativo e alta acumulação de biomassa (ROONEY; AYDIN, 1999; ROONEY et al., 2007; MURPHY et al., 2011).

O tipo ideal de sorgo para produção de biocombustíveis é altamente dependente do processo de conversão que irá ser utilizado, visto que híbridos de sorgo granífero fornecem amido para conversão, enquanto o sorgo sacarino acumula açúcares no colmo. Cada tipo de sorgo tem um sistema diferente de produção e nenhuma outra espécie tem a flexibilidade de produzir grandes quantidades de amido, açúcar ou celulose (ROONEY et al., 2007).

Celulose, hemicelulose e lignina, principais componentes químicos do sorgo biomassa, são as mais abundantes fontes renováveis de energia da Terra e podem ser utilizadas como

alternativa para o uso do petróleo e outros recursos fósseis. Atualmente, o melhoramento deste tipo de sorgo visa à produção de biomassa diferenciada para características lignocelulósicas para fins energéticos, com foco em etanol de segunda geração, dado que a dissociação dos polímeros presentes na matéria lignocelulósica é um dos principais entraves para a conversão da biomassa em açúcares fermentescíveis (HABYARIMANA et al., 2004; MURRAY et al., 2008).

Neste contexto, os programas de melhoramento genético buscam estratégias que os auxiliem a promover alterações no conteúdo e na composição da lignina. Tais esforços baseiam-se no fato de que, para a conversão da biomassa em etanol de segunda geração, quanto menor o conteúdo de lignina ou diferenciação na sua composição, mais eficiente o processo de conversão. Já para a cogeração de energia, o aumento do conteúdo de lignina possibilita o aumento do poder calorífico da biomassa e, conseqüentemente, maior produção de energia (EMBRAPA, 2011).

2.2 Melhoramento Genético do Sorgo Biomassa

O melhoramento genético é um processo que procura alterar geneticamente as plantas de modo a atender às necessidades humanas. O aumento da população mundial implica na necessidade crescente de uma maior produção de alimentos e recursos energéticos, sendo o melhoramento genético de plantas uma importante ferramenta utilizada para aumento de produtividade através do desenvolvimento de cultivares mais adaptadas, resistentes a pragas e doenças, entre outras características de interesse. O sucesso de um programa de melhoramento está relacionado à sua capacidade de seleção de indivíduos superiores, ou seja, depende da habilidade em prever os valores genéticos a partir dos valores fenotípicos e/ou dados genotípicos (BUENO; MENDES; CARVALHO, 2006; BERNARDO, 2014a).

O melhoramento genético na cultura do sorgo iniciou-se através da manipulação da enorme variabilidade existente na espécie, primeiramente objetivando-se a produção de grãos, xarope e forragem. A descoberta da macho esterilidade citoplasmática em sorgo granífero permitiu que, em meados da década de 1950, fossem desenvolvidos os primeiros híbridos de sorgo, os quais contribuíram de forma efetiva para o aumento da produtividade (ASSEFA; STAGGENBORG, 2010). Nos Estados Unidos o aumento da produtividade média de sorgo granífero foi de 0,7 Mg ha⁻¹ no início da década de 90 para mais de 3,5 Mg ha⁻¹ no início do século 21 (ASSEFA; STAGGENBORG, 2010). Vários fatores agrônômicos e genéticos

contribuíram para este aumento, sendo que, segundo Duvick (1999), aproximadamente 35 a 40% deste ganho foi devido ao desenvolvimento de híbridos.

Quando o sorgo biomassa emergiu na década de 1980, como uma cultura dedicada à produção bioenergética, os programas de melhoramento de sorgo granífero já contavam com uma experiência de mais de 30 anos na produção de híbridos e acumulavam conhecimentos genéticos de características desejáveis, contribuindo, portanto, para o avanço do melhoramento de sorgo biomassa. Sendo assim, a proposta do sorgo biomassa ser um híbrido que combina características de sorgo granífero e sacarino surgiu como uma maneira lógica de aumentar o potencial de produção através da utilização da heterose para explorar o desenvolvimento vegetativo e fenológico (MONK et al., 1984). Entretanto, o melhoramento de sorgo biomassa permaneceu inativo por um período e apenas recentemente, por volta dos anos 2000, foi reativado, impulsionado pelas crescentes preocupações com a demanda energética mundial (REDDY et al., 2004).

Atualmente, estão entre os principais focos dos programas de melhoramento voltados à produção bioenergética a maximização da produção de biomassa por unidade de área plantada, visando minimizar o uso de área cultivável, além de alterações nas proporções e estruturas dos vários polímeros que constituem a parede celular, uma vez que um dos principais problemas para a conversão de biomassa para açúcares fermentescíveis é a dissociação dos vários polímeros presentes na matéria lignocelulósica. De modo geral, o melhoramento de plantas tem sido implementado visando aumentar a produtividade, qualidade e eficiência de conversão da biomassa, seja através da seleção entre progênies obtidas do cruzamento de parentais com características desejáveis, ou através de transgênicos e mutantes (VERMERRIS et al., 2007).

Neste contexto, os programas de melhoramento genético de sorgo biomassa podem ser muito beneficiados pelo desenvolvimento de métodos que reduzam o tempo gasto para completar um ciclo de melhoramento e possibilitem uma seleção mais precoce e eficiente de genótipos superiores, entre os quais a seleção genômica tem grande potencial de utilização.

2.3 Recursos Genéticos e Genômicos Disponíveis

O sorgo é uma espécie diploide ($2n = 20$), para a qual já está disponível uma montagem de referência do genoma, cujo tamanho é relativamente pequeno, em torno de 700 Mpb, tornando-o uma espécie muito indicada como modelo para estudos de análises genômicas em espécies geneticamente próximas, como a cana-de-açúcar e o milho

(PATERSON et al., 2009). O seu genoma está anotado, com modelos gênicos preditos, aos quais também foram atribuídas funções putativas ou hipotéticas, isto é, preditas computacionalmente. Além disso, é possível encontrar na literatura científica QTLs (*quantitative trait loci*, ou locos que controlam caracteres quantitativos) já mapeados para o sorgo.

Existem atualmente 136 QTLs mapeados para diferentes caracteres agronômicos de interesse, entre eles o comprimento de perfilho basal, número de perfilhos basais, comprimento de colmo, ângulo da folha e largura da inflorescência (YOUENS-CLARK et al., 2010; PATERSON, 2008; PATERSON et al., 2009). Esta disponibilidade de recursos e informações facilita os esforços de melhoramento da espécie, visto que é possível corroborar resultados prévios e fazer uso de anotações do genoma para, por exemplo, fazer inferências a respeito de potenciais genes candidatos.

2.4 Genotipagem por Sequenciamento

A estratégia de genotipagem por sequenciamento, ou GBS (*genotyping-by-sequencing*), é um sistema *multiplex* para genotipagem do genoma, o qual permite a identificação de milhares de polimorfismos de nucleotídeo único (*single nucleotide polymorphisms*, ou SNPs), bem como pequenas inserções e deleções (*indels*), a custos reduzidos (POLAND; RIFE, 2012). A flexibilidade do sistema GBS, no que diz respeito às espécies, populações e objetivos da pesquisa aos quais pode ser aplicado, torna-o uma ferramenta bastante atrativa para estudos genéticos em plantas, como, por exemplo, estudos de diversidade genética e de genética de populações, construção de mapas genéticos, identificação de QTLs, mapeamento associativo ou GWAS (*genome wide association study*), estudos da complexidade genômica, determinação do nível de ploidia, caracterização de germoplasmas e seleção genômica (DESCHAMPS et al., 2012; POLAND et al., 2012; CROSSA et al., 2013; LU et al., 2013; SONAH et al., 2013; WARD et al., 2013).

Primeiramente desenvolvida pelo IGD (*Institute for Genomic Diversity*), da Universidade de Cornell, a técnica genericamente denominada de GBS consiste na construção de bibliotecas genômicas com base na redução da complexidade do genoma, através do uso de enzimas de restrição sensíveis à metilação, as quais reduzem a representatividade de regiões repetitivas do genoma. Os fragmentos selecionados resultantes são submetidos ao sequenciamento para a geração de dezenas de milhões de sequências curtas em plataformas de sequenciamento de nova geração (*next generation sequencing*, ou NGS), com destaque para a

plataforma Illumina. Após a redução de complexidade via enzimas de restrição e antes do sequenciamento, as amostras de DNA a serem genotipadas recebem *barcodes*, que são adaptadores com sequências indexadoras, os quais permitem que mais tarde as sequências geradas para cada amostra sejam rastreadas. Desta forma, até 96 amostras podem ser sequenciadas conjuntamente em cada canaleta de sequenciamento, o que reduz substancialmente os custos. Finalmente, através da aplicação de rotinas de bioinformática, são detectados polimorfismos de presença ou ausência entre amostras derivados da variabilidade na distribuição dos sítios de restrição, bem como SNPs entre sequências em comum entre as amostras (ELSHIRE; GLAUBITZ; SUN, 2011).

O objetivo principal da técnica GBS é simultaneamente descobrir polimorfismo e obter informações genótípicas de toda a população de interesse. Esta combinação faz do sistema GBS uma técnica rápida e flexível, a qual já encontra extensa aplicação em uma ampla diversidade de espécies, como, por exemplo, cevada, trigo, milho, arroz, uva, cacau, sorgo, feijão e soja, sendo que pesquisas em muitas outras espécies estão em progresso (ELSHIRE; GLAUBITZ; SUN, 2011; HANSEY et al., 2012; POLAND et al., 2012; MORRIS et al., 2013; LU et al., 2013; SONAH et al., 2013). Além disso, existem rotinas de bioinformática (*pipelines*) específicas para permitir a aplicação do GBS em espécies que possuam, ou não, uma montagem de referência do genoma. Quando a espécie em estudo apresenta um genoma de referência disponível, a obtenção de SNPs pode ser realizada alinhando-se os fragmentos sequenciados a tal montagem de referência do genoma, aplicando-se assim o chamado *Discovery Pipeline*. Já para espécies que não possuem o genoma sequenciado foi proposta uma abordagem denominada de *Universal Network Enabled Analysis Kit*, ou UNEAK Pipeline (LU et al., 2013).

Neste contexto, dentre as principais vantagens da técnica de GBS destacam-se: flexibilidade, simplicidade, rapidez, eficiente sistema de *barcodes*, baixo custo por unidade de observação, aplicabilidade em espécies que não possuam montagem de referência do genoma, *pipelines* de código-fonte aberto e, principalmente, a possibilidade de realização simultânea da descoberta de polimorfismo e sua aplicação no conjunto de dados de interesse, o que combina as duas principais etapas da genotipagem em um único passo (ELSHIRE; GLAUBITZ; SUN, 2011; POLAND; RIFE, 2012; DE DONATO et al., 2013). Segundo Poland e Rife (2012), umas das aplicações mais promissoras da metodologia de genotipagem por sequenciamento será na área do melhoramento de plantas, fornecendo uma ferramenta rápida e de baixo custo para a genotipagem, permitindo, assim, que a seleção genômica seja implementada em larga escala em programas de melhoramento.

No entanto, a elevada quantidade de dados perdidos, devido à baixa cobertura do sequenciamento e à distribuição não uniforme das leituras, são algumas desvantagens da técnica (DAVEY et al., 2011; BEISSINGER et al., 2013). Porém, é importante destacar que a quantidade de dados perdidos só se torna relevante dependendo do objetivo do estudo realizado. Nos casos em que esta efetivamente representa um problema, a imputação de dados mostra-se uma alternativa viável para solucioná-lo. Por exemplo, o uso da imputação tem aumentado o poder de estudos de mapeamento associativo (MARCHINI et al., 2007; MARCHINI; HOWIE, 2010), e em estudos de seleção genômica a imputação permite o uso de genotipagem de baixa densidade sem maiores perdas na capacidade preditiva (HABIER et al., 2009; WEIGEL et al., 2010; DASSONNEVILLE et al., 2011; MULDER et al., 2012).

2.5 Desequilíbrio de Ligação

Desequilíbrio de ligação (DL), também denominado desequilíbrio da fase gamética, desequilíbrio gamético e associação alélica, pode ser definido como a associação não aleatória, ou preferencial, entre alelos de diferentes locos. O interesse em estudar o desequilíbrio de ligação em plantas aumentou recentemente de forma acentuada, devido, principalmente, ao avanço das tecnologias genômicas, as quais possibilitam a rápida identificação de polimorfismos em todo o genoma, permitindo, assim, a utilização do DL em uma ampla variedade de pesquisas (RAFALSKI; MORGANTE, 2004; GUPTA et al., 2005). No contexto da seleção genômica é essencial a existência de DL, pois este garante que os efeitos estimados dos marcadores sejam consistentes entre indivíduos da população de interesse. Deste modo, este é um dos fatores responsáveis por alterar a acurácia dos modelos, o que torna essencial o entendimento do DL antes de a seleção genômica ser aplicada.

Os termos ligação e desequilíbrio de ligação são frequentemente utilizados como sinônimos, mas, embora relacionados, são termos distintos. Ligação refere-se à herança correlacionada de locos através da ligação física em um cromossomo, enquanto desequilíbrio de ligação refere-se à correlação entre alelos em uma população. A confusão acontece, pois locos intimamente ligados podem estar em elevado DL. Porém, deve ficar claro que o DL pode ocorrer não apenas entre alelos de dois ou mais locos fisicamente ligados, presentes no mesmo cromossomo ou haplótipo, mas também entre alelos de dois ou mais locos que não estão fisicamente ligados, ou seja, os quais estão muito distantes em um cromossomo ou mesmo em cromossomos diferentes (FLINT-GARCIA; THORNSBERRY, BUCKLER, 2003).

Vários fatores relacionados à genética de populações, os quais alteram a frequência alélica e/ou geram recombinação, afetam os padrões de DL. Como exemplos, citam-se: mutação, seleção, deriva genética, tamanho populacional e sistema reprodutivo. Processos de mutação são responsáveis por gerar polimorfismos, os quais podem culminar em DL. Seleção natural e seleção artificial ambas alteram as frequências alélicas, podendo gerar DL entre locos ligados e não ligados. A redução brusca de tamanho de uma população, fenômeno conhecido como “gargalo-de-garrafa” ou *bottleneck*, seguido por deriva genética, provoca a perda de alelos e haplótipos, alterando suas frequências, o que por sua vez gera associações preferencias entre alelos. O DL é proporcional à fração de recombinação, portanto, visto que em espécies autógamas a taxa de recombinação é usualmente menor, extensos blocos de DL são criados quando comparados a espécies alógamas (FLINT-GARCIA; THORNSBERRY, BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; ZHONG et al.; 2009).

Gupta, Rustgi e Kulwal (2005) relataram, de forma revisada, estimativas da extensão do DL em diferentes espécies de plantas, sendo este inferior a 3 cM em beterraba sacarina, 500 pb em uva, 0,4-7 kpb em milho e 10 cM em cana-de-açúcar, as quais são espécies alógamas. No entanto, para espécies autógamas este intervalo é de 50 kpb em soja, 100 kpb em arroz, 250 kpb em *Arabidopsis thaliana*, 200 kpb em alface, 10-20 cM em cevada e 0,3-1 cM em batata. No caso do sorgo, segundo Hamblin et al. (2004), este intervalo é maior ou igual a 10 kpb.

Para o sucesso da seleção genômica é essencial que pelo menos um marcador esteja em forte DL com determinado QTL de interesse; caso contrário, espera-se que a acurácia decaia rapidamente nas gerações seguintes à estimação dos efeitos dos marcadores. Porém, mesmo na ausência de DL entre marcadores e QTL, a acurácia dos valores genéticos genômicos estimados pode ser superior a zero, dado que o efeito dos marcadores também pode capturar relações de parentesco genético. Deste modo, os métodos para estimar os efeitos dos marcadores utilizam informações de parentesco genético entre indivíduos, bem como informações de DL, sendo variável a forma como cada método utiliza essas informações (HABIER; FERNANDO; DEKKERS, 2007; ZHONG et al., 2007; JANNINK; LORENZ; IWATA, 2010). Nos casos em que o efeito dos marcadores captura informações de parentesco genético, a acurácia dos efeitos genéticos genômicos estimados decai mais rapidamente depois da estimação, em decorrência do decaimento do parentesco genético em cada geração. Em contrapartida, informações do DL são mais persistentes, fato que acentua sua importância em estudos de seleção genômica.

Em estudos de simulação, Habier, Fernando e Dekkers (2007) mostraram que o modelo RR-BLUP foi mais efetivo em capturar informações de parentesco genético, pois ajustou um maior número de marcas no modelo preditivo. De modo contrário, o modelo Bayes-B, o qual permite que muitas marcas tenham efeito igual a zero, conforme será detalhado posteriormente, foi mais efetivo em capturar DL entre marcadores e QTLs. Consequentemente, a acurácia do modelo Bayesiano permaneceu inalterada por mais gerações quando comparado à acurácia do modelo RR-BLUP. Nas condições simuladas, 61% e 79% das acurácias obtidas foram devido ao DL e 39% e 21% foram devido à captura de efeitos genéticos, para os modelos RR-BLUP e Bayes-B, respectivamente.

Além disso, a densidade de marcadores necessária para seleção genômica é reflexo da extensão do DL da população em questão. Quanto maiores os blocos de DL, menor a quantidade de marcadores necessária para garantir que pelo menos uma marca esteja em DL com o QTL de interesse (NAKAYA; ISOBE, 2012). Estudos realizados em cevada por Zhong et al. (2009), evidenciaram que existe uma clara interação entre a extensão do DL e a densidade de marcadores. Os resultados encontrados mostraram que, em condições de alta densidade de marcadores, as capacidades preditivas em delineamento com baixo DL foram apenas marginalmente inferiores àsquelas de delineamento com alto DL. No entanto, em condições de baixa densidade de marcadores, essa diferença foi considerável.

2.6 Seleção Genômica

A seleção genômica, também conhecida como *Genome-Wide Selection* (GWS), foi inicialmente proposta em 2001 por Meuwissen e colaboradores, como uma forma de aumentar a eficiência da seleção assistida por marcadores e acelerar o ciclo do melhoramento genético, através do uso simultâneo de marcadores genéticos cobrindo todo o genoma. Este método permaneceu discreto por alguns anos, devido ao fato de os marcadores moleculares disponíveis na época serem de número limitado e obtidos a custos elevados. Porém, com o surgimento das tecnologias de sequenciamento de nova geração e a possibilidade da descoberta de um grande número de marcadores do tipo SNP, a custos cada vez mais reduzidos, o interesse científico a respeito da seleção genômica foi retomado (BERNARDO; YU, 2007; DEKKERS, 2007; GODDARD; HAYES, 2007; HABIER; FERNANDO; DEKKERS, 2007).

Grande parte dos caracteres de importância agrônômica, como produtividade, qualidade e algumas formas de resistências a doenças, são controlados por muitos genes,

sendo, portanto, conhecidos como caracteres quantitativos ou poligênicos (COLLARD et al., 2005). As regiões do genoma responsáveis pela expressão de caracteres poligênicos são designadas por *quantitative trait loci* (QTLs) (PATERSON et al., 1988; LIU, 1998). Com o advento dos marcadores moleculares, tornou-se possível mapear tais regiões cromossômicas que afetam esses caracteres quantitativos. Subsequentemente, foi proposta a seleção assistida por marcadores como uma maneira de explorar marcadores ligados a QTLs de interesse, para auxiliar o trabalho de seleção dos melhoristas (BERNARDO, 2008; LORENZ et al., 2011) (BERNARDO, 2008; LORENZ et al., 2011). O uso de informações diretamente em nível de DNA prometia reformular os programas de melhoramento genético e promover ganhos de seleção mais rápidos, quando comparados à seleção baseada apenas em dados fenotípicos. No entanto, apesar do intenso trabalho de pesquisa, na prática a implementação da seleção assistida mostrou-se limitada e os ganhos genéticos pequenos (DEKKERS, 2004). Dentre as principais limitações apresentadas pela seleção assistida por marcadores estão o fato de o mapeamento de populações biparentais usados na maioria dos estudos de QTL não ter aplicação direta em programas de melhoramento, além de os métodos estatísticos usados para identificar locos de interesse e realizar a seleção assistida por marcadores serem inadequados para o melhoramento de caracteres poligênicos controlados por muitos locos de pequeno efeito (MEUWISSEN et al., 2001, GODDARD; HAYES, 2007; HEFFNER; SORRELLS; JANNINK, 2009).

Neste contexto, a seleção genômica surgiu como uma maneira de contornar essas limitações. O ponto chave deste método é a predição simultânea dos efeitos genéticos de milhares de marcadores dispersos em todo o genoma, de forma a capturar os efeitos de todos os locos, de pequenos e de grandes efeitos, para explicar o máximo possível da variação genética de um caráter quantitativo. Assim como no mapeamento de QTLs, a condição fundamental para tal é que haja desequilíbrio de ligação (ZHONG et al., 2009; NAKAYA; ISOBE, 2012). Além disso, o fato de serem consideradas informações de todos os marcadores do genoma, e não apenas daqueles associados ao caráter por meio da análise de QTLs, elimina as etapas de construção de mapa genético e mapeamento de QTLs, o que torna o processo de seleção genômica mais rápido e barato (BERNARDO, 2008; LORENZ et al., 2011).

A seleção genômica foi primeiramente empregada no melhoramento animal, em que o alto custo de fenotipagem e a impossibilidade de fazer uso de repetições de indivíduos tornaram-na atrativa. Vários estudos de simulação relataram a sua eficiência e levantaram questões a ela relacionadas (DEKKERS, 2007; LONG et al., 2007). Posteriormente, estudos empíricos comprovaram as expectativas teóricas do sucesso da seleção genômica no

melhoramento animal (LEE et al., 2008; LEGARRA et al., 2008; VANRADEN et al., 2009), atraindo também a atenção de melhoristas de plantas (BERNADOR; YU, 2007; HEFFNER; SORRELLS; JANNINK, 2009; JANNINK; LORENZ; IWATA, 2010). Assim, estudos de simulação em diversas culturas vêm sendo realizados, os quais têm comprovado a superioridade em termos de acurácia da seleção genômica sobre a seleção assistida por marcadores tradicional e a seleção baseada apenas em informações de parentesco (BERNARD; YU, 2007; BERNARD, 2009, 2014b; ZHONG et al., 2009; LORENZANA; BERNARD, 2009; MAYOR; BERNARD, 2009; GRATTAPAGLIA; RESENDE, 2010; CROSSA et al., 2010, 2013, 2014; ZHAO et al., 2013).

Até o momento, grande parte das publicações que envolvem seleção genômica em plantas é baseada em dados simulados. No entanto, nos últimos anos as avaliações empíricas têm se tornado mais importantes (MASSMAN; JUNG; BERNARD, 2013). Para que a seleção genômica se torne realidade em um programa de melhoramento de plantas, muitas decisões devem ser tomadas, como, por exemplo, a escolha do modelo estatístico, o tamanho da população de treinamento, os tipos de marcadores, a densidade de marcadores, sua integração às práticas corriqueiras do programa de melhoramento, dentre outras. Nos últimos anos vários estudos têm produzido valiosas informações que podem ser usadas para auxiliar o melhorista na tomada dessas decisões. No entanto, ainda há muito a ser explorado no contexto da seleção genômica, como, por exemplo, lidar com a interação genótipo por ambiente, o uso de índices de seleção e os tipos de população nas quais esta pode ser aplicada, como por exemplo, a predição entre painéis de genótipos em fases iniciais de programas de melhoramento. O fato de ser um tema recente, o qual tem o potencial de revolucionar o uso das informações de marcadores moleculares em programas de melhoramento genético, torna a seleção genômica um empolgante campo de estudo (KONING; SWALVE, 2009; JONAS; KONING, 2013; DE LOS CAMPOS et al., 2013; DESTA; ORTIZ, 2014; HESLOT; JANNINK; SORRELLS, 2015).

2.6.1 Populações de Treinamento, Teste e Seleção

Para que a seleção genômica seja realizada, é usual que três populações sejam definidas: populações de treinamento, teste e seleção, as quais serão detalhadas a seguir.

A estimação dos efeitos dos marcadores é realizada com base em dados genotípicos e fenotípicos de indivíduos pertencentes a uma amostra da população de trabalho, denominada população de treinamento. A população de treinamento usualmente contempla um grande

número de marcadores avaliados em diversos indivíduos, cujo número depende da acurácia desejada, os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Equações de estimação de valores genéticos genômicos são obtidas para cada caráter de interesse. Os marcadores associados aos locos que controlam esses caracteres são detectados, bem como são estimados os seus efeitos (GODDARD; HAYES, 2007; KOLBEHDARI; SCHAEFFER; ROBINSON, 2007; LONG et al., 2007; LEGARRA; MISZTAL, 2008).

Em seguida, a capacidade preditiva do modelo estatístico, contendo os efeitos dos marcadores, é verificada pelo ajuste em uma população de validação, também denominada população de teste, para a qual também há disponibilidade de dados genotípicos e fenotípicos. Nessa etapa, são estimados os valores genéticos com base apenas em dados genotípicos (*Genomic Estimated Breeding Values*, GEBVs), utilizando o modelo estatístico ajustado na população de treinamento.

De modo geral, utiliza-se a técnica da validação cruzada (*cross-validation*) para a composição das populações de treinamento e validação. O procedimento baseia-se no particionamento do conjunto de dados em k subconjuntos, em geral mutuamente exclusivos, dos quais $k - 1$ são utilizados em cada iteração do método para a estimação dos parâmetros do modelo e correspondem à população de treinamento. O subconjunto restante é empregado na avaliação da capacidade preditiva do modelo e corresponde à população de teste ou validação (KOHAVI, 1995). A validação cruzada permite que todos os indivíduos sejam usados tanto para treino quanto para teste, visto que em cada uma das k instâncias de validação os indivíduos de um determinado subconjunto são os componentes da população de teste. Através da validação cruzada, a avaliação da capacidade preditiva de cada modelo é realizada por meio da correlação entre a estimativa dos valores genéticos com base em dados genotípicos (GEBVs) e os valores genéticos com base em observações fenotípicas (*Estimated Breeding Values*, EBVs), $r(GEBV:EBV)$ (DE LOS CAMPOS et al., 2009; RESENDE et al., 2010).

Caso as correlações entre os GEBVs e os EBVs sejam altas, o modelo ajustado na população de treinamento pode ser utilizado por melhoristas para a predição dos valores genéticos de indivíduos da população de seleção, a qual é constituída por indivíduos para os quais há apenas informações genotípicas, mas que compartilhem a base genética da população de treinamento (JANNINK; LORENZ; IWATA, 2010; CROSSA et al., 2011).

2.6.2 Acurácia da Seleção Genômica

A acurácia dos valores genéticos genômicos estimados (GEBVs) é avaliada através da correlação destes com os valores genéticos estimados com base em observações fenotípicas (*Estimated Breeding Values*, EBVs), denotada por $r(GEBV:EBV)$. Os EBVs podem ser obtidos de diversas formas, utilizando-se, por exemplo, modelos mistos para a obtenção de predições lineares não-viesadas (*Best Linear Unbiased Predictions*, BLUPs), ou apenas através da média fenotípica. Esta correlação proporciona uma estimativa da acurácia seletiva e, portanto, relaciona a acurácia do GEBV com a resposta de seleção (LORENZ et al., 2011). A acurácia da seleção genômica é definida como a correlação entre GEBVs e o verdadeiro valor genético (*True Breeding Values*, TBVs), denotada por $r(GEBV:TBV)$. No entanto, na prática não é possível medir os TBVs, logo, uma estimativa da acurácia, denominada capacidade preditiva, é realizada através da $r(GEBV:EBV)$ (LORENZ et al., 2011).

A resposta da seleção genômica é resultado dos vários fatores que afetam a acurácia dos GEBVs. Estes fatores são inter-relacionados e podem ou não estar sob o controle do melhorista. Entre os fatores que não podem ser controlados estão a herdabilidade do caráter, a arquitetura genética e a extensão do desequilíbrio de ligação. No entanto, pode-se ter controle sobre a densidade de marcadores, número de indivíduos da população de treinamento, parentesco genético entre as populações de treinamento e teste e o modelo de seleção genômica utilizado. Entre esses últimos fatores, o tamanho da população de treinamento e o parentesco genético entre as populações de treinamento e teste são os fatores mais importantes que afetam a acurácia. O modelo de seleção genômica utilizado também é importante, pois o número de locos que controlam o caráter e a distribuição de seus efeitos são tratados de formas diferentes em cada modelo. Consequentemente, quanto melhor o modelo utilizado se encaixar à arquitetura genética do caráter, maior será a acurácia (LORENZ et al., 2011; DE LOS CAMPOS et al., 2013; DESTA; ORTIZ, 2014).

2.6.3 Modelos Estatísticos

A seleção genômica foi proposta como uma forma de explorar altas densidades de marcadores cobrindo todo o genoma para estimar valores genéticos de caracteres quantitativos (MEUWISSEN; HAYES; GODDARD, 2001). Porém, como consequência disto, o número p de efeitos a serem preditos é comumente muito superior ao número n de indivíduos avaliados.

Tal situação é conhecida como o problema do “*large p, small n*”. A abordagem dos quadrados mínimos ordinários para ajuste do modelo trata os efeitos dos marcadores como fixos; portanto, como o número de marcas excede o número de indivíduos, não existem graus de liberdade suficientes para estimar os efeitos de todos os marcadores simultaneamente, devido à existência de multicolinearidade (JANNINK; LORENZ; IWATA, 2010; LORENZ et al., 2011).

Para solucionar este problema, vários modelos estatísticos têm sido propostos para aplicação em seleção genômica. A principal distinção entre tais modelos é a suposição a respeito da distribuição subjacente dos efeitos dos QTLs. Devido a particularidades da arquitetura genética de caracteres quantitativos, isto é, diferenças no número de QTLs, suas posições no genoma, magnitude dos efeitos, interações alélicas (dominância) e entre locos (epistasia) (JIANG; ZENG, 1995; ZENG et al., 1999; ZENG, 2001), diferentes modelos de seleção genômica podem ser adequados para caracteres fenotípicos distintos. Assim, é importante ajustar e comparar diversos modelos estatísticos para cada caráter isoladamente, de modo a se escolher o mais apropriado a cada um, para posterior seleção de genótipos sem avaliação fenotípica.

A seguir, encontram-se descrições dos vários modelos para seleção genômica utilizados na presente dissertação.

2.6.3.1 Modelo RR-BLUP/G-BLUP

O método RR-BLUP regressão aleatória (*random regression*) do tipo BLUP, calcula os valores genéticos dos indivíduos através de preditores lineares não enviesados, ou BLUP (HENDERSON, 1984). A técnica da regressão aleatória é um tipo especial de regressão de cuneeira (*Ridge Regression*). Os estimadores associados à regressão aleatória e regressão de cuneeira promovem *shrinkage*, ou encolhimento, cuja magnitude é ditada por uma função da quantidade do parâmetro de penalização λ .

O modelo estatístico para realização de seleção genômica, neste caso, é definido por:

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{S}_1\mathbf{g}_1 + \mathbf{S}_2\mathbf{g}_2 + \cdots + \mathbf{S}_p\mathbf{g}_p + \mathbf{e} \\ &= \mathbf{1}\mu + \mathbf{S}\mathbf{g} + \mathbf{e}, \end{aligned} \quad (1)$$

em que $\mathbf{y}$ é o vetor de observações fenotípicas; $\mathbf{1}$ é um vetor de 1's; μ é a média geral do caráter; $\mathbf{S}_j$ contém os genótipos codificados do SNP j , para $j = 1, \dots, p$, em que p é o número total de SNPs no modelo; $\mathbf{g}_j$ é o efeito de substituição alélica do SNP j , para $j = 1, \dots, p$; $\mathbf{e}$ é

um vetor de erros, com $\mathbf{e} \sim N(0, \sigma_e^2)$; $\mathbf{S}$ é uma matriz em que as colunas contêm os genótipos $\mathbf{S}_j$; e $\mathbf{g}$ é um vetor com os efeitos dos marcadores. Os efeitos dos SNPs são considerados efeitos aleatórios, com distribuição normal e variância comum, isto é, a variância explicada por cada SNP é dada por $\sigma_j^2 = \sigma_g^2/p$, em que σ_g^2 representa a variação genética total. Portanto, tem-se que $\mathbf{g}_j \sim N(0, \sigma_g^2/p)$ (MEUWISSEN; HAYES; GODDARD, 2001, 2013). Este modelo mais simples supõe que cada QTL explique uma porção igual da variância genética.

Tal modelo pode ser reescrito como:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{e},$$

em que $\mathbf{u} = \mathbf{S}_1\mathbf{g}_1 + \mathbf{S}_2\mathbf{g}_2 + \dots + \mathbf{S}_p\mathbf{g}_p$, com $\mathbf{u} \sim N(0, \sigma_g^2\mathbf{SS}'/p)$. A matriz $\mathbf{SS}'/p$ representa uma matriz de parentesco genômico, comumente denotada por $\mathbf{G}$, de modo que este modelo recebe o nome de G-BLUP (GODDARD; HAYES, 2009; MEUWISSEN; HAYES; GODDARD, 2013).

Devido à inclusão dos marcadores no modelo como efeitos aleatórios, o modelo misto resultante exige a estimação de apenas um parâmetro referente aos efeitos dos marcadores, ou seja, a variância genética σ_g^2 . Por esta razão, os BLUPs dos efeitos dos marcadores podem ser calculados mesmo que o número de SNPs seja maior do que o número de indivíduos, o que geralmente ocorre para aplicações de seleção genômica.

Os valores genéticos genômicos dos indivíduos, ou GEBVs, podem ser obtidos por:

$$GEBV_i = \sum_{j=1}^p S_{ij}\hat{g}_j,$$

em que S_{ij} representa o genótipo do indivíduo i para o SNP j .

2.6.3.2 Família de Modelos Bayesianos

A suposição de que todos os efeitos de marcadores originam-se de uma mesma distribuição normal, com variância comum, pode não ser realista na prática, pois muitos dos efeitos estimados são muito próximos de zero. Na realidade, há evidências de que alguns QTLs possuem efeitos marcadamente elevados (LORENZ; COHEN, 2012), de modo que SNPs ligados a esses locos devem ter efeito mais pronunciado no modelo de seleção genômica (MEUWISSEN; HAYES; GODDARD, 2013). Em contrapartida, é possível que centenas a milhares de QTLs possuam efeito reduzido, assim como é também possível que alguns SNPs não possuam efeito algum no modelo.

Para contornar as limitações do modelo RR-BLUP/G-BLUP, foram propostos diversos modelos Bayesianos, os quais fazem outras suposições a respeito da distribuição *a priori* dos efeitos dos QTLs. Em outras palavras, a variância explicada pelo j -ésimo loco, σ_j^2 , origina-se de uma distribuição *a priori* denotada por $p(\sigma_j^2)$. Isto permite que locos distintos expliquem proporções distintas da variação genética, as quais são estimadas durante o ajuste do modelo e combinam a informação *a priori* com a verossimilhança dos dados. Os vários modelos Bayesianos propostos na literatura também podem ser definidos pela equação (1) acima, mas são diferenciados do modelo G-BLUP e entre si por adotar distribuições diferentes para $p(\sigma_j^2)$, conforme será detalhado a seguir (MEUWISSEN; HAYES; GODDARD, 2001).

2.6.3.2.1 Modelos BayesA

Este primeiro modelo Bayesiano atribui à variância explicada por cada SNP uma distribuição *a priori* do tipo qui-quadrado invertida, o que resulta em distribuição *a posteriori* da mesma família. O mesmo é válido para a variância do erro, σ_e^2 . Tal distribuição *a posteriori* não permite que a estimação seja realizada algebricamente, pois depende dos efeitos desconhecidos g_j . Portanto, é necessário fazer uso de técnicas de otimização numérica, com destaque para métodos baseados em cadeias MCMC (*Markov Chain Monte Carlo*) (ROBERT; CASELLA, 2004). Dadas essas variâncias, os efeitos dos marcadores são amostrados a partir de uma distribuição normal (MEUWISSEN; HAYES; GODDARD, 2001).

A combinação dessas distribuições faz com que os efeitos dos SNPs tenham uma distribuição t de Student. Consequentemente, é possível que alguns SNPs tenham efeito maior do que teriam sob suposição de normalidade, já que a distribuição t possui caudas mais volumosas que a distribuição normal (MEUWISSEN; HAYES; GODDARD, 2013). Observa-se, por conseguinte, que tal modelo é mais adequado à observação de que alguns QTLs possuem efeitos bastante pronunciados, quando comparado ao modelo G-BLUP.

2.6.3.2.2 Modelos BayesB

O método Bayes B parte do princípio que muitos locos não explicam a variação genética em absoluto, de modo que apenas alguns locos restantes são responsáveis por toda a variância. Assim, assume-se uma distribuição *a priori* com certa densidade (π) para $\sigma_j^2 = 0$ e,

consequentemente, $(1 - \pi)$ para $\sigma_j^2 > 0$ (com distribuição qui-quadrado invertida como no modelo Bayes A) (MEUWISSEN; HAYES; GODDARD, 2001). Em termos práticos, isto significa que uma fração π de SNPs não possui efeito algum, enquanto que a fração restante $(1 - \pi)$ tem efeito amostrado de uma distribuição t de Student.

2.6.3.2.3 Modelos BayesC π

Ambos os modelos Bayes A e B são fortemente influenciados pela escolha dos hiperparâmetros das distribuições *a priori* e, especificamente no caso de Bayes B, a probabilidade π é usualmente considerada conhecida (GIANOLA et al., 2009). Para contornar essas limitações, Habier et al. (2011) propuseram o modelo Bayes C π , em que uma variância única de efeitos é empregada em conjunto para todos os SNPs com efeito diferente de zero, ao invés de variâncias individuais para cada loco. Isto faz com que o modelo seja menos influenciado pela escolha da distribuição *a priori* $p(\sigma_j^2)$. Adicionalmente, o modelo C π trata a proporção π como desconhecida e a estima a partir dos dados. É importante ressaltar que o pacote BGLR, utilizado para ajuste de modelos de seleção genômica no ambiente estatístico R, atualmente também permite que esta probabilidade seja considerada como desconhecida no caso do modelo Bayes B (PÉREZ; DE LOS CAMPOS, 2014b).

Como resultado dessas modificações, a suposição é de que os efeitos dos SNPs são amostrados a partir de uma distribuição normal, e não t de Student, o que desfavorece a ocorrência de efeitos de elevada magnitude (MEUWISSEN et al., 2013). Resultados obtidos a partir de simulações e dados reais mostram que Bayes B e C π fornecem acurácias similares, porém o ajuste do modelo Bayes C π é computacionalmente mais simples (HABIER et al., 2011).

2.6.3.2.4 Modelos BayesRR

O modelo BRR (*Bayesian Ridge Regression*) é a versão bayesiana do modelo RR-BLUP/G-BLUP. Os efeitos das marcas seguem uma distribuição normal, assim como acontece no método RR-BLUP/G-BLUP, mas se assume que a variância dos marcadores e a variância do erro seguem uma distribuição *a priori* do tipo qui-quadrado invertida. Isto influencia o valor do parâmetro de penalização λ , o qual é tratado como uma variável aleatória a ser estimada.

2.6.3.2.5 Modelos Lasso Bayesiano

Park e Casella (2008) apresentaram uma versão bayesiana do método LASSO (TIBSHIRANI, 1996), denominada Lasso Bayesiano (*Bayesian Lasso* – BL). A proposta inicial do BL tratou π como desconhecido, o que foi posteriormente incorporado à implementação do modelo Bayes B no BGLR. No método BL uma distribuição controlada por λ é declarada para a variância de todos os SNPs.

A distribuição assumida para os efeitos genéticos dos marcadores é uma exponencial dupla, a qual apresenta maior massa de densidade no valor zero e caudas mais robustas, exercendo maior encolhimento sobre coeficientes de regressão próximos de zero e menor encolhimento sobre coeficientes de regressão distantes de zero. Embora a distribuição t coloque uma maior massa de densidade no valor zero quando comparada à *priori* Gaussiana do método BRR, esta massa de densidade no valor zero é ainda maior na distribuição exponencial dupla (Figura 1). Por isso, nos casos em que a maioria das marcas tem efeito muito pequeno, uma *priori* com distribuição exponencial dupla é mais adequada do que uma com distribuição Gaussiana ou t de Student (DE LOS CAMPOS et al., 2009; LI; SILLANPÄÄ, 2012). Entretanto, Park e Casella (2008) mostraram empiricamente que o BL pode não ser o método mais adequado para dados contendo apenas um pequeno número de QTLs com grandes efeitos.

Figura 1 - Densidades *a priori* dos coeficientes de regressão empregadas em diversos modelos Bayesianos

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Material Genético

O material genético utilizado neste trabalho é proveniente da Embrapa Milho e Sorgo, sendo composto por um painel de sorgo biomassa contendo 200 linhagens, das quais 150 constituem sorgo biomassa e 50 sorgo sacarino. Esse painel foi elaborado por pesquisadores da Embrapa e engloba genótipos de origens diversas, como, por exemplo, materiais do banco de germoplasma da Embrapa, coleções núcleo do CIRAD (França) e ICRISAT (Índia), além de materiais pertencentes aos programas de melhoramento de sorgo da Embrapa. Esses materiais foram selecionados com base em características agronômicas importantes para a produção e qualidade da biomassa e em dados disponíveis em publicações.

Os 200 materiais do painel de sorgo biomassa podem ser subdivididos em dois sub-painéis. Em detalhes, o sub-painel 1 é composto por materiais diversos, originados do banco de germoplasma da Embrapa e coleções núcleo do CIRAD (França) e ICRISAT (Índia), sendo composto por 50 linhagens de sorgo biomassa e 50 linhagens sacarinas. O sub-painel 2 contém os 100 genótipos restantes, compostos por linhagens de sorgo biomassa, oriundas do banco de germoplasma e do programa de melhoramento da Embrapa Milho e Sorgo, os quais foram adicionados posteriormente ao painel de sorgo biomassa e caracterizam-se por apresentar menor variabilidade genética, maior sensibilidade ao fotoperíodo e altos níveis de celulose, quando comparados aos genótipos do sub-painel 1.

3.1.2 Dados Moleculares

Os 200 indivíduos do painel de sorgo biomassa foram genotipados por sequenciamento no IGD (*Institute for Genomic Diversity*) da Universidade de Cornell (EUA), onde a técnica de GBS foi desenvolvida, gerando marcadores do tipo SNP (ELSHIRE; GLAUBITZ; SUN, 2011). Para tanto, a obtenção de SNPs foi realizada alinhando-se os fragmentos sequenciados à última versão do genoma de referência do sorgo (PATERSON et al., 2009), aplicando-se assim o chamado *Discovery Pipeline*.

Uma das características da técnica GBS é o grande número de dados perdidos. Portanto, visando contornar esta limitação, os dados gerados via GBS foram imputados

utilizando o software NPUTE (ROBERTS et al., 2007). Para cada cromossomo foram testados diversos tamanhos de janelas de imputação, sendo que as janelas com maiores acurácias foram escolhidas. Posteriormente, os marcadores com frequência alélica do alelo alternativo (*Minor Allele Frequency*, ou MAF) abaixo de 5% foram eliminados.

Os dados genotípicos foram então codificados, para cada SNP, atribuindo-se o valor “-1” aos indivíduos homozigotos para o alelo de referência, “0” aos indivíduos heterozigotos e “1” aos indivíduos homozigotos para o alelo alternativo. Tal codificação foi realizada através de códigos próprios, no ambiente estatístico R (R Core Team, 2013).

3.1.2.1 Análise Descritiva dos Dados Genotípicos

A tecnologia de GBS é ainda bastante recente e, consequentemente, pouco se sabe a respeito da natureza dos dados genotípicos obtidos, em especial no tocante a particularidades de cada espécie biológica. Com o objetivo de obter informações descritivas dos dados genotípicos, bem como permitir a filtragem de SNPs e/ou indivíduos potencialmente problemáticos, algumas estatísticas relevantes foram extraídas deste conjunto de dados genotípicos, dentre elas a frequência de genótipos heterozigotos para cada SNP, a frequência de dados perdidos por genótipos e por SNP e a frequência alélica do alelo alternativo. É importante ressaltar que tais valores foram calculados a partir dos dados não imputados, visto que tal prática elimina os dados perdidos e re-codifica indivíduos heterozigotos como homozigotos para um ou outro alelo, consequentemente alterando as frequências alélicas. Todos esses cálculos foram realizados no ambiente R.

Realizou-se também uma análise de componentes principais, de modo a averiguar o agrupamento dos indivíduos do painel, bem como verificar a possível existência de estrutura populacional entre os genótipos. Para tanto, utilizou-se o pacote *pcaMethods* (STACKLIES et al., 2007).

3.1.3 Dados Fenotípicos

Os dados fenotípicos foram avaliados em ensaios conduzidos na área experimental da Embrapa Milho e Sorgo, em Sete Lagoas – MG. Os 100 indivíduos do sub-painel 1 foram avaliados nos anos agrícolas de 2011 e 2012 a partir de três ensaios, utilizando-se o delineamento experimental de látice (10 x 10). Os 100 indivíduos do sub-painel 2 foram

avaliados no ano agrícola de 2011 também a partir de três ensaios, utilizando-se novamente o delineamento experimental de látice (10 x 10).

As seguintes características morfo-agronômicas foram avaliadas: i) florescimento: número de dias da semeadura até o início da liberação de pólen em 50% das plantas da parcela; ii) altura de plantas (AP): altura média, em metros, das plantas de cada parcela, medidas da superfície do solo ao ápice da panícula; e iii) produção de massa verde total (PMV) e massa seca total (PMS): determinadas em kg/parcela, através da pesagem de todas as plantas (completas) de cada parcela, colhidas na maturidade fisiológica do grão. Para a determinação da matéria seca foi retirada uma amostra de biomassa verde das parcelas, as quais foram armazenadas em estufa a 65°C por 72 horas. Posteriormente, através da diferença entre os pesos secos e úmidos, obteve-se a porcentagem de matéria seca de cada parcela. Os dados de PMV e PMS foram convertidos para Mg.ha⁻¹.

Adicionalmente, a biomassa foi caracterizada em laboratório, através da determinação das proporções de fibra em detergente ácido (FDA) e neutro (FDN), celulose, hemicelulose e lignina. Para a obtenção da proporção de FDA, FDN e lignina adotaram-se as respectivas metodologias do manual ANKOM Fiber Analyzer: *Technology Method 5* (2006), *Technology Method 6* (2006) e *Method for determing acid lignin in beakers* (2010). As proporções de hemicelulose e celulose foram estimadas pela diferença entre os valores de FDN e FDA, e pelos valores de FDN e lignina, respectivamente.

3.2 Métodos

3.2.1 Análise Fenotípica

Para a análise dos dados fenotípicos, foram ajustados modelos mistos apropriados para os delineamentos experimentais empregados para avaliação dos genótipos dos sub-painéis 1 e 2. Para os genótipos do sub-painel 1 o modelo misto empregado para cada caráter foi:

$$y_{ijkm} = \mu + \beta S_{ikm} + a_m + r_{km} + b_{jkm} + g_i + \varepsilon_{ijkm} \quad (2)$$

em que y_{ijkm} é o valor fenotípico do genótipo i , no bloco j , na repetição k e no ano m ; μ é o efeito fixo referente ao intercepto; S_{ikm} é o estande do genótipo i , na repetição k , no ano m , e β é o efeito fixo do estande; a_m é o efeito fixo do ano m ; r_{km} é o efeito aleatório da repetição k no ano m ; b_{jkm} é o efeito aleatório do bloco j , na repetição k , no ano m ; g_i é o efeito aleatório do genótipo i ; e ε_{ijkm} é o efeito residual aleatório.

Para os indivíduos do sub-painel 2, o modelo misto empregado para cada caráter foi:

$$y_{ijk} = \mu + \beta S_{ik} + r_k + b_{jk} + g_i + \varepsilon_{ijk} \quad (3)$$

em que y_{ijk} é o efeito aleatório fenotípico do genótipo i , no bloco j , na repetição k ; S_{ik} é o estande do genótipo i , na repetição k , e β é o coeficiente angular fixo correspondente; r_k é o efeito aleatório da repetição k ; b_{jk} é o efeito aleatório do bloco j , na repetição k ; e ε_{ijk} é o efeito residual aleatório.

Assumiu-se que o efeito aleatório de genótipo, g_i , segue uma distribuição normal com média zero e variância σ_g^2 . Os efeitos de repetição r_{km} ou r_k , bloco b_{jkm} ou b_{jk} e resíduo ε_{ijkm} ou ε_{ijk} também seguem uma distribuição normal centrada em zero. As matrizes de variância-covariância (VCOV) para estes efeitos foram estruturadas para os fatores de agrupamentos convenientes: ano e/ou repetição e/ou bloco (Tabela1). Brevemente, a matriz identidade modela uma variância em comum para todos os níveis do fator; o modelo diagonal permite variâncias distintas, sem haver correlação entre as variáveis; o modelo de simetria composta heterogênea assume a existência de variâncias distintas, com correlação em comum entre as diversas variáveis; por fim, a matriz não estruturada permite que todos os elementos da matriz sejam diferentes, isto é, variâncias e covariâncias distintas para todas as variáveis. Os diferentes modelos para estrutura de VCOV foram comparados primeiramente para os efeitos genéticos, representados pela matriz **G**, seguindo o critério BIC (*Bayesian Information Criterion*) (SCHWARZ, 1978). Uma vez determinada a melhor estrutura de (co)variâncias genéticas, estruturas semelhantes foram comparadas para os efeitos residuais, denotados pela matriz **R**. Uma vez modeladas as matrizes de variância-covariância e ajustados os modelos, foi possível obter os BLUPs dos genótipos para cada caráter analisado.

Tabela 1 – Modelos de estrutura de variância-covariância (VCOV) comparados para os efeitos genéticos

Modelo	Matriz de variância-covariância
Identidade (ID)	$\begin{bmatrix} \sigma_g^2 + \sigma_{ge}^2 & 0 & \cdots & 0 \\ 0 & \sigma_g^2 + \sigma_{ge}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_g^2 + \sigma_{ge}^2 \end{bmatrix}$
Diagonal (DIAG)	$\begin{bmatrix} \sigma_{g1}^2 + \sigma_{ge1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{g2}^2 + \sigma_{ge2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{gj}^2 + \sigma_{gej}^2 \end{bmatrix}$
Simetria Composta Heterogênea (CS _{Het})	$\begin{bmatrix} \sigma_{g1}^2 + \sigma_{ge1}^2 & \sigma_g^2 & \cdots & \sigma_g^2 \\ \sigma_g^2 & \sigma_{g2}^2 + \sigma_{ge2}^2 & \cdots & \sigma_g^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_g^2 & \sigma_g^2 & \cdots & \sigma_{gj}^2 + \sigma_{gej}^2 \end{bmatrix}$
Não Estruturada (UNS)	$\begin{bmatrix} \sigma_{g1}^2 + \sigma_{ge1}^2 & \sigma_{g12}^2 & \cdots & \sigma_{g1j}^2 \\ \sigma_{g21}^2 & \sigma_{g2}^2 + \sigma_{ge2}^2 & \cdots & \sigma_{g2j}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{gj1}^2 & \sigma_{gj2}^2 & \cdots & \sigma_{gj}^2 + \sigma_{gej}^2 \end{bmatrix}$

σ_g^2 e σ_{ge}^2 : componentes de variância para o efeito principal de genótipos e para a interação genótipo x ambiente, respectivamente.

Todas essas análises foram realizadas no software GenStat versão 16 (PAYNE; WELHAM; HARDING, 2013).

3.2.2 Ajuste dos Modelos de Seleção Genômica

Os modelos de seleção genômica foram ajustados no software R, utilizando o pacote BGLR (PÉREZ; DE LOS CAMPOS, 2014a) para ajuste dos modelos bayesianos: Bayes A, Bayes B, Bayes C π , Bayes Lasso e Bayes RR; e o pacote rrBlup (ENDELMAN, 2011) para o ajuste do modelo RR-BLUP. Foram utilizadas 40.000 iterações para os modelos bayesianos, sendo 20.000 descartadas como *burn-in*, e assumidos parâmetros padrão para as *prioris* do modelos. Visando explorar ao máximo as informações contidas nos dados e atender às necessidades do programa de melhoramento de sorgo biomassa da Embrapa, diferentes estratégias de ajuste de modelos foram adotadas, as quais serão detalhadas a seguir.

3.2.2.1 Análise Conjunta dos Dois Sub-Painéis

Inicialmente, desconsiderando a divisão do painel de genótipos em dois sub-conjuntos, os 200 genótipos foram analisados conjuntamente, de modo a aproveitar as informações de todos os indivíduos genotipados para a predição dos efeitos dos marcadores, além de permitir que a população de treino fosse a maior possível.

Desse modo, a técnica da validação cruzada foi empregada no painel de sorgo biomassa com dez subconjuntos, sendo que os 200 genótipos foram particionados em dez grupos mutuamente exclusivos, contendo 20 genótipos cada. Para cada uma das dez instâncias de validação, o conjunto de treino foi utilizado para ajustar os modelos de seleção genômica, a fim de se estimar os efeitos dos marcadores com base em informações genotípicas e fenotípicas. No respectivo conjunto de teste, tais efeitos foram empregados para estimação dos valores genéticos com base apenas nas informações genotípicas (GEBVs). Ao final, as capacidades preditivas dos modelos estatísticos ajustados foram verificadas através da correlação entre o valor genético predito com base em dados genotípicos e o valor genético predito com base em dados fenotípicos (EBVs).

3.2.2.2 Seleção Genômica entre Sub-painéis

Devido ao fato de as avaliações fenotípicas para rendimento e qualidade de biomassa serem laboriosas e de custo elevado, é interessante investigar a aplicabilidade de modelos de seleção genômica ajustados em painéis de genótipos já selecionados, para prever o comportamento de genótipos que se encontram em fases iniciais de melhoramento, com o objetivo de poupar tempo e recursos. Visto que os programas brasileiros de melhoramento de sorgo biomassa são ainda recentes, a possibilidade de utilizar genótipos oriundos de outros programas de melhoramento, os quais já foram selecionados, para prever o comportamento de genótipos relevantes aos sistemas de produção brasileiros torna-se uma alternativa interessante. Dessa forma, os dados fenotípicos e genotípicos dos 100 indivíduos do sub-painel 1 foram utilizados para a obtenção dos efeitos dos marcadores SNPs, os quais, em seguida, foram empregados para a estimação dos valores genéticos com base nos GEBVs dos 100 indivíduos do sub-painel 2. Posteriormente, as capacidades preditivas foram obtidas através da correlação entre os GEBVs dos indivíduos do sub-painel 2 e seus respectivos EBVs.

3.2.2.3 Simulações de Redução do Número de Marcadores

Visando avaliar o reflexo do número de marcadores na eficácia da seleção genômica, foram simulados vários cenários variando-se o número de marcadores utilizados. Posteriormente, para cada cenário simulado, o modelo RR-BLUP foi ajustado, as capacidades preditivas foram obtidas e comparadas. Optou-se por ajustar apenas o modelo RR-BLUP devido à sua menor complexidade computacional e maior rapidez de execução, quando comparado aos outros modelos.

A seleção do número de marcadores a serem utilizados foi realizada seguindo duas abordagens diferentes. Em um primeiro cenário, esta redução foi aleatória, objetivando-se não privilegiar os marcadores de maiores efeitos, mas apenas avaliar o efeito causado pelo número de marcadores utilizados para o ajuste do modelo. Mais detalhadamente, a partir do conjunto de todas as marcas, realizou-se o descarte aleatório de metade destas e o modelo foi novamente ajustado. De modo a avaliar a variância das capacidades preditivas, esse procedimento foi realizado por dez vezes, sempre tomando-se como ponto de partida o conjunto de todas as marcas disponíveis. Em seguida, o número de marcas foi reduzido a um quarto do valor inicial e o modelo ajustado. Novamente, dez amostras independentes foram obtidas e avaliadas. Estes procedimentos foram aplicados iterativamente, com redução gradual do número de marcas seguindo potências de dois, até se atingir o menor número de marcas avaliado, constituído por 16.

Em um segundo cenário, a redução do número de marcadores foi realizada de forma a sempre se selecionar os marcadores de maiores efeitos, objetivando-se, assim, avaliar o efeito de se retirar dos modelos ajustados as marcas de efeitos pequenos, as quais potencialmente apenas acrescentam ruído ao modelo. Portanto, inicialmente o modelo com todas as marcas foi ajustado; em seguida, a metade das marcas com efeito de maior magnitude foi selecionada e utilizada para ajustar novamente o modelo. Sucessivamente, modelos foram ajustados com metade das marcas do modelo anterior, até atingir o mínimo de 16 marcas. Nesse cenário, não foi possível realizar dez simulações com cada número de marcas avaliado, pois as marcas não foram escolhidas aleatoriamente, mas selecionadas de acordo com a magnitude de seus efeitos, isto é, o módulo do efeito predito.

3.2.3 Relação entre Capacidade Preditiva e Herdabilidade

Os coeficientes de herdabilidade no sentido amplo foram estimados para todos os caracteres avaliados, em cada sub-painel separadamente e em conjunto, através da equação:

$$H^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \frac{\sigma_e^2}{n})}; \text{ em que } n \text{ é o número de repetições, } \sigma_g^2 \text{ é o componente de variância genética}$$

e σ_e^2 é o componente de variância residual, ambos estimados pelo software GenStat versão 16 (PAYNE; WELHAM; HARDING, 2013). De posse dessas herdabilidades, foi possível comparar as capacidades preditivas obtidas por cada modelo ajustado com a herdabilidade da característica correspondente, objetivando-se assim avaliar a influência da herdabilidade da característica na capacidade preditiva dos modelos ajustados.

3.2.4 Comparação dos Modelos

Após o ajuste dos modelos de seleção genômica, as capacidades preditivas obtidas foram comparadas entre modelos, para cada caráter fenotípico avaliado. Posteriormente, comparou-se a co-ocorrência dos genótipos superiores selecionados por cada modelo de seleção genômica com a seleção fenotípica via BLUP. Com o objetivo de comparar a possibilidade de aplicação da seleção genômica na detecção de genótipos superiores, com elevada intensidade de seleção, à sua aplicação na eliminação de genótipos inferiores, com menor intensidade de seleção, os modelos ajustados para cada caráter foram utilizados para realizar a seleção dos 10%, 30% e 50% genótipos superiores. Para cada um dos níveis de intensidade de seleção utilizados, contabilizaram-se quantos genótipos foram simultaneamente selecionados pelo valor de BLUP e por cada modelo de seleção genômica.

3.2.5 Enriquecimento Funcional

Os efeitos preditos dos marcadores na análise conjunta dos dois sub-painéis foram utilizados para realizar uma análise de enriquecimento funcional. O objetivo era verificar a existência de efeitos de maior magnitude para um ou outro determinado grupo funcional. Para tanto, os marcadores foram agrupados com base no banco de dados do *Gene Ontology* (GO),

o qual fornece uma ontologia de termos definidos representando as propriedades de um produto gênico.

Através da plataforma Phytozome (GOODSTEIN et al., 2012), obteve-se a anotação do genoma do sorgo, a qual inclui modelos gênicos preditos e anotações de termos GO para cada produto gênico predito (PATERSON et al., 2009). A cada SNP situado dentro de um produto gênico qualquer foram transferidos todos os termos GO correspondentes, de modo que alguns SNPs foram associados a termos em particular. Para cada termo GO encontrado neste conjunto de dados dos SNPs, os marcadores relevantes, isto é, aqueles que estavam anotados com o termo em questão, foram inicialmente discriminados. Em seguida, aplicou-se o teste de Kolmogorov-Smirnov (FRANK; MASSEY, 1951) para comparar a distribuição dos efeitos destes marcadores com a distribuição dos efeitos dos marcadores remanescentes. Como o intuito era detectar termos GO em que os efeitos dos marcadores eram superiores, em média, aos efeitos do conjunto de todos os marcadores, aplicou-se um teste de significância unilateral, no qual foram comparados os valores absolutos dos efeitos, ou seja, o módulo do efeito predito.

Posteriormente, realizou-se o controle do erro tipo I para múltiplos testes através da correção de FDR (*False Discovery Rate*) (BENJAMINI; HOCHBERG, 1995), de modo que apenas os termos GO com efeitos significativos (p -valor ajustado $< 0,01$) foram selecionados. Por fim, foram selecionados os termos GO relacionados a marcadores com efeitos de magnitude significativamente maior, e a existência de enriquecimento funcional foi visualizada através do método REVIGO, disponível *online* (SUPEK et al., 2011).

4 RESULTADOS

4.1 Análise Fenotípica

Para cada caráter fenotípico, foram testados e comparados 36 modelos de estrutura de VCOV para os efeitos genéticos representados pela matriz **G** e 64 modelos para os efeitos residuais, denotados pela matriz **R**, para os indivíduos do sub-painel 1. Para os indivíduos do sub-painel 2 foram testados e comparados 4 modelos de estrutura de VCOV para a matriz **G** e 16 modelos para a matriz **R**. As estruturas de VCOV de melhor ajuste, com os respectivos valores do critério BIC (*Bayesian Information Criterion*) (SCHWARZ, 1978) estão apresentadas nas Tabelas 2 e 3.

Tabela 2 – Modelos de estruturas de variância-covariância (VCOV) selecionados para as matrizes **G** e **R**, com respectivos valores do critério BIC, para os indivíduos do sub-painel 1

Fenótipo	Matriz G		Matriz R	BIC
	Rep/Ano	Bloco/Rep/Ano	Parcela/Bloco/Rep/Ano	
Altura	ID	ID⊗DIAG	ID⊗DIAG⊗ID	647,11
Celulose	ID	ID⊗ID	ID⊗DIAG⊗ID	3498,04
FDA	ID	ID⊗ID	ID⊗DIAG⊗ID	3670,39
FDN	ID	DIAG⊗DIAG	DIAG⊗ID⊗ID	3952,86
Florescimento	ID	ID⊗ID	DIAG⊗DIAG⊗ID	3970,13
Hemicelulose	ID	DIAG⊗DIAG	DIAG⊗DIAG⊗UNS	3220,98
Lignina	ID	ID⊗ID	ID⊗DIAG⊗CS _{HET}	2047,34
PMS	ID	DIAG⊗ID	DIAG⊗CS _{HET} ⊗DIAG	3469,74
PMV	ID	ID⊗ID	DIAG⊗ID⊗ID	4663,75

Tabela 3 – Modelos de estruturas de variância-covariância (VCOV) selecionados para as matrizes **G** e **R**, com respectivos valores do critério BIC, para os indivíduos do sub-painel 2

Fenótipo	Matriz G	Matriz R	BIC
	Bloco/Rep	Parcela/Bloco/Rep	
Altura	ID	ID $\otimes$ ID	367,30
Celulose	ID	ID $\otimes$ ID	1536,43
FDA	ID	ID $\otimes$ ID	1723,11
FDN	ID	ID $\otimes$ ID	1754,76
Florescimento	ID	DIAG $\otimes$ ID	2116,85
Hemicelulose	ID	ID $\otimes$ ID	1351,72
Lignina	ID	ID $\otimes$ ID	1134,64
PMS	ID	DIAG $\otimes$ DIAG	1442,22
PMV	ID	DIAG $\otimes$ ID	1994,65

4.2 Análise Descritiva dos Dados Genotípicos

A genotipagem por sequenciamento dos 200 indivíduos do painel de sorgo biomassa, realizada pelo IGD (Institute for Genomic Diversity) da Universidade de Cornell (EUA), gerou um total de 1.024.892 SNPs, os quais foram distribuídos pelos 10 cromossomos do sorgo (Figura 2). Observa-se que a maior parte dos marcadores concentrou-se principalmente nos cromossomos 1, 2 e 3, com 159.941, 130.833 e 137.091 SNPs, respectivamente. Em contrapartida, o cromossomo 8 apresentou o menor número de marcadores, 67.622.

Figura 2 – Distribuição dos 1.024.892 marcadores pelos 10 cromossomos do sorgo

4.2.1 Frequência de Genótipos Heterozigotos

Por meio da análise da frequência de genótipos heterozigotos dos dados não imputados, observa-se que a grande maioria de SNPs apresentou proporção muito baixa de genótipos heterozigotos, o que de fato já era esperado por se tratar de um painel de linhagens (Figura 3). A proporção de genótipos heterozigotos por SNP variou de 0 a 0,95, com média igual a 0,042 e mediana 0,022. Apesar de estarem presentes em pequenas quantidades, alguns genótipos heterozigotos foram observados e codificados devidamente.

Figura 3 – Frequência de indivíduos heterozigotos por marcador, para os 200 genótipos do painel de sorgo biomassa

4.2.2 Frequência de Dados Perdidos

Uma das primeiras etapas a serem realizadas em estudos de seleção genômica é a averiguação dos SNPs quanto à ocorrência de dados perdidos, os quais poderão posteriormente ser desconsiderados ou imputados. O cálculo da proporção de dados perdidos, para cada SNP, indicou que a porcentagem de dados perdidos variou de 0 a 0.995, com média de 0,346, mediana 0,285 e moda igual a 0,046 (Figura 4). Analisando a proporção de dados perdidos por indivíduo, pode-se observar que a maior parte dos 200 indivíduos apresentaram entre 20 e 50% de dados perdidos (Figura 5). Esta última medida variou de 0,196 a 0,662, com média 0,346 e mediana 0,341. Como o sorgo é uma espécie diploide e autógama, além de haver disponível um genoma de referência bem montado, optou-se por realizar a imputação dos dados perdidos e utilizar os dados imputados para todas as análises subsequentes.

Figura 4 – Distribuição de dados perdidos por marcador, para os 200 genótipos do painel de sorgo biomassa

Figura 5 – Distribuição de dados perdidos por indivíduo, de um total de 1.024.892 marcadores

4.2.3 Frequência do Alelo Alternativo (*Minor Allele Frequency* - MAF)

O resultado da distribuição da frequência alélica do alelo alternativo para os 1.024.892 SNPs, com base nos dados imputados, é ilustrado pela Figura 6. Por meio dela pode-se observar que a maior parte dos SNPs apresentaram MAF inferior a 5%. Em geral, marcadores com MAF muito baixa podem representar falsos positivos resultantes de erros de sequenciamento os quais devem ser eliminados (GLAUBTIZ et al., 2014). Mesmo quando efetivamente representam variação genética existente, esses marcadores são pouco informativos. Sendo assim, optou-se por utilizar um valor mínimo de MAF de 5% e, após filtragem com este critério, restaram 258.220 SNPs, os quais foram utilizados para os ajustes dos modelos de seleção genômica.

Figura 6 – Distribuição da frequência do alelo alternativo (MAF – *minor allele frequency*) para 1.024.892 marcadores encontrados no painel de genótipos de sorgo biomassa

4.2.4 Análise de Componentes Principais

A análise de componentes principais é um método da estatística multivariada que objetiva a obtenção de um pequeno número de combinações lineares (componentes principais) de um conjunto de variáveis, que retenham o máximo possível da variação contida nas variáveis originais. Através dela pode-se descrever o conjunto de dados constituído por 200 indivíduos caracterizados por 258.220 SNPs, de modo a obter informações de agrupamento entre genótipos. A Figura 7 apresenta o gráfico do tipo *biplot* da análise de componentes principais, no qual se pode observar que o primeiro componente explicou 12,42% da variação dos dados, enquanto o componente dois explicou 4,09% dessa variação. Nota-se a formação de dois grupos de indivíduos, os quais refletem a subdivisão dos genótipos em dois sub-painéis, fato que por sua vez reflete a composição desses sub-painéis quanto à variabilidade dos genótipos. Em detalhes, os pontos pretos representam indivíduos pertencentes ao sub-painel 1, enquanto que pontos em cinza representam indivíduos

pertencentes ao sub-painel 2. Observa-se que os genótipos do sub-painel 1 situaram-se predominantemente à direita do gráfico *biplot*, ao passo que os genótipos do sub-painel 2 ficaram à esquerda. Assim, nota-se que o primeiro componente principal separa os indivíduos pertencentes aos dois sub-painéis.

Figura 7 - Análise de componentes principais dos 200 indivíduos do painel de sorgo biomassa caracterizados por 258.220 marcadores

4.3 Ajuste dos Modelos de Seleção Genômica

4.3.1 Análise Conjunta dos Dois Sub-Painéis

Os resultados das capacidades preditivas da análise conjunta dos dois sub-painéis estão na Tabela 4. Observa-se que os caracteres FDA, FDN, celulose e lignina apresentaram as maiores capacidades preditivas, enquanto os caracteres hemicelulose e florescimento apresentaram as menores.

Para cada um dos nove fenótipos avaliados, todos os modelos de predição ajustados obtiveram capacidades preditivas similares. O modelo RR-BLUP foi o que apresentou maior capacidade preditiva para todos os caracteres avaliados; em contrapartida, o modelo BayesLasso apresentou os menores valores. Em geral, as diferenças entre os modelos foram de poucos pontos percentuais. Por exemplo, para o caráter altura, os valores variaram apenas

de 0,79 a 0,77, na comparação entre o melhor e pior modelo preditivo ajustado, respectivamente. Comportamento semelhante foi observado para os demais caracteres.

Tabela 4 – Capacidades preditivas obtidas por diferentes modelos de seleção genômica, aplicados à seleção conjunta dos genótipos de sorgo dos dois sub-painéis, para nove fenótipos avaliados

Fenótipo	Modelo Preditivo					
	BayesB	BayesA	BayesRR	BayesC	BayesLasso	RRBLUP
Altura	0,78	0,78	0,78	0,78	0,77	0,79
Celulose	0,82	0,82	0,82	0,81	0,81	0,83
FDA	0,84	0,84	0,84	0,84	0,83	0,85
FDN	0,85	0,84	0,84	0,84	0,83	0,85
Florescimento	0,59	0,59	0,59	0,59	0,57	0,61
Hemicelulose	0,68	0,68	0,68	0,68	0,67	0,69
Lignina	0,81	0,81	0,81	0,81	0,80	0,82
PMS	0,72	0,72	0,72	0,72	0,71	0,73
PMV	0,77	0,78	0,77	0,77	0,77	0,79

4.3.1.1 Relação entre Capacidade Preditiva e Herdabilidade

A herdabilidade de um caráter é um dos fatores responsáveis por afetar a acurácia dos modelos de seleção genômica. Em geral, quanto maior o coeficiente de herdabilidade do caráter, maior espera-se que seja a capacidade preditiva do modelo. A Tabela 5 apresenta os coeficientes de herdabilidade no sentido amplo, obtidos através da análise fenotípica conjunta dos indivíduos dos dois sub-painéis. Os caracteres PMV e altura apresentaram as maiores herdabilidades, 0,95 e 0,94, respectivamente, enquanto os caracteres hemicelulose e celulose apresentaram as menores herdabilidades, 0,30 e 0,78, respectivamente. É notável que as menores herdabilidades foram obtidas para caracteres mensurados pela diferença entre os valores de FDN e FDA, e pelos valores de FDN e lignina, respectivamente.

A Figura 8 apresenta a relação entre os coeficientes de herdabilidade avaliados e as capacidades preditivas, para os diversos caracteres. Em linhas gerais, não se observa tendência de aumento da capacidade preditiva à medida que aumenta a herdabilidade do caráter. Os caracteres que apresentaram as maiores capacidades preditivas exibiram herdabilidade na ordem de 0,80, à exceção do caráter florescimento, o qual apresentou as

capacidades preditivas mais baixas, mesmo com herdabilidade de 0,82. Apesar de o caráter hemicelulose apresentar a menor herdabilidade, igual a 0,30, este não apresentou a menor capacidade preditiva. De maneira semelhante, apesar de os caracteres PMV e altura apresentarem as maiores herdabilidades, estes não apresentaram as maiores capacidades preditivas.

Tabela 5 – Coeficientes de herdabilidades dos fenótipos avaliados, obtidos pela análise conjunta dos dois sub-painéis

Fenótipo	Herdabilidade
Altura	0,94
Celulose	0,78
FDA	0,82
FDN	0,79
Florescimento	0,82
Hemicelulose	0,30
Lignina	0,77
PMS	0,89
PMV	0,95

Figura 8 – Relação entre coeficientes de herdabilidade e capacidades preditivas, na análise conjunta dos dois sub-painéis

4.3.1.2 Simulações de Redução do Número de Marcadores

A redução do número de marcadores empregados para ajuste dos modelos de seleção genômica, através da seleção dos marcadores de maior efeito, possibilitou a obtenção de capacidades preditivas superiores àquelas obtidas quando todos os SNPs foram utilizados, para todos os caracteres fenotípicos (Figura 9). De maneira geral, para todos os caracteres avaliados, o número ótimo de marcadores situou-se no intervalo de 2.018 a 64.556 marcadores. Dentro deste intervalo, as capacidades preditivas mantiveram-se praticamente constantes, variando, por exemplo, de 0,96 a 0,97 para o caráter altura. Para o número ótimo de marcas, as capacidades preditivas atingiram valores bastante altos, por exemplo, de 0,98 para os caracteres PMS e FDN, 0,97 para os caracteres altura, PMV, FDA e lignina, e 0,95

para os caracteres florescimento, celulose e hemicelulose. Números excessivamente reduzidos de marcadores, inferiores a 2.018, resultaram em capacidades preditivas menores, variando de 0,57 para o caráter hemicelulose a 0,78 para o caráter FDN, ao se utilizar o mínimo de 16 marcas.

De modo contrário, a redução do número de marcadores realizada aleatoriamente evidenciou que as capacidades preditivas foram máximas e a variância mínima utilizando-se todos os marcadores para o ajuste do modelo de seleção genômica. Assim, quanto menor o número de marcadores utilizados, menor a capacidade preditiva e maior a variância obtida (Figura 10). A capacidade preditiva manteve-se praticamente inalterada até o modelo com 4.035 marcas, em geral, depois caiu lentamente e, por fim, de maneira mais acentuada para números pequenos de marcadores.

Figura 9 – Simulação da redução do número de marcas para ajuste dos modelos de seleção genômica, selecionando-se as marcas de maiores efeitos, na análise conjunta dos dois sub-painéis

Figura 10 – Boxplot da simulação da redução do número de marcas realizada aleatoriamente, para seleção genômica conjunta do painel. Dados de 10 repetições

4.3.2 Seleção Genômica Entre Sub-Painéis

As capacidades preditivas da seleção genômica aplicada entre sub-painéis foram bastante variáveis entre os nove caracteres avaliados (Tabela 6). O fenótipo altura apresentou as maiores capacidades preditivas, as quais não apresentaram grandes variações entre modelos, variando de 0,66, no caso do modelo BayesA, a 0,65 para os demais. Os demais fenótipos apresentaram capacidades preditivas menores, acompanhadas de maiores variações entre modelos. Em um caso extremo, para o fenótipo hemicelulose, foram obtidas capacidades preditivas com valores negativos, -0,002 e -0,001, no caso dos modelos BayesC e RRBLUP, respectivamente.

O modelo estatístico resultante na maior capacidade preditiva foi diferente para cada um dos fenótipos avaliados. O modelo BayesRR forneceu a melhor capacidade preditiva para os caracteres celulose, hemicelulose e PMV; o BayesLasso para os caracteres FDA, FDN e lignina; BayesA para o caráter altura; BayesC para o caráter florescimento e BayesB para o caráter PMS.

De modo geral, as capacidades preditivas resultantes da aplicação da seleção genômica entre sub-painéis foram consideravelmente inferiores, quando comparadas àquelas obtidas aplicando-se a seleção genômica na análise conjunta do painel.

Tabela 6 – Capacidades preditivas obtidas por diferentes modelos de seleção genômica, aplicados à seleção entre sub-painéis, para nove fenótipos avaliados

Fenótipo	Modelo Preditivo					
	BayesB	BayesA	BayesRR	BayesC	BayesLasso	RRBLUP
Altura	0,65	0,66	0,65	0,65	0,65	0,65
Celulose	0,20	0,20	0,25	0,19	0,23	0,12
FDA	0,25	0,28	0,29	0,28	0,32	0,21
FDN	0,16	0,15	0,17	0,16	0,22	0,10
Florescimento	0,24	0,23	0,23	0,24	0,22	0,23
Hemicelulose	0,005	0,008	0,012	-0,002	0,008	-0,001
Lignina	0,31	0,35	0,32	0,33	0,35	0,31
PMS	0,36	0,35	0,34	0,35	0,35	0,35
PMV	0,24	0,24	0,25	0,24	0,25	0,24

4.3.2.1 Relação entre Capacidade Preditiva e Herdabilidade

Os coeficientes de herdabilidade no sentido amplo, obtidos por meio da análise fenotípica dos indivíduos do sub-painel 1, foram similares aos obtidos com a análise conjunta dos dois sub-painéis (Tabela 7). Apesar de as variâncias ambientais terem sido menores no sub-painel 1, as variâncias genéticas na análise conjunta dos dois sub-painéis foram maiores. Os caracteres altura e PMV também apresentaram as maiores herdabilidades, iguais a 0,96 e 0,87, respectivamente, enquanto os caracteres hemicelulose e celulose também apresentaram as menores herdabilidades, 0,36 e 0,73, respectivamente.

Através das relações entre os coeficientes de herdabilidade avaliados e as capacidades preditivas, observa-se uma tendência de aumento da capacidade preditiva com o aumento na herdabilidade do caráter, ou seja, existiu uma correlação positiva entre herdabilidade e capacidade preditiva (Figura 11). O fenótipo altura apresentou a maior herdabilidade e a maior capacidade preditiva; analogamente, o fenótipo hemicelulose apresentou a menor herdabilidade e a menor capacidade preditiva.

Tabela 7 – Coeficientes de herdabilidades dos fenótipos avaliados obtidas pela análise do sub-painel 1

Fenótipo	Herdabilidade
Altura	0,96
Celulose	0,73
FDA	0,78
FDN	0,76
Florescimento	0,81
Hemicelulose	0,36
Lignina	0,78
PMS	0,83
PMV	0,87

Figura 11 – Relação entre coeficientes de herdabilidades e capacidades preditivas, na análise entre sub- painéis

4.3.2.1 Simulações de Redução do Número de Marcadores

Os resultados da redução do número de marcadores utilizados para ajuste dos modelos de seleção genômica, selecionando-se sempre aqueles de maiores efeitos estão representados na Figura 12. Observa-se que, para os fenótipos altura, florescimento, lignina e PMV, a capacidade preditiva foi máxima utilizando-se todos os marcadores, mantendo-se praticamente constante até 252 SNPs para os caracteres florescimento e PMV, 1.009 SNPs para o caráter altura e 8.070 SNPs para o caráter lignina, quando, então, apresentou tendência de queda. No entanto, nota-se que esta tendência de queda foi irregular, visto que a capacidade preditiva oscilou substancialmente.

Em contrapartida, os fenótipos celulose, FDA, FDN, hemicelulose e PMS apresentaram picos para números reduzidos de marcadores, nos quais a capacidade preditiva foi máxima. Estes picos foram observados em 252 marcadores para os fenótipos celulose e FDA, 126 marcadores para o fenótipo hemicelulose e 32 marcadores para os fenótipos FDN e PMS. Diferentemente do que aconteceu para a seleção conjunta dos dois sub-painéis, nestes picos as capacidades preditivas não foram muito superiores àquelas obtidas no modelo com todas as marcas e, além disso, foram observados em regiões de grande oscilação da capacidade preditiva, visto que o número de marcadores era bastante reduzido.

A redução do número de marcadores realizada aleatoriamente entre sub-painéis também evidenciou que as capacidades preditivas médias foram máximas e a variância mínima utilizando-se todos os marcadores para ajuste do modelo (Figura 13). Entretanto, para os caracteres em que a capacidade preditiva era muito baixa com todas as marcas, como celulose, hemicelulose e FDN, não foi observada forte tendência de queda ao se utilizar números reduzidos de marcadores.

Figura 12 - Simulação da redução do número de marcas selecionando-se as marcas de maiores efeitos, na análise entre sub-painéis

Figura 13 - Boxplot da simulação da redução do número de marcas realizada aleatoriamente, para seleção genômica entre subpainéis. Dados de 10 repetições

4.4 Comparações dos Modelos

Uma maneira alternativa de comparar a seleção fenotípica via BLUP com a seleção genômica é por meio do índice de co-ocorrência entre as melhores e piores plantas selecionadas por estas duas metodologias. Além disso, por meio do índice de co-ocorrência, pode-se realizar comparações entre os diversos modelos de seleção genômica, através da avaliação da sobreposição entre os genótipos por eles selecionados.

O índice de co-ocorrência entre as 10, 30 e 50 melhores plantas selecionadas pelo BLUP e pelos diversos modelos de seleção genômica ajustados entre sub-painéis, correspondendo às intensidades de seleção de 10%, 30% e 50%, respectivamente, é apresentado na Figura 14. Os índices de co-ocorrência dos genótipos selecionados com intensidade de seleção de 10% estão representados no gráfico pela cor azul escura, conforme legenda. Observa-se que, com intensidade de seleção de 10%, para alguns dos fenótipos relacionados à caracterização da biomassa, celulose, FDN e hemicelulose, não houve co-ocorrência alguma entre as plantas selecionadas pelo BLUP e por quaisquer dos modelos de seleção genômica ajustados. Para os fenótipos FDA, florescimento e PMS, a co-ocorrência apresentada foi muito baixa. No caso do fenótipo FDA, genótipos selecionados pelo BLUP foram também selecionados apenas pelos modelos BayesC e BayesLasso, e o mesmo foi observado para o modelo BayesC no fenótipo PMS. De maneira contrastante, todos os modelos ajustados para o fenótipo florescimento foram consistentes. O fenótipo altura apresentou os maiores valores de co-ocorrência, seguido por lignina e PMV. Mesmo nestes casos, no entanto, é importante frisar que os genótipos selecionados pelos modelos de seleção genômica e pelos BLUPs apresentaram apenas moderada sobreposição, não sendo superior a 40% em situação alguma.

Em contrapartida, o índice de co-ocorrência das 10 melhores plantas selecionadas, entre os diversos modelos ajustados, foi em geral muito alto. Na realidade, tal fato foi observado para todos os caracteres fenotípicos avaliados. Por exemplo, para o caráter altura, a comparação de todos os modelos de seleção genômica, dois a dois, revelou índices de co-ocorrência nunca inferiores a 70%, isto é, os dez genótipos superiores selecionados por cada modelo incluíram sete genótipos em comum. Para algumas combinações de modelos, valores iguais a 90% foram obtidos.

Os índices de co-ocorrência entre as 30 melhores plantas selecionadas pelo BLUP e pelos diversos modelos de seleção genômica ajustados entre sub-painéis estão representados no gráfico pela cor azul. De modo análogo aos resultados encontrados utilizando-se

intensidade de seleção de 10%, os fenótipos relacionados à caracterização da biomassa, celulose, FDA, FDN e hemicelulose apresentaram os menores índices de co-ocorrência. Os valores, nesta situação, foram sempre maiores que os obtidos sob seleção mais intensa. Em outras palavras, houve maior concordância entre os modelos de seleção genômica e os BLUPs fenotípicos com intensidade de seleção de 30% do que com intensidade de 10%.

Finalmente, os índices de co-ocorrência entre as 50 melhores plantas selecionadas (intensidade de seleção de 50%), pelo BLUP e pelos diversos modelos de seleção genômica ajustados entre sub-painéis, estão representados no gráfico pela cor azul clara. Observa-se que, de maneira geral, estes foram superiores aos apresentados pelos esquemas de seleção de maior intensidade. As únicas exceções foram observadas para o fenótipo lignina, no caso dos modelos BayesA e BayesRR, em que os índices de co-ocorrência com intensidade de seleção de 50% foram ligeiramente inferiores aos obtidos com intensidade de seleção de 30%. O fenótipo altura apresentou os maiores índices de co-ocorrência, sendo estes de 78% para os modelos BayesA, BayesC, BayesLasso, BayesRR e RRBLUP, e de 74% para o modelo BayesB. De fato, mesmo os fenótipos relacionados à caracterização da biomassa, como celulose, FDA, FDN, hemicelulose e lignina, apresentaram índices de co-ocorrência substancialmente mais elevados.

O índice de co-ocorrência entre as 20, 60 e 100 melhores plantas selecionadas pelo BLUP e pelos diversos modelos de seleção genômica na análise conjunta dos dois sub-painéis, correspondendo a intensidades de seleção de 10%, 30% e 50%, respectivamente, foram maiores quando comparados aos obtidos na seleção genômica entre sub-painéis (Figuras 14 e 15). Observa-se que, entre as 20 melhores plantas selecionadas (intensidade de seleção de 10%), o fenótipo florescimento apresentou os menores índices de co-ocorrência, sendo este de 0% para o modelo BayesB (Figura 15). Para a mesma intensidade de seleção, o caráter lignina apresentou os maiores valores de co-ocorrência, sendo 60% para os modelos BayesA e RRBLUP, seguido pelo caráter FDN com 50% de co-ocorrência nos modelos BayesB e RRBLUP.

Utilizando-se intensidade de seleção de 30%, o fenótipo hemicelulose apresentou os menores índices de co-ocorrência, 23,33% nos modelos BayesC, BayesLasso, BayesRR e RRBLUP. O fenótipo lignina apresentou os maiores índices de co-ocorrência, 66,66% para todos os modelos ajustados. Os valores, nesta situação, nem sempre foram maiores que no caso anterior. Por exemplo, para o caráter hemicelulose houve maior concordância entre os modelos BayesC, BayesLasso e RRBLUP e os BLUPs fenotípicos com intensidade de seleção de 10% quando comparado à intensidade de seleção de 30%.

Por fim, em concordância com os resultados apresentados pela Figura 14, o índice de co-ocorrência entre as 100 melhores plantas selecionadas (intensidade de seleção de 50%) pelo BLUP e pelos diversos modelos de seleção genômica ajustados na análise conjunta dos dois painéis foi em geral superior ao observado com maiores intensidades de seleção. Novamente, o fenótipo lignina apresentou índices maiores com intensidade de seleção de 30% quando comparado à intensidade de seleção de 50%, no caso dos modelos BayesA, BayesB, BayesLasso e RRBLUP. Assim como na análise entre sub-painéis, o caráter altura apresentou os maiores índices de co-ocorrência, sendo estes de 74% em todos os modelos ajustados.

Figura 14 - Índices de co-ocorrência (%) entre as 10, 30 e 50 melhores plantas selecionadas pelo BLUP calculado com base apenas nos valores fenotípicos e pelos valores genéticos genômicos estimados pelos diversos modelos de seleção genômica, quando aplicados à seleção entre sub-painéis

Figura 15 - Índices de co-ocorrência (%) entre as 20, 60 e 100 melhores plantas selecionadas pelo BLUP calculado com base apenas nos valores fenotípicos e pelos valores genéticos genômicos estimados pelos diversos modelos de seleção genômica, quando aplicados na análise conjunta dos dois sub-painéis

4.5 Enriquecimento Funcional

Por meio da anotação dos SNPs, foram encontrados 1.119 termos GO relacionados aos SNPs presentes dentro dos modelos gênicos preditos. O número de SNPs associados a um

mesmo termo GO variou de 1 a 14.625, com média de 214,9 e mediana 26. Adicionalmente, o número de termos GO associados a um mesmo SNP variou de 1 a 40, com média 3,27 e mediana 3.

Os resultados da análise de enriquecimento funcional revelaram de 47 a 27 termos significativos para os diferentes fenótipos. Os caracteres hemicelulose e PMV apresentaram o maior número de termos enriquecidos, enquanto que o caráter florescimento exibiu o menor número de termos significativos. As listas de termos GO relacionados a cada caráter avaliado, com os respectivos *p*-valores ajustados, encontram-se no Anexo A.

A formação de *clusters* para os termos GO relacionados a cada caráter avaliado é representada por meio de gráficos do tipo *TreeMap*, produzidos pelo método REVIGO (SUPEK et al., 2011). Os tamanhos dos retângulos refletem a frequência de cada termo GO no banco de dados, os quais são agrupados por similaridade funcional e representados por meio de cores.

Para o caráter altura, destacam-se termos relacionados à síntese e metabolismo de biomoléculas, como aminoácidos, ácidos graxos, nucleotídeos e proteínas (*glycine biosynthesis*, *glutamine metabolic process*, *fatty acid biosynthetic process*, *dTMP biosynthetic process*, *nucleotide biosynthetic process*, *RNA splicing*, *regulation of translational fidelity*), metabolismo de carboidratos (*cellular carbohydrate metabolism*), processos catabólicos de macromoléculas da parede celular (*cell wall macromolecule catabolic process*) e divisão celular (*regulation of DNA replication*). Nota-se que a ocorrência desses termos é coerente com células em intensa atividade de divisão celular, como é esperado para o caráter altura (Figura 16).

Os caracteres pesos de matéria verde e seca, PMS e PMV, apresentaram enriquecimento de termos relacionados à biossíntese de aminoácidos (*glycine biosynthesis*, *methionine biosynthetic process*, *cellular amino acid metabolic process*), biossíntese de lipídios (*steroid biosynthetic process*, *lipid A biosynthetic process*, *fatty acid biosynthetic process*) e transporte de carboidratos (*carbohydrate transport*) (Figuras 23 e 24). Para o caráter florescimento, foram encontrados termos relacionados às modificações de proteínas (*protein deacetylation*, *protein ubiquitination*), ativação e desativação de enzimas (*dephosphorylation*) e silenciamento gênico (*gene silencing*) (Figura 20). Os caracteres ligados à composição de fibra, FDA e FDN, exibiram termos relacionados aos carboidratos (*carbohydrate transport*, *carbohydrate metabolism*, *cell wall macromolecule catabolic process* e *polysaccharide catabolic process*), termos possivelmente relacionados ao transporte de substâncias para fora da célula (*protein targeting to Golgi*, *protein localization*, *protein*

secretion, transmembrane transport). Também foram observados termos relacionados à biosíntese de biomoléculas (*glycine biosynthetic process, histidine biosynthetic process e methionine biosyntheses*) (Figuras 18 e 19).

Termos relacionados aos carboidratos, com destaque para transporte (*carbohydrate transport*), direcionamento de biomoléculas (*retrograde vesicle-mediated transport, Golgi to ER*), ancoramento de proteínas à membrana plasmática (*GPI-anchor metabolism*) e modificações na parede celular (*cell wall modification*), além de termos relacionadas à biosíntese de proteínas e outras moléculas (*glycine biosynthetic process, methionine biosynthetic process e cellular amino acid biosynthetic process*), destacaram-se entre aqueles relacionados ao caráter hemicelulose (Figura 21).

Para o caráter celulose, destacaram-se aqueles relacionados ao transporte de carboidratos e ancoragem à membrana (*carbohydrate transport, protein transport, citrate transport e GPI-anchor metabolic process*), movimento de microtúbulos (*microtubule-based movement*), além de biossíntese de biomoléculas (*methionine biosynthetic process, cellular amino acid biosynthetic process e purine nucleotide biosynthetic process*) (Figura 17). Finalmente, para o caráter lignina destacaram-se aqueles relacionados à exocitose (*exocytosis*), processos catabólicos de macromoléculas da parede celular (*cell wall macromolecule catabolic process*), metabolismo secundário (*secondary metabolism*), ancoragem (*GPI-anchor metabolic process*) e biossíntese de macromoléculas, em especial aminoácidos da família aromática (*aromatic amino acid family biosynthetic process*) (Figura 22).

Por fim, diversos caracteres apresentaram termos relacionados à transdução de sinais (*small GTPase mediated signal transduction, intracellular signal transduction e regulation of ARF protein signal transduction*).

Figura 16 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter altura

Figura 17 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter celulose

Figura 18 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter FDA

Figura 19 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter FDN

Figura 20 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter florescimento

Figura 21 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter hemicelulose

Figura 22 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter lignina

Figura 23 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter PMS

Figura 24 – Representação gráfica “TreeMap” do enriquecimento funcional dos efeitos preditos na análise conjunta dos dois sub-painéis, para o caráter PMV

5 DISCUSSÃO

Atualmente, vários trabalhos que aplicam seleção genômica em diversas espécies vegetais vêm sendo desenvolvidos (BERNARDO; YU, 2007; HEFFNER; SORRELLS; JANNINK, 2009), entretanto, até o presente momento, não foram encontrados relatos na literatura de estudos de seleção genômica aplicados a sorgo biomassa. O presente trabalho mostrou que esta técnica pode ser aplicada com sucesso em programas de melhoramento de sorgo biomassa visando à produção de etanol de segunda geração, apresentando capacidades preditivas altas para os diversos caracteres avaliados, contribuindo assim, para uma seleção mais precoce e eficiente de genótipos superiores.

O surgimento das tecnologias de sequenciamento de nova geração e a possibilidade da descoberta de um grande número de marcadores do tipo SNP, a custos cada vez mais reduzidos, tornou possível a ampla cobertura do genoma com marcadores. Dentre as diversas técnicas de genotipagem disponíveis, o sistema GBS fornece uma ferramenta rápida e de baixo custo (POLAND; RIFE, 2012). Apesar de a elevada quantidade de dados perdidos ser uma das conhecidas desvantagens desta técnica (DAVEY et al., 2011; BEISSINGER et al., 2013), devido ao grande número de marcadores gerados, foi possível encontrar 412.874 marcadores com pequenas proporções de dados perdidos (menos de 20% de genótipos não determinados). Visto que o sorgo é uma espécie diploide e apresenta um genoma de referência bem montado, a imputação dos dados possibilitou o uso da genotipagem via GBS, sem maiores perdas na capacidade preditiva, em concordância com resultados já publicados (HABIER; FERNANDO; DEKKERS, 2009; WEIGEL et al., 2010; DASSONNEVILLE et al., 2011; MULDER et al., 2012). De fato, mesmo após filtragem dos marcadores com MAF mínima de 5%, foi possível obter uma densa cobertura do genoma, com 258.220 SNPs, o que corresponde a aproximadamente 1 SNP a cada 2,8 kpb. Segundo Hamblin et al. (2004), a extensão do DL em sorgo é maior ou igual a 10 kpb, logo, a densidade de marcadores utilizada neste estudo torna bastante provável a existência de pelo menos uma marca em DL com cada QTL de interesse.

Estudos de seleção genômica em plantas têm sido baseados principalmente em dados simulados utilizando populações de melhoramento (BERNARDO; YU, 2007; CROSSA et al., 2010; ZHAO et al., 2013). De maneira contrastante, neste trabalho foram utilizados painéis de diversidade, isto é, coleções de genótipos com ampla variedade genética. Ao se realizar a predição entre painéis de genótipos de origens diversas, objetivou-se reduzir o tempo necessário para selecionar genótipos de sorgo biomassa em fases iniciais do programa de

melhoramento da Embrapa, por meio da predição baseada em genótipos já selecionados de outros programas de melhoramento. Os resultados obtidos foram satisfatórios (Tabela 6), tendo sido obtidas capacidades preditivas medianas para alguns caracteres, a saber: 0,36, 0,35, 0,32, 0,25 e 0,24 para os caracteres PMS, lignina, FDA, PMV e florescimento, respectivamente. Para o caráter altura, foi observada uma capacidade preditiva de 0,66. Entretanto, para os caracteres celulose, FDN e hemicelulose, as capacidades preditivas encontradas foram baixas e até mesmo negativas. De maneira geral, espera-se que caracteres de herança mais complexa, com menor herdabilidade, apresentem menor capacidade preditiva do que caracteres mais simples e de alta herdabilidade. Nos resultados encontrados por Hayes et al. (2010), caracteres complexos controlados por muitos loci de pequenos efeitos apresentaram menores capacidades preditivas do que caracteres menos complexos. Predições genômicas são mais acuradas para caracteres com alta herdabilidade (COMBS; BERNARDO, 2013; LORENZ, 2013). Logo, as correlações entre herdabilidades e capacidades preditivas encontradas (Figura 11) refletem bem o papel desta em estudos de seleção genômica.

Um dos requisitos necessários para o sucesso da seleção genômica é que os genótipos os quais constituem as populações de treinamento e teste sejam altamente aparentados, garantindo assim que os efeitos estimados na população de treinamento sejam consistentes com os existentes na população de teste, permitindo, portanto, uma seleção efetiva (JANNINK; LORENZ; IWATA, 2010; JONAS; KONING, 2013). A predição na análise conjunta dos dois sub-painéis se mostrou mais efetiva, apresentando capacidades preditivas mais elevadas, quando comparada à predição realizada entre sub-painéis. De fato, a análise de componentes principais (Figura 7) refletiu a sub-divisão dos 200 genótipos em dois sub-painéis relacionados à variabilidade dos genótipos, sendo que o sub-painel 1 conta com materiais originados de coleções núcleo do CIRAD (França) e ICRISAT (Índia), além de genótipos de sorgo sacarino e granífero, enquanto o sub-painel 2 é composto principalmente por materiais oriundos do programa de melhoramento da Embrapa Milho e Sorgo. Portanto, a predição entre sub-painéis contou com populações de treinamento e teste menos aparentadas, resultando assim, em menores capacidades preditivas (Tabela 6). Não obstante, valores razoáveis de co-ocorrência de genótipos para seleção entre sub-painéis foram obtidos, em particular para seleção de menor intensidade, para alguns caracteres.

O tamanho e composição da população de treino são importantes fatores os quais podem ser manipulados pelos melhoristas, por meio da união de conjuntos de dados de treinamento, para maximizar o uso de informações fenotípicas e genotípicas disponíveis, gerando assim maiores populações de treinamento e aumentando a capacidade preditiva

(HAYES et al., 2009; DE ROOS; HAYES; GODDARD, 2009; ASORO et al., 2011; LORENZ; SMITH, JANNINK, 2012; TECHNOW; BURGER; MELCHINGER, 2013). Em nosso estudo, ao se realizar a análise conjunta dos dois sub-painéis, uma maior população de treinamento foi utilizada, isto é, todos os 200 genótipos foram combinados em um único grupo de indivíduos, empregando-se a seguir o sistema de validação cruzada para predição dos efeitos. Deste modo, o fato de as capacidades preditivas encontradas (Tabela 4) terem sido superiores quando comparadas àquelas obtidas na análise entre sub-painéis, pode ser parcialmente explicado pelo tamanho da população de treinamento utilizada em cada cenário, bem como à maior similaridade entre genótipos nas populações de treinamento e teste no caso da análise conjunta.

Os efeitos dos marcadores moleculares podem capturar informações de parentesco genético entre indivíduos, bem como informações de desequilíbrio de ligação (HABIER; FERNANDO; DEKKERS, 2007; ZHONG et al., 2009; JANNINK; LORENZ; IWATA, 2010). Avaliando o reflexo do número de marcadores de maiores efeitos na capacidade preditiva dos modelos ajustados na análise conjunta dos dois sub-painéis, as capacidades preditivas máximas foram obtidas utilizando-se pequenas densidades de marcadores, por exemplo, 2.018 marcas para o caráter hemicelulose (Figura 9). É possível que estes marcadores tenham capturado principalmente informações de parentesco genético, ao invés de desequilíbrio de ligação, de modo que a grande densidade de marcadores acrescentou ruído ao modelo. Entretanto, quando não foram selecionadas as marcas de maiores efeitos (Figura 10), as médias obtidas através da redução aleatória do número de marcas comprovaram que as capacidades preditivas máximas foram obtidas utilizando-se a máxima densidade de marcadores, isto é, reduções aleatórias do número de marcas levaram a menores capacidades preditivas. Do ponto de vista prático, isto indica que um número reduzido de marcadores explica grande parte da variação genética, abrindo perspectivas para a utilização de um subconjunto relativamente pequeno dos SNPs de sorgo, assim como é feito para outras espécies vegetais e animais, mediante a construção de chips específicos de SNPs (MATUKUMALLI et al., 2009; YU et al., 2013; WANG et al., 2014). A análise entre sub-painéis exibiu resultados similares quando as marcas foram eliminadas aleatoriamente (Figura 13), mas não quando foram escolhidas as de maiores efeitos (Figura 12).

A importância da fenotipagem de qualidade se torna ainda maior em estudos de seleção genômica, dado que esta tem impacto em todas as etapas do processo, desde a predição dos efeitos dos marcadores na população de treinamento até a seleção dos genótipos superiores. Sendo assim, a precisão com que as medidas fenotípicas são tomadas para cada

caráter reflete nas capacidades preditivas dos modelos (HESLOT; JANNINK; SORRELLS, 2015). O caráter florescimento é de extrema importância para o melhoramento de sorgo biomassa – de fato, a caracterização genética de um sistema regulatório responsável pelo controle do fotoperiodismo e do tempo de florescimento possibilitou o desenvolvimento de híbridos de sorgo biomassa altamente sensíveis ao fotoperíodo e com florescimento tardio, os quais exibem prolongado crescimento vegetativo e alta acumulação de biomassa (ROONEY; AYDIN, 1999; ROONEY et al., 2007; MURPHY et al., 2011). Na prática, sua fenotipagem é fortemente influenciada pelas condições ambientais presentes e pela época de semeadura. Portanto, a baixa capacidade preditiva apresentada para tal caráter na predição entre sub-painéis, igual a 0,22, pode ter sido influenciada pela baixa precisão experimental dos dados fenotípicos obtidos no sub-painel 2 quando comparada à do sub-painel 1, visto que aquele foi avaliado em apenas uma safra, e a época de semeadura do sub-painel 2 pode ter influenciado o florescimento do mesmo. Além disso, é importante ressaltar que o sub-painel 2 é composto por genótipos mais sensíveis ao fotoperíodo do que o sub-painel 1.

Os caracteres FDA, FDN e lignina foram caracterizados por meio de técnicas laboratoriais. Na análise conjunta dos dois sub-painéis estes apresentaram capacidades preditivas altas, iguais a 0,85 para FDA e FDN e 0,82 para lignina. Entretanto, na predição entre sub-painéis, as capacidades preditivas apresentadas foram baixas, iguais a 0,32, 0,22 e 0,35, respectivamente. Isto sugere que dificuldades técnicas na caracterização fenotípica dos materiais, bem como o fato de os dois sub-painéis possuírem bases genéticas distintas, podem ter influenciado os resultados.

Para os caracteres celulose e hemicelulose, as capacidades preditivas encontradas na predição entre os dois sub-painéis foram muito baixas, iguais a 0,25 e 0,012, respectivamente. Na realidade, alguns modelos estatísticos resultaram em valores negativos, como foi o caso dos modelos BayesC e RRBLUP para o caráter hemicelulose. Como as proporções de hemicelulose e celulose foram estimadas pela diferença entre os valores de FDN e FDA, e pelos valores de FDN e lignina, respectivamente, são caracteres derivados, obtidos indiretamente a partir de medidas de outros caracteres. Consequentemente, espera-se que as variâncias das estimativas sejam maiores, justificando assim as baixas capacidades preditivas encontradas.

Apesar das capacidades preditivas relativamente baixas na seleção entre os sub-painéis, a seleção genômica pode ser aplicada utilizando-se intensidades de seleção mais brandas, fornecendo resultados úteis ao melhoramento. No entanto, vale ressaltar que a baixa capacidade preditiva pode reduzir o número de ciclos de seleção possíveis de serem realizados

com o modelo treinado. Ao comparar a seleção fenotípica via BLUP com a seleção genômica por meio do índice de co-ocorrência entre as melhores plantas selecionadas por estas duas metodologias, para diferentes intensidades de seleção, observa-se que os maiores índices de co-ocorrência foram encontrados para seleção mais branda e não para seleção intensa (Figuras 14 e 15). Não obstante, a utilização da seleção genômica para descarte dos genótipos inferiores seria de grande utilidade, por promover economia de tempo e de recursos em programas de melhoramento, em particular em suas fases iniciais.

Os índices de co-ocorrência encontrados entre os diversos modelos de seleção genômica ajustados foram muito altos, indicando que, apesar de cada modelo apresentar diferentes suposições a respeito da distribuição subjacente dos efeitos dos QTLs, a diferença entre as plantas selecionadas por eles foi muito pequena. Em estudo realizado com a cultura do milho, Bernardo e Yu (2007) relataram que os modelos Bayesianos exibiram pouca ou até mesmo nenhuma vantagem, quando comparados ao modelo RRBLUP. Devido à menor complexidade computacional e maior rapidez de execução do modelo RRBLUP, sua aplicação corriqueira em programas de melhoramento pode ser considerada uma alternativa viável. Todavia, é interessante que diversos modelos de seleção genômica sejam ajustados e comparados, para que a seleção seja realizada baseando-se no melhor modelo para cada caráter. Entretanto, limitações de recursos computacionais e tempo podem inviabilizar tal comparação e exigir a aplicação de um modelo mais simples, como o RRBLUP, o qual pode gerar resultados satisfatórios em um programa de melhoramento.

Uma das principais vantagens de implementar a seleção genômica em programas de melhoramento é a redução do tempo necessário para lançamento de novos cultivares. A seleção genômica pode reduzir o ciclo de melhoramento por meio da predição precoce do desempenho fenotípico de um conjunto de genótipos, para diversos caracteres de interesse, reduzindo o tempo necessário para liberação de novos materiais (MEUWISSEN; HAYES; GODDARD, 2001; BERNARDO; YU, 2007; BERNARDO, 2008; CROSSA et al., 2011). Ademais, além de reduzir o tempo de um ciclo de melhoramento, a seleção genômica também pode ser usada para prever o desempenho fenotípico de genótipos com relação a caracteres de caracterização da biomassa, como FDA, FDN e lignina, os quais possuem avaliações fenotípicas caras e laboriosas.

A seleção genômica vem revolucionando o uso da seleção assistida por marcadores moleculares no melhoramento de plantas, principalmente devido à sua abordagem diferenciada quando comparada a estudos de mapeamento de QTL e GWAS. O mapeamento de QTLs tem como objetivo mapear regiões cromossômicas que afetam caracteres fenotípicos

de interesse, a fim de possibilitar a exploração de marcadores ligados a estes (BERNARDO, 2008; LORENZ et al., 2011). Porém, o fato de o mapeamento de populações biparentais usados na maioria dos estudos de QTL não ter aplicação direta em programas de melhoramento, além de os métodos estatísticos usados para identificar locos de interesse e realizar a seleção assistida por marcadores serem inadequados para o melhoramento de caracteres poligênicos controlados por muitos locos de pequeno efeito, limitam o uso desta técnica em programas de melhoramento (MEUWISSEN; HAYES; GODDARD, 2001; GODDARD; HAYES, 2007; HEFFNER; SORRELLS; JANNINK, 2009). Os estudos de GWAS, assim como a estratégia da seleção genômica, baseiam-se na ampla cobertura do genoma por marcadores, mas têm como objetivo principal a identificação de regiões cromossômicas associadas a um determinado caráter, ou seja, visam ao estudo da arquitetura genética dos caracteres (INGVARSSON; STREET, 2011; HUANG; HAN, 2014). A seleção genômica não visa à avaliação da influência de genes e/ou marcadores individuais, mas trata o agregado de todo o genoma de maneira semelhante ao que a genética quantitativa tem feito historicamente, porém utilizando informações de marcadores densamente distribuídos por todo o genoma, ou seja, tem como objetivo a seleção aplicada diretamente ao melhoramento de plantas (BERNARDO; YU, 2007; DEKKERS, 2007; GODDARD; HAYES, 2007; HABIER; FERNANDO; DEKKERS, 2007).

Apesar disso, é concebível que marcadores localizados próximos a genes responsáveis por determinadas funções biológicas apresentem efeitos de maior magnitude, fato este que não foi ainda relatado na literatura científica. Assim, os efeitos preditos dos marcadores em estudos de seleção genômica podem ser utilizados para análises de enriquecimento funcional, a fim de verificar a existência de efeitos maiores para um ou outro determinado grupo funcional. Neste trabalho, os efeitos preditos dos marcadores na análise conjunta dos dois subpainéis foram utilizados para realizar o estudo de enriquecimento funcional, de tal maneira que os resultados dos modelos ajustados também fornecessem informações biológicas a respeito do controle genético dos caracteres de interesse, bem como do modo com que a seleção genômica atua sobre os genomas dos indivíduos.

Os resultados encontrados para os diversos caracteres avaliados sugeriram associações interessantes, as quais devem ser investigadas com maiores detalhes em estudos futuros. Por exemplo, foram observados termos relacionados aos carboidratos, com destaque para transporte (*carbohydrate transport*), direcionamento (*retrograde vesicle-mediated transport, Golgi to ER*), ancoramento de proteínas à membrana plasmática (*GPI-anchor metabolism*) e modificações na parede celular (*cell wall modification*) para o caráter hemicelulose (Figura

21). Levando-se em conta o fato de que, após sintetizada no complexo de Golgi, a hemicelulose é transportada para a membrana plasmática, a presença de termos relacionados ao transporte está em concordância com o esperado (PAULY et al., 2013). A presença de termos relacionados ao movimento de microtúbulos (*microtubule-based movement*) para o caráter celulose (Figura 17) possivelmente tem ligação com o fato de que, na síntese de celulose, sua deposição é guiada por microtúbulos que estão adjacentes ou diretamente conectados ao complexo de síntese (DELMER; AMOR, 1995). A biossíntese de lignina se inicia no citosol com a síntese de monolignóis glicosilados a partir do aminoácido fenilalanina, o qual é um aminoácido aromático, justificando assim a presença de termos relacionados a processos biossintéticos de aminoácidos da família aromática (Figura 22) (BOERJAN; RALPH; BAUCHER, 2003).

Diversos termos relacionados à biossíntese de macromoléculas foram detectados para os diversos caracteres fenotípicos estudados, o que fornece evidências de que tais fenótipos são, ao menos parcialmente, resultado da maior atividade metabólica dos indivíduos. A seleção dos genótipos de acordo com seus GEBVs, indiretamente, atuaria mais fortemente sobre esses SNPs, visto que seus efeitos foram de maior magnitude que os demais. Assim sendo, a abordagem de enriquecimento funcional tenta conciliar a aplicação da seleção genômica no melhoramento genético à elucidação da arquitetura genética dos caracteres, ficando situada entre os extremos da seleção genômica e dos estudos de mapeamento de QTL, GWAS e seleção assistida por marcadores.

Os diversos modelos de seleção genômica aqui empregados deram origem a resultados satisfatórios e diretamente aplicáveis ao melhoramento genético, potencialmente capazes de reduzir o tempo necessário para o lançamento de novas cultivares de sorgo biomassa, aumentando o potencial de esta se destacar como uma importante cultura bioenergética. Por fim, as análises de enriquecimento funcional abrem perspectivas para o uso da seleção genômica na elucidação da arquitetura genética de caracteres quantitativos.

6 CONCLUSÕES

Os métodos de seleção genômica aplicados ao painel de sorgo biomassa resultaram em capacidades preditivas satisfatórias. Entretanto, quando aplicados à análise entre sub-painéis, resultaram em menores capacidades preditivas, sendo, portanto, indicados para realizar o descarte dos genótipos inferiores. As capacidades preditivas mostraram-se robustas na condição de menor disponibilidade de marcadores moleculares. Contudo, a utilização de grandes números de marcas permitem o ajuste de modelos com subconjuntos das mesmas, por exemplo, a partir da seleção daquelas de maior efeito, com potenciais ganhos nas capacidades preditivas. Por fim, os efeitos preditos dos marcadores em estudos de seleção genômica podem ser utilizados para análises de enriquecimento funcional, de modo a fornecer informações biológicas relevantes a respeito dos modelos ajustados.

REFERÊNCIAS

ANTONOPOULOU, G.; GAVALA, H.N.; SKIADAS, I.V.; ANGELOPOULOS, K.; LYBERATOS, G. Biofuels generation from sweet sorghum : Fermentative hydrogen production and anaerobic digestion of the remaining biomass. **Bioresource Technology**, Essex, v. 99, p. 110–119, Jan. 2007.

ASORO, F.G.; NEWELL, M.A.; BEAVIS, W.D.; SCOTT, M.P.; JANNINK, J.-L. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. **The Plant Genome Journal**, Madison, v. 4, p. 132-144, July 2011.

ASSEFA, Y.; STAGGENBORG, S.A. Grain sorghum yield with hybrid advancement and changes in agronomic practices from 1957 through 2008. **Agronomy Journal**, Madison, v. 102, n. 2, p. 703–706, Mar. 2010.

BEISSINGER, T.M.; HIRSCH, C.N.; SEKHON, R.S.; FOERSTER, J.M.; JOHNSON, J.M.; MUTTONI, G.; VAIÇÇANCOURT, B.; BUELL, C.R.; KAEPLER, S.M.; DE LEON, N. Marker density and read depth for genotyping populations using genotyping-by-sequencing. **Genetics**, London, v. 193, n. 4, p. 1073–1081, Apr. 2013.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society B**, Oxford, England, v. 57, n. 1, p. 289–300, Mar. 1995.

BERNARDO, R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. **Crop Science**, Madison, v. 48, n. 5, p. 1649–1664, Sept. 2008.

_____. Genomewide selection for rapid introgression of exotic germplasm in maize. **Crop Science**, Madison, v. 49, n. 2, p. 419–425, Mar. 2009.

_____. **Essentials of plant breeding**. Woodbury: Stemma, 2014a. 260 p.

_____. Genomewide selection when major genes are known. **Crop Science**, Madison, v. 54, n. 1, p. 68, Jan. 2014b.

BERNARDO, R.; YU, J. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, Madison, v. 47, n. 3, p. 1082–1090, May 2007.

BOERJAN, W.; RALPH, J.; BAUCHER, M. Lignin biosynthesis. **Annual review of plant biology**, Palo Alto, v. 54, p. 519-546, Jan. 2003.

BUENO, L.C.S.; MENDES, A.N.G.; CARVALHO, S.P.C. **Melhoramento genético de plantas: princípios e procedimentos**. 2.ª ed. Lavras: Editora UFLA, 2006. 319 p.

COLLARD, B.C.Y.; JAHUFER, M.Z.Z.; BROUWER, J.B.; PANG, E.C.K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. **Euphytica**, Wageningen, v. 142, n. 1/2, p. 169–196, Feb. 2005.

COMBS, E.; BERNARDO, R. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. **The Plant Genome Journal**, Madison, v. 6, p. 1-7, Mar. 2013.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Levantamentos de safra: grãos**. Disponível em: <<http://www.conab.gov.br/conteudos.php?a=1253&t=>>. Acesso em: 10 jan. 2015.

CROSSA, J.; PÉREZ, P.; DE LOS CAMPOS, G.; MAHUKU, G.; DREISIGACKER, S.; MAGOROKOSHO, C. Genomic selection and prediction in plant breeding. **Journal of Crop Improvement**, London, v. 25, n. 3, p. 239–261, Apr. 2011.

CROSSA, J.; PÉREZ, P.; HICKEY, J.; BURGUEÑO, J.; OMELLA, L.; CERÓN-ROJAS, J.; ZHANG, X.; DREISIGACKER, S.; BABU, R.; LI, Y.; BONNETT, D.; MATHEWS, K. Genomic prediction in CIMMYT maize and wheat breeding programs. **Heredity**, London, v. 112, n. 1, p. 48–60, Jan. 2014.

CROSSA, J.; BEYENE, Y.; KASSA, S.; PÉREZ, P.; HICKEY, J.M.; CHEN, C.; DE LOS CAMPOS, G.; BURGUEÑO, J.; WINDHAUSEN, V.S.; BUCKLER, E.; JANNINK, J.-L.; LOPEZ CRUZ, M.; BABU, R. Genomic prediction in maize breeding populations with genotyping-by-sequencing. **G3**, Bethesda, v. 3, p. 1903–1926, Nov. 2013.

CROSSA, J.; CAMPOS, G.D.L.; PÉREZ, P.; GIANOLA, D.; BURGUEÑO, J.; ARAUS, J.L.; ARAUS, J.L.; MAKUMBI, D.; SINGH, R.P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BANZIGER, M.; BRAUN, H.-J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, London, v. 186, n. 2, p. 713–24, Jan. 2010.

DASSONNEVILLE, R.; BRØNDUM, R.F.; DRUET, T.; FRITZ, S.; GUILLAUME, F.; GULDBRANDTSEN, B.; LUND, M.S.; DUCROCQ, V.; SU, G. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. **Journal of Dairy Science**, Champaign, v. 94, n. 7, p. 3679–3686, Apr. 2011.

DAVEY, J.W.; HOHENLOHE, P.A.; ETTER, P.D.; BOONE, J.Q.; CATCHEN, J.M.; BLAXTER, M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature reviews. Genetics**, London, v. 12, n. 7, p. 499–510, July 2011.

DE DONATO, M.; PETERS, S.O.; MITCHELL, S.E.; HUSSAIN, T.; IMUMORIN, I.G. Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. **PloS one**, San Francisco, v. 8, n. 5, p. 1-10, May 2013.

DE LOS CAMPOS, G.; HICKEY, J.M.; PONG-WONG, R.; DAETWYLER, H.D.; CALUS, M.P.L. Whole genome regression and prediction methods applied to plant and animal breeding. **Genetics**, London, v. 193, n. 2, p. 327–345, Feb. 2013.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J.M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, London, v. 182, p. 375–385, Mar. 2009.

DE ROOS, A.P.W.; HAYES, B.J.; GODDARD, M.E. Reliability of genomic predictions across multiple populations. **Genetics**, London, v. 183, n. 4, p.1545-1553, Dec. 2009.

DEKKERS, J.C. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, Champaign, v. 82, p.313-328, Feb. 2004.

DEKKERS, J.C.M. Prediction of response to marker-assisted and genomic selection using selection index theory. **Journal of Animal Breeding and Genetics**, Berlin, v. 124, n. 1961, p. 331–341, Oct. 2007.

DELMER, D.P.; AMOR, Y. Cellulose biosynthesis. **The Plant cell**, Rockville, v. 7, n. 7, p. 987-1000, Oct. 1995.

DESCHAMPS, S.; LLACA, V.; MAY, G.D. Genotyping-by-sequencing in plants. **Biology**, Basel, Switzerland, v. 1, n. 3, p. 460–483, Sept. 2012.

DESTA, Z.A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in Plant Science**, Oxford, v. 19, n. 9, p. 592–601, June 2014.

DOGGETT, H. **Sorghum**. 2nd ed. New York: Longman Scientific & Technical, 1988. 512 p.

ELSHIRE, R.J.; GLAUBITZ, J.C.; SUN, Q.; POLAND, J.A.; KAWAMOTO, K.; BUCKLER, E.S.; MITCHELL, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS one**, San Francisco, v. 6, p. 1–10, May 2011.

EMBRAPA. Sorgo sacarino: tecnologia agronomica e industrial para alimentos e energia. **Agroenergia em Revista**, Brasília, v. 3, p. 52, ago. 2011.

_____. **Sistemas de produção**. Disponível em: <<http://www.cnpms.embrapa.br/>>. Acesso em: 09 jan. 2015.

ENDELMAN, J.B. ridge regression and other kernels for genomic selection with r package rrBLUP. **The Plant Genome Journal**, Madison, v. 4, p. 250-255, Nov. 2011.

FAO. **FAOSTAT**: production. Disponível em: <<http://faostat3.fao.org/compare/E>>. Acesso em: 10 jan. 2015.

FLINT-GARCIA, S.A.; THORNSBERRY, J.M.; BUCKLER, E.S. Structure of linkage disequilibrium in plants. **Annual Review of Plant Biology**, Palo Alto, v. 54, p. 357–374, Jan. 2003.

FRANK, J.; MASSEY, J. The Kolmogorov-Smirnov test for goodness of fit. **Journal of the American Statistical Association**, Philadelphia, v. 46, n. 253, p. 68–78, 1951.

GIANOLA, D.; DE LOS CAMPOS, G.; HILL, W.G.; MANFREDI, E.; FERNANDO, R. Additive genetic variability and the Bayesian alphabet. **Genetics**, London, v. 183, p. 347–363, Sept. 2009.

GLAUBITZ, J.C.; CASSTEVEN, T.M.; LU, F.; HARRIMAN, J.; ELSHIRE, R.J.; SUN, Q.; BUCLKER, E.S. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. **PLoS one**, San Francisco, v. 9, n. 2, p. 1–11, Feb. 2014.

GODDARD, M.E.; HAYES, B.J. Genomic selection. **Journal of Animal Breeding and Genetics**, Berlin, v. 124, p. 323–330, Oct. 2007.

_____. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, London, v. 10, p. 381–391, June 2009.

GOODSTEIN, D.M.; SHU, S.; HOWSON, R.; NEUPANE, R.; HAYES, R.D.; FAZO, J.; MITROS, T.; DIRKS, W.; HELLSTEN, U.; PUTNAM, N.; ROKHSAR, D.S. Phytozome: a comparative platform for green plant genomics. **Nucleic Acids Research**, New York, v. 40, p. 1178–1186, Nov. 2012.

GRATTAPAGLIA, D.; RESENDE, M.D.V. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, Berlin, v. 7, n. 2, p. 241–255, Oct. 2010.

GUPTA, P.K.; RUSTGI, S.; KULWAL, P.L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. **Plant Molecular Biology**, Dordrecht, v. 57, n. 4, p. 461–485, Jan. 2005.

HABIER, D.; FERNANDO, R.L.; DEKKERS, J.C.M. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, London, v. 177, n. 4, p. 2389–2397, Dec. 2007.

_____. Genomic selection using low-density marker panels. **Genetics**, London, v. 182, n. 1, p. 343–353, May 2009.

HABIER, D.; FERNANDO, R.L.; KIZILKAYA, K.; GARRICK, D.J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, London, v. 12, n. 1, p. 186, May 2011.

HABYARIMANA, E.; LAURETI, D.; NINNO, M. de; LORENZONI, C. Performances of biomass sorghum [*Sorghum bicolor* (L .) Moench] under different water regimes in Mediterranean region. **Industrial Crops and Products**, New York ,v. 20, p. 23–28, Dec. 2004.

HAMBLIN, M.T.; MITCHELL, S.E.; WHITE, G.M.; GALLEGU, J.; KUKATLA, R.; WING, R.A.; PATERSON, A.H.; KRESOVICH, S. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. **Genetics**, London, v. 167, p. 471–483, May 2004.

HANSEY, C.N.; VAILLANCOURT, B.; SEKHON, R.S.; DE LEON, N.; KAEPLER, S.M.; BUELL, C.R. Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. **PloS one**, San Francisco, v. 7, n. 3, p. 1–10, Mar. 2012.

HAYES, B.J.; BOWMAN, P.J.; CHAMBERLAIN, A.C.; GODDARD, M.E. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, Champaign, v. 92, p.433-443, Oct. 2009.

HAYES, B.J.; PRYCE, J.; CHAMBERLAIN, A.J.; BOWMAN, P.J.; GODDARD, M.E. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage and type in holstein cattle as contrasting model traits. **PloS Genetics**, San Francisco, v. 6, n.9, p. 1-11, Sept. 2010.

HEFFNER, E.L.; SORRELLS, M.E.; JANNINK, J.-L. Genomic selection for crop improvement. **Crop Science**, Madison, v. 49, n. 1, p. 1–12, Jan. 2009.

HENDERSON, C. **Applications of linear models in animal breeding**. Guelph: University of Guelph, 1984. 462 p.

HESLOT, N.; JANNINK, J.-L.; SORRELLS, M.E. Perspectives for genomic selection applications and research in plants. **Crop Science**, Madison, v. 55, p. 1–12, Jan. 2015.

HONS, F.M.; MORESCO, R.F.; WIEDENFELD, R.P.; COTHREN, J.T. Applied nitrogen and phosphorus effects on yield and nutrient uptake by high-energy sorghum produced for grain and biomass. **Agronomy Journal**, Madison, v. 78, p. 1069–1078, Nov. 1986.

HUANG, X.; HAN, B. Natural variations and genome-wide association studies in crop plants. **Annual Review of Plant Biology**, Palo Alto, v. 65, p. 531-551, Jan. 2014.

INGVARSSON, P.K.; STREET, N.R. Association genetics of complex traits in plants. **New Phytologist**, Cambridge, v. 189, n. 4, p. 909-992, Mar. 2011.

JANNINK, J.-L.; LORENZ, A.J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**, London, v. 9, n. 2, p. 166–177, Feb. 2010.

JIANG, C.; ZENG, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. **Genetics**, London, v. 140, p. 1111–27, July 1995.

JONAS, E.; KONING, D.-J. Does genomic selection have a future in plant breeding? **Trends in biotechnology**, Amsterdam, v. 31, n. 9, p. 497–504, Sept. 2013.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: FOURTEENTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995, Montreal. **Proceedings...** San Francisco: Morgan Kaufmann Publ., 1995. v. 2, p. 1137-1143.

KOLBEHDARI, D.; SCHAEFFER, L.R.; ROBINSON, J.A.B. Estimation of genome-wide haplotype effects in half-sib designs. **Journal of Animal Breeding and Genetics**, Berlin, v. 124, p. 356–361, Oct. 2007.

KÖNIG, S.; SWALVE, H.H. Application of selection index calculations to determine selection strategies in genomic breeding programs. **Journal of Dairy Science**, Champaign, v. 92, n. 10, p. 5292–5303, July 2009.

LEE, S.H.; VAN DER WERF, J.H.J.; HAYES, B.J.; GODDARD, M.E.; VISSCHER, P.M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS Genetics**, San Francisco, v. 4, n. 10, p. 1–11, Oct. 2008.

LEGARRA, A.; MISZTAL, I. Technical note: computing strategies in genome-wide selection. **Journal of Dairy Science**, Champaign v. 91, p. 360–366, Jan. 2008.

LEGARRA, A.; ROBERT-GRANIÉ, C.; MANFREDI, E.; ELSEEN, J.M. Performance of genomic selection in mice. **Genetics**, London, v. 180, p. 611–618, Sept. 2008.

LI, Z.; SILLANPÄÄ, M.J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and Applied Genetics**, Berlin, v. 125, p. 419–435, Nov. 2012.

LIU, B. **Statistical genomics: linkage, mapping and QTL analysis**. Boca Raton: CRC Press, 1998. 611 p.

LONG, N.; GIANOLA, D.; ROSA, G.J.M.; WEIGEL, K.A.; AVENDAÑO, S. Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, Berlin, v. 124, p. 377–389, Sept. 2007.

LORENZ, A.J. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. **G3**, Bethesda v. 3, p. 481–491, Mar. 2013.

LORENZ, K.; COHEN, B.A. Small- and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency. **Genetics**, London, v. 192, p. 1123–1132, Nov. 2012.

LORENZ, A.J.; SMITH, K.P.; JANNINK, J.–L. Potential and optimization of genomic selection for *Fusarium* head blight resistance in six-row barley. **Crop Science**, Madison, v. 52, p. 1610–1621, July 2012.

LORENZ, A.J.; CHAO, S.; ASORO, F.G.; ELLIOT, L.H.; HAYASHI, T.; IWATA, H.; SMITH, K.P.; SORRELLS, M.E.; JANNINK, J.–L. Genomic selection in plant breeding : knowledge and prospects. **Advances in Agronomy**, San Diego, v. 110, p. 77–123, June 2011.

LORENZANA, R.E.; BERNARDO, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and Applied Genetics**, Berlin, v. 120, p. 151–161, Feb. 2009.

LU, F.; LIPKA, A.E.; GLAUBITZ, J.; ELSHIRE, R.; CHERNEY, J.H.; CASLER, M.D.; BUCKLER, E.S.; COSTICH, D.E. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. **PLoS genetics**, San Francisco, v. 9, n. 1, p. 1–14, Jan. 2013.

MARCHINI, J.; HOWIE, B. Genotype imputation for genome-wide association studies. **Nature Genetics**, London, v. 11, n. 7, p. 499–511, July 2010.

MARCHINI, J.; HOWIE, B.; MYERS, S.; MCVEAN, G.; DONNELLY, P. A new multipoint method for genome-wide association studies by imputation of genotypes. **Nature Genetics**, London, v. 39, n. 7, p. 906–913, July 2007.

MASSMAN, J.M.; JUNG, H.J.G.; BERNARDO, R. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. **Crop Science**, Madison, v. 53, p. 58–66, Jan. 2013.

MATUKUMALLI, L.K.; LAWLEY, C.T.; SCHABEL, R.D.; TAYLOR, J.F.; ALLAN, M.F.; HEATON, M.P.; O'CONNELL, J.; MOORE, S.S.; SMITH, T.P.L.; SONSTEGARD, T.S.; VAN TASSELL, C.P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, San Francisco, v. 4, n. 4, p. 1-13, Apr. 2009.

MAYOR, P.J.; BERNARDO, R. Genomewide selection and marker-assisted recurrent selection in doubled haploid versus F populations. **Crop Science**, Madison, v. 49, p. 1719–1725, Apr. 2009.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, London, v. 157, p. 1819–1829, Apr. 2001.

_____. Accelerating improvement of livestock with genomic selection. **Annual Review of Animal Biosciences**, Palo Alto, v. 1, p. 221–37, 2013.

MONK, R.L.; MILLER, F.R.; MCBEE, G.G. Sorghum improvement for energy production. **Biomass**, London, v. 6, p. 145–153, June 1984.

MORRIS, G. P.; RAMU, P.; DESHPANDE, S. P.; HASH, C.T.; SHAH, T.; UPADHYAYA, H.D. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **PNAS**, Washington, v. 110, n. 2, p. 453–458, Jan. 2013.

MULDER, H.A.; CALUS, M.P.L.; DRUET, T.; SCHROOTEN, C. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. **Journal of Dairy Science**, Champaign, v. 95, n. 2, p. 876–889, Oct. 2012.

MURPHY, R.L.; KLEIN, R.R.; MORISHIGE, D.T.; BRADY, J.; ROONEY, W.; MILLER, F.R.; DUGAS, D.V.; KLEIN, P.E.; MULLET, J.E. Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. **Proceedings of the National Academy of Sciences**, Washington, v. 108, n. 39, p. 16469–16474, Sept. 2011.

MURRAY, S.C.; SHARMA, A.; ROONEY, W.L.; KLEIN, P.E.; MULLET, J.E.; MITCHELL, SE.; KRESOVICH, S. Genetic improvement of sorghum as a biofuel feedstock : I. QTL for stem sugar and grain nonstructural carbohydrates. **Crop Science**, Madison, v. 48, p. 2165–2179, Dec. 2008.

NAKAYA, A.; ISOBE, S.N. Will genomic selection be a practical method for plant breeding? **Annals of Botany**, London, v. 110, n. 6, p. 1303–1316, May 2012.

OLSON, S.N.; RITTER, K.; ROONEY, W.; KEMANIAN, A.; MCCARL, A.B.; ZHANG, Y.; HALL, S.; PACKER, D.; JOHN, M. High biomass yield energy sorghum: developing a genetic model for C4 grass bioenergy. **Biofuels, Bioproducts and Biorefining**, Chichester, v. 6, p. 640–655, June. 2012.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, New York, v. 103, p. 681-686, June 2008.

PATERSON, A.H.; LANDER, E.S.; HEWITT, J.D.; PETERSON, S.; LINCOLN, S.E.; TANKSLEY, S.D. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. **Nature**, London, v. 335, p. 721-726, Oct. 1988.

PATERSON, A.H.; BOWERS, J.E.; BRUGGMANN, R.; DUBCKAK, I.; GRIMWOOD, J.; GUNDLACH, H.; HABERER, G.; HELLSTEN, U.; MITROS, T.; POLIAKOV, A.; SCHMUTZ, J.; SPANNAGL, M.; TANG, H.; WANG, X.; BHARTI, A.K.; CHAPMAN, J.; FELTUS, F.A.; GOWIK, U.; GROGORIEV, I.V.; LYONS, E.; MAHER, C.; MARTIS, M.; NARECHANIA, A.; OTILLAR, R.P.; PENNING, B.W.; SALAMOV, A.; WANG, Y.; ZHANG, L.; CARPITA, N.C.; FREELING, M.; GINGLE, A.R.; HASH, C.T.; KELLER, B.; KLEIN, P.; KRESOVICH, S.; MCCANN, M.C.; MING, R.; PETERSON, D.G.; MEHBOOB, R.; WARE, D.; MAYER, K.F.X.; MESSING, J.; ROKHSAR, D.S. The Sorghum bicolor genome and the diversification of grasses. **Nature**, London, v. 457, n. 7229, p. 551–556, Jan. 2009.

PAYNE, R. ; WELHAM, S.; HARDING, S. **A guide to REML in GenStat**. 16th ed. Rothamsted: VSN International, 2013. 97 p.

PAULY, M.; GILLE, S.; LIU, L.; MANSOORI, N.; SOUZA, A.; SCHULTINK, A. Hemicellulose biosynthesis. **Planta**, Berlin, v. 238, n. 4, p. 627-642, June 2013.

PÉREZ, P.; DE LOS CAMPOS, G. **BGLR**: a statistical package for whole genome regression and prediction; R package version 1.0.3. 2014a.

. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, London, v. 198, p. 483-495, Oct. 2014b.

POLAND, J.A.; RIFE, T.W. Genotyping-by-sequencing for plant breeding and genetics. **The Plant Genome Journal**, Madison, v. 5, n. 3, p. 92–102, Nov. 2012.

POLAND, J.A.; BROWN, P.J.; SORRELLS, M.E.; JANNINK, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **PloS one**, San Francisco, v. 7, n. 2, p. 1–8, Feb. 2012.

R FOUNDATION FOR STATISTICAL COMPUTING. **R**: a language and environmet for statistical computing. Vienna: R Core Team, 2013.

RAFALSKI, A.; MORGANTE, M. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. **Trends in Genetics**, London, v. 20, n. 2, p. 103–111, Feb. 2004.

REDDY, B.V.S.; RAMESH, S.; REDDY, P.S. Sorghum breeding research at ICRISAT-goals, strategies, methods and accomplishments. **International Sorghum and Millets Newsletter**, Patancheru, v. 45, p. 5–12, Apr. 2004.

RESENDE, M.D.V.; JUNIOR, M.F.R.R.; AGUIAR, A.M.; ABAD, J.I.M.; MISSIAGGIA, A.A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. **Computação da Seleção Genômica Ampla (GWS)**. Colombo: EMBRAPA, 2010. 79 p. (Comunicado Técnico, 210).

ROBERT, C.P.; CASELLA, G. **Monte Carlo statistical methods**. New York: Springer, 2004. 580 p.

ROBERTS, A.; MCMILLAN, L.; WANG, W.; PARKER, J.; RUSYN, I.; THREADGILL, D. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. **Bioinformatics**, London, v. 23, p. 401–407, Feb. 2007.

ROONEY, W.L.; AYDIN, S. Genetic control of a photoperiod-sensitive response in *Sorghum bicolor* (L.) Moench. **Crop Science**, Madison, v. 39, p. 397–400, Mar. 1999.

ROONEY, W.L.; BLUMENTHAL, J.; BEAN, B.; MULLET, J.E. Designing sorghum as a dedicated bioenergy feedstock. **Biofuels, Bioproducts and Biorefining**, Chichester, v. 1, p. 147–157, July 2007.

SCHIMIDT, A.A.P. **O sorgo**. São Paulo: Ícone 1987. 63 p.

SCHWARZ, G. Estimating the dimension of a model. **Annals of Statistics**, Philadelphia v. 6, p. 461–464, Mar. 1978.

SONAH, H.; BASTIEN, M.; IQUIRA, E.; TARDIVEL, A.; LÉGARÉ, A.; BOYLE, B.; NORMANDEAU, E.; LAROCHE, J.; LAROSE, S.; JEAN, M.; BELZILE, F. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. **PloS One**, San Francisco, v. 8, n. 1, p. 1–9, Jan. 2013.

STACKLIES, W.; REDESTIG, H.; SCHOLZ, M.; WALTHER, D.; SELBIG, J. pcaMethods - A bioconductor package providing PCA methods for incomplete data. **Bioinformatics**, London, v. 23, n. 9, p. 1164–1167, Mar. 2007.

SUPEK, F.; BOŠNJAK, M.; ŠKUNCA, N.; ŠMUC, T. Revigo summarizes and visualizes long lists of gene ontology terms. **PloS One**, San Francisco, v. 6, n. 7, p. 1–9, 2011.

TECHNOW, F.; BURGER, A.; MELCHINGER, A. E. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. **G3**, Bethesda, v. 3, p. 197–203, Feb. 2013.

TIBSHIRANI, R. Regression selection and shrinkage via the Lasso. **Journal of the Royal Statistical Society B**, Oxford, v. 58, p. 267–288, Jan. 1996.

VANRADEN, P.M.; TASSELL, C.P. van; WIGGANS, G.R.; SONSTEGARD, T.S.; SCHNABEL, R.D.; TAYLOR, J.F.; SCHENKEL, F.S. Invited review: reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, Champaign, v. 92, n. 1, p. 16–24, Aug. 2009.

VERMERRIS, W.; SABALLOS, A.; EJETA, G.; MOSIER, N.S.; LADISCH, M.R.; CARPITA, N.C. Molecular breeding to enhance ethanol production from corn and sorghum stover. **Crop Science**, Madison, v. 47, p. 143–153, Dec. 2007.

WANG, S.; WONG, D.; FORREST, K.; ALLE, A.; CHAO, S.; HUANG, B.E.; MACCAFERRI, M.; SALVI, S.; MILNER, S.G.; CATTIVELLI, L.; MASTRANGELO, A.M.; WHAN, A.; STEPHEN, S.; BARKER, G.; WIESEKE, R.; PLIESKE, J.; LILLEMO, M.; MATHER, D.; APPELS, R.; DOLFERUS, R.; BROWN-GUEDIRA, G.; KOROL, A.; AKHUNOVA, A.R.; FEUILLET, C.; SALSE, J.; MORGANTE, M.; POZNIAK, C.; LUO, M.C.; DVORAK, J.; MORELL, M.; DUBCOVSKY, J.; GANAL, M.; TUBEROSA, R.; LAWLEY, C.; MIKOULITCH, I.; CAVANAGH, C.; EDWARDS, K. J.; HAYDEN, M.; AKHUNOV, E. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. **Plant Biotechnology Journal**, Oxford, v. 12, n. 6, p. 787–796, Apr. 2014.

WARD, J.A.; BHANGOO, J.; FERNÁNDEZ-FERNÁNDEZ, F.; MOORE, P.; SWANSON, J.D.; VIOLA, R.; VELASCO, R.; BASSIL, N.; WEBER, C.A.; SARGENTE, D.J. Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. **BMC Genomics**, London, v. 14, p. 114–125, Jan. 2013.

WEIGEL, K.A.; DE LOS CAMPOS, G.; VAZQUEZ, A.I.; ROSA, G.J.M.; GIANOLA, D.; VAN TASSEL, C.P. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. **Journal of Dairy Science**, Champaign, v. 93, n. 11, p. 5423–35, July 2010.

WET, J.M. .; HUCKABAY, J. The origin of *Sorghum bicolor*. II. Distribution and domestication. **Evolution**, Lawrence, v. 21, n. 4, p. 787–802, Dec. 1967.

YOUENS-CLARK, K.; BUCKLER, E.; CASSTEVENS, T.; CHEN, C.; DECLERCK, G.; DERWENT, P.; DHARMAWARDHANA, P.; JAISWAL, P.; KERSEY, P.; KARTHIKEYAN, A.S.; LU, J.; MCCOUCH, S.R.; REN, L.; SPOONER, W.; STEIN, J.C.; THOMASON, J.; WEI, S.; WARE, D. Gramene database in 2010: updates and extensions. **Nucleic Acids Research**, Oxford, v. 39, p. 1085-1094, Jan. 2010. Disponível em: <[http://www.gramene.org/db/qtl/qtl display?query=&vocabulary=qtl&search box name=query&search box id=qtl search for&search field=&species=Sorghum+bicolor&x=11&y=14](http://www.gramene.org/db/qtl/qtl%20display?query=&vocabulary=qtl&search%20box%20name=query&search%20box%20id=qtl%20search%20for&search%20field=&species=Sorghum+bicolor&x=11&y=14)>. Acesso em: 23 ago. 2014.

YU, H.; XIE, W.; LI, J.; ZHOU, F.; ZHANG, Q. A whole-genome SNP array (RICE6K) for genomic breeding in rice. **Plant Biotechnology Journal**, Oxford, v. 12, n. 1, p. 28–37, Mar. 2013.

ZENG, Z. B.; KAO, C. H.; BASTEN, C. . Estimating the genetic architecture of quantitative traits. **Genetical Research**, London, v. 74, p. 279–289, Dec. 1999.

ZENG, Z. **Statistical methods for mapping quantitative trait loci**. Raleigh: North Carolina State University, 2001.

ZHAO, Y.; ZENG, J.; FERNANDO, R.; REIF, J.C. Genomic prediction of hybrid wheat performance. **Crop Science**, Madison, v. 53, n. 3, p. 802, Mar. 2013.

ZHONG, S.; DEKKERS, J.C.M.; FERNANDO, R.L.; JANNINK, J.-L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. **Genetics**, London, v. 182, n. 1, p. 355–364, Mar. 2009.

ANEXOS

Anexo A – Tabelas de enriquecimento funcional , descrição dos termos e p-valores para os nove caracteres avaliados

Tabela 1 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter altura

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006914	autophagy	-17.3798
GO:0006099	tricarboxylic acid cycle	-13.6689
GO:0006545	glycine biosynthetic process	-7.6176
GO:0009165	nucleotide biosynthetic process	-7.6176
GO:0006231	dTMP biosynthetic process	-7.6176
GO:0019748	secondary metabolic process	-7.3768
GO:0002098	tRNA wobble uridine modification	-7.2991
GO:0007264	small GTPase mediated signal transduction	-6.2913
GO:0006355	regulation of transcription, DNA-templated	-5.0966
GO:0044262	cellular carbohydrate metabolic process	-4.6246
GO:0006457	protein folding	-4.5249
GO:0006275	regulation of DNA replication	-4.1681
GO:0055114	oxidation-reduction process	-3.9880
GO:0006814	sodium ion transport	-3.8151
GO:0006629	lipid metabolic process	-3.8067
GO:0009432	SOS response	-3.8067
GO:0006807	nitrogen compound metabolic process	-3.7034
GO:0016998	cell wall macromolecule catabolic process	-3.6930
GO:0006450	regulation of translational fidelity	-3.6846
GO:0006511	ubiquitin-dependent protein catabolic process	-3.5538
GO:0006855	drug transmembrane transport	-3.4770
	proteasome-mediated ubiquitin-dependent protein	
GO:0043161	catabolic process	-3.4712
GO:0005992	trehalose biosynthetic process	-3.4712
GO:0006468	protein phosphorylation	-3.3795

Tabela 1 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter altura

(conclusão)		
Termo GO	Descrição	Log ₁₀ p-valor
GO:0006633	fatty acid biosynthetic process	-3.1069
GO:0005975	carbohydrate metabolic process	-3.0324
GO:0008104	protein localization	-2.9421
GO:0006821	chloride transport	-2.9346
GO:0006438	valyl-tRNA aminoacylation	-2.7075
GO:0016310	phosphorylation	-2.5784
GO:0071805	potassium ion transmembrane transport	-2.4858
GO:0006857	oligopeptide transport	-2.4146
GO:0015991	ATP hydrolysis coupled proton transport	-2.2582
GO:0032312	regulation of ARF GTPase activity	-2.2196
GO:0006541	glutamine metabolic process	-2.1909
GO:0042254	ribosome biogenesis	-2.1838
GO:0006505	GPI anchor metabolic process	-2.1838
GO:0006265	DNA topological change	-2.1749
GO:0008380	RNA splicing	-2.0268
GO:0006887	exocytosis	-2.0088

Tabela 2 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter celulose

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006164	purine nucleotide biosynthetic process	-11.1704
GO:0009086	methionine biosynthetic process	-11.1704
GO:0008643	carbohydrate transport	-11.0350
GO:0006479	protein methylation	-8.9496
GO:0009165	nucleotide biosynthetic process	-8.7900
GO:0006545	glycine biosynthetic process	-8.7900
GO:0006231	dTMP biosynthetic process	-8.7900
GO:0006505	GPI anchor metabolic process	-7.4857
GO:0019748	secondary metabolic process	-7.4081
GO:0002098	tRNA wobble uridine modification	-6.9267
GO:0009245	lipid A biosynthetic process	-6.0224
GO:0017038	protein import	-6.0172
GO:0006450	regulation of translational fidelity	-5.0014
GO:0000042	protein targeting to Golgi	-4.7601
GO:0008104	protein localization	-4.6377
GO:0016458	gene silencing	-4.5736
GO:0008652	cellular amino acid biosynthetic process	-4.2333
GO:0045132	meiotic chromosome segregation	-3.7856
GO:0006568	tryptophan metabolic process	-3.5243
GO:0006629	lipid metabolic process	-3.3786
GO:0006511	ubiquitin-dependent protein catabolic process	-3.2360
GO:0035556	intracellular signal transduction	-3.0139
GO:0006457	protein folding	-2.7568
GO:0000272	polysaccharide catabolic process	-2.6761
GO:0007018	microtubule-based movement	-2.6606
GO:0043631	RNA polyadenylation	-2.6289
GO:0000160	phosphorelay signal transduction system	-2.6235
GO:0000105	histidine biosynthetic process	-2.5728

Tabela 2 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter celulose

		(conclusão)
Termo GO	Descrição	Log ₁₀ p-valor
GO:0015746	citrate transport	-2.5558
GO:0006367	transcription initiation from RNA polymerase II promoter	-2.5287
GO:0050790	regulation of catalytic activity	-2.4318
GO:0006265	DNA topological change	-2.4219
GO:0007275	multicellular organismal development	-2.3305
GO:0006281	DNA repair	-2.2150
GO:0016567	protein ubiquitination	-2.1925
GO:0006665	sphingolipid metabolic process	-2.1244
GO:0009306	protein secretion	-2.1244
GO:0006184	GTP catabolic process	-2.1244
GO:0006468	protein phosphorylation	-2.1175
GO:0016070	RNA metabolic process	-2.0647
GO:0033615	mitochondrial proton-transporting ATP synthase complex assembly	-2.0505
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	-2.0298
GO:0009439	cyanate metabolic process	-2.0162
GO:0007205	protein kinase C-activating G-protein coupled receptor signaling pathway	-2.0030

Tabela 3 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter FDA

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0009086	methionine biosynthetic process	-10.5897
GO:0006164	purine nucleotide biosynthetic process	-10.1723
GO:0006505	GPI anchor metabolic process	-8.2279
GO:0008643	carbohydrate transport	-8.0221
GO:0009165	nucleotide biosynthetic process	-8.0221
GO:0006545	glycine biosynthetic process	-8.0221
GO:0006231	dTMP biosynthetic process	-8.0221
GO:0000042	protein targeting to Golgi	-7.5591
GO:0019748	secondary metabolic process	-7.1147
GO:0002098	tRNA wobble uridine modification	-6.9049
GO:0006479	protein methylation	-6.8903
GO:0008104	protein localization	-6.4304
GO:0045132	meiotic chromosome segregation	-5.9434
GO:0009245	lipid A biosynthetic process	-5.8032
GO:0006568	tryptophan metabolic process	-4.0300
GO:0006457	protein folding	-3.9837
GO:0006450	regulation of translational fidelity	-3.9837
GO:0008652	cellular amino acid biosynthetic process	-3.6722
GO:0016458	gene silencing	-3.6189
GO:0016070	RNA metabolic process	-3.5984
GO:0007275	multicellular organismal development	-3.5745
GO:0044237	cellular metabolic process	-3.5225
GO:0005975	carbohydrate metabolic process	-3.3157
GO:0006665	sphingolipid metabolic process	-3.2722
GO:0007018	microtubule-based movement	-2.9685
GO:0006694	steroid biosynthetic process	-2.9602
GO:0006914	autophagy	-2.8175
GO:0000105	histidine biosynthetic process	-2.7986

Tabela 3 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter FDA

(conclusão)		
Termo GO	Descrição	Log ₁₀ p-valor
GO:0006857	oligopeptide transport	-2.7932
GO:0015746	citrate transport	-2.2934
GO:0016998	cell wall macromolecule catabolic process	-2.2818
GO:0006629	lipid metabolic process	-2.2532
GO:0006281	DNA repair	-2.2532
GO:0016567	protein ubiquitination	-2.2525
GO:0006511	ubiquitin-dependent protein catabolic process	-2.1561
GO:0000272	polysaccharide catabolic process	-2.0270

Tabela 4 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter FDN

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0009086	methionine biosynthetic process	-10.0352
GO:0009165	nucleotide biosynthetic process	-8.6164
GO:0006545	glycine biosynthetic process	-8.6164
GO:0006231	dTMP biosynthetic process	-8.6164
GO:0006505	GPI anchor metabolic process	-8.0797
GO:0019748	secondary metabolic process	-7.4554
GO:0035556	intracellular signal transduction	-7.3590
GO:0009245	lipid A biosynthetic process	-7.0484
GO:0000042	protein targeting to Golgi	-6.3269
GO:0002098	tRNA wobble uridine modification	-6.3225
GO:0006479	protein methylation	-5.9321
GO:0008652	cellular amino acid biosynthetic process	-5.4471
GO:0044237	cellular metabolic process	-4.9953
GO:0015746	citrate transport	-4.8038
GO:0006164	purine nucleotide biosynthetic process	-4.6627
GO:0006450	regulation of translational fidelity	-4.6390

Tabela 4 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter FDN (continuação)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006568	tryptophan metabolic process	-4.6276
GO:0008643	carbohydrate transport	-4.4306
GO:0050790	regulation of catalytic activity	-3.9836
GO:0006694	steroid biosynthetic process	-3.8794
GO:0006457	protein folding	-3.7901
GO:0016310	phosphorylation	-3.7376
GO:0009306	protein secretion	-3.5812
GO:0000105	histidine biosynthetic process	-3.4544
GO:0009439	cyanate metabolic process	-3.4065
GO:0007205	protein kinase C-activating G-protein coupled receptor signaling pathway	-3.3920
GO:0007275	multicellular organismal development	-3.3360
GO:0016458	gene silencing	-3.2865
GO:0006414	translational elongation	-2.9439
GO:0005975	carbohydrate metabolic process	-2.7998
GO:0033615	mitochondrial proton-transporting ATP synthase complex assembly	-2.6477
GO:0006665	sphingolipid metabolic process	-2.5872
GO:0009435	NAD biosynthetic process	-2.5854
GO:0008104	protein localization	-2.4409

Tabela 4 - Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter FDN

(conclusão)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0051258	protein polymerization	-2.4409
GO:0006562	proline catabolic process	-2.3945
GO:0055085	transmembrane transport	-2.3925
GO:0006281	DNA repair	-2.3334
GO:0007186	G-protein coupled receptor signaling pathway	-2.3065
GO:0006184	GTP catabolic process	-2.2164
GO:0006511	ubiquitin-dependent protein catabolic process	-2.2051
GO:0043631	RNA polyadenylation	-2.1910

Tabela 5 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter florescimento

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006914	autophagy	-14.8386
GO:0006099	tricarboxylic acid cycle	-9.3129
GO:0019748	secondary metabolic process	-6.7119
GO:0009086	methionine biosynthetic process	-5.8493
GO:0000042	protein targeting to Golgi	-5.8342
GO:0016311	dephosphorylation	-5.4017
GO:0006629	lipid metabolic process	-4.9991
GO:0006164	purine nucleotide biosynthetic process	-4.4650
GO:0016458	gene silencing	-3.8238
GO:0006801	superoxide metabolic process	-3.8123
GO:0006415	translational termination	-3.5518
GO:0035556	118elulose118lar signal transduction	-3.0070
GO:0006465	signal peptide processing	-2.8821
GO:0046168	glycerol-3-phosphate catabolic process	-2.8006
GO:0006529	asparagine biosynthetic process	-2.7748
GO:0016567	protein ubiquitination	-2.5177
GO:0006370	7-methylguanosine mRNA capping	-2.4556
GO:0030244	118elulose biosynthetic process	-2.4072
GO:0006097	glyoxylate cycle	-2.4072

Tabela 5 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter florescimento

(conclusão)		
Termo GO	Descrição	Log ₁₀ p-valor
GO:0006887	exocytosis	-2.4072
GO:0048193	Golgi vesicle transport	-2.4072
GO:0051258	protein polymerization	-2.4070
GO:0006184	GTP catabolic process	-2.3405
GO:0007018	microtubule-based movement	-2.2034
GO:0006476	protein deacetylation	-2.0599
GO:0006098	pentose-phosphate shunt	-2.0532
GO:0007165	signal transduction	-2.0439

Tabela 6 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o carácter hemicelulose

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
	GPI anchor metabolic process	-12.9537
GO:0006414	translational elongation	-8.1554
GO:0043043	peptide biosynthetic process	-7.9833
GO:0019748	secondary metabolic process	-7.3401
GO:0016310	phosphorylation	-7.1428
GO:0006520	cellular amino acid metabolic process	-7.1428
GO:0006545	glycine biosynthetic process	-6.3322
GO:0006231	dTMP biosynthetic process	-6.3322
GO:0009245	lipid A biosynthetic process	-5.4936
GO:0008652	cellular amino acid biosynthetic process	-5.3943
GO:0008643	carbohydrate transport	-5.3376
GO:0006418	tRNA aminoacylation for protein translation	-5.3115
GO:0045454	cell redox homeostasis	-4.9982
GO:0006807	nitrogen compound metabolic process	-4.9670
GO:0050790	regulation of catalytic activity	-4.6875
GO:0006075	(1->3)-beta-D-glucan biosynthetic process	-3.9761
GO:0016311	dephosphorylation	-3.9761
GO:0006013	mannose metabolic process	-3.8465
GO:0015746	citrate transport	-3.7458
GO:0007264	small GTPase mediated signal transduction	-3.6379
GO:0009308	amine metabolic process	-3.4316

Tabela 6 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter hemicelulose (conclusão)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0033615	mitochondrial proton-transporting ATP synthase complex assembly	-3.3498
GO:0009086	methionine biosynthetic process	-3.3288
GO:0009439	cyanate metabolic process	-3.1665
GO:0015991	ATP hydrolysis coupled proton transport	-2.8458
GO:0015992	proton transport	-2.8341
GO:0035556	intracellular signal transduction	-2.8341
GO:0006631	fatty acid metabolic process	-2.7020
GO:0006468	protein phosphorylation	-2.6374
GO:0046034	ATP metabolic process	-2.5528
GO:0006909	phagocytosis	-2.5160
GO:0016070	RNA metabolic process	-2.3104
GO:0005975	carbohydrate metabolic process	-2.2954
GO:0006006	glucose metabolic process	-2.1623
GO:0042545	cell wall modification	-2.1347
GO:0006450	regulation of translational fidelity	-2.1207
GO:0007186	G-protein coupled receptor signaling pathway	-2.1096
GO:0007275	multicellular organismal development	-2.0820
GO:0042254	ribosome biogenesis	-2.0769
GO:0002098	tRNA wobble uridine modification	-2.0769
GO:0015996	chlorophyll catabolic process	-2.0338
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	-2.0338
GO:0006659	phosphatidylserine biosynthetic process	-2.0338
GO:0009058	biosynthetic process	-2.0233
GO:0034968	histone lysine methylation	-2.0092

Tabela 7 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o carácter lignina

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0000042	protein targeting to Golgi	-8.1492
GO:0006479	protein methylation	-6.0499
GO:0006914	autophagy	-6.0069
GO:0006505	GPI anchor metabolic process	-4.6561
GO:0006508	proteolysis	-3.9769
GO:0007275	multicellular organismal development	-3.7581
GO:0002098	tRNA wobble uridine modification	-3.7435
GO:0006665	sphingolipid metabolic process	-3.5402
GO:0009058	biosynthetic process	-3.4422
GO:0006099	tricarboxylic acid cycle	-3.4422
GO:0016998	cell wall macromolecule catabolic process	-3.2022
GO:0006887	exocytosis	-3.1507
GO:0009245	lipid A biosynthetic process	-3.1417
GO:0009165	nucleotide biosynthetic process	-3.1417
GO:0006545	glycine biosynthetic process	-3.1417
GO:0006231	dTMP biosynthetic process	-3.1417
GO:0008652	cellular amino acid biosynthetic process	-3.1039
GO:0006783	heme biosynthetic process	-2.8325

Tabela 7 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter lignina

(conclusão)		
Termo GO	Descrição	Log ₁₀ p-valor
GO:0006164	purine nucleotide biosynthetic process	-2.7692
GO:0009086	methionine biosynthetic process	-2.7208
GO:0015746	citrate transport	-2.6951
GO:0019748	secondary metabolic process	-2.6551
GO:0006694	steroid biosynthetic process	-2.6551
GO:0044262	cellular carbohydrate metabolic process	-2.4117
GO:0016070	RNA metabolic process	-2.3700
GO:0009073	aromatic amino acid family biosynthetic process	-2.3091
GO:0006418	tRNA aminoacylation for protein translation	-2.1371
GO:0019538	protein metabolic process	-2.0356

Tabela 8 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMS

(continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0008643	carbohydrate transport	-12.4026
GO:0006545	glycine biosynthetic process	-7.8023
GO:0009165	nucleotide biosynthetic process	-7.8023
GO:0006231	dTMP biosynthetic process	-7.8023
GO:0019748	secondary metabolic process	-6.4185
GO:0009245	lipid A biosynthetic process	-6.1747
GO:0006468	protein phosphorylation	-5.3019
GO:0008104	protein localization	-5.2492
GO:0007264	small GTPase mediated signal transduction	-5.1619
GO:0016310	phosphorylation	-5.0115
GO:0006857	oligopeptide transport	-4.6992
GO:0032012	regulation of ARF protein signal transduction	-4.5143
GO:0045454	cell redox homeostasis	-4.4070
GO:0006633	fatty acid biosynthetic process	-4.1598
GO:0006511	ubiquitin-dependent protein catabolic process	-4.1192
GO:0006450	regulation of translational fidelity	-3.8583
GO:0044267	cellular protein metabolic process	-3.7983

Tabela 8 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMS (continuação)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006506	GPI anchor biosynthetic process	-3.6976
GO:0006520	cellular amino acid metabolic process	-3.6976
GO:0007275	multicellular organismal development	-3.6697
GO:0006505	GPI anchor metabolic process	-3.6618
GO:0006807	nitrogen compound metabolic process	-3.5766
GO:0001522	pseudouridine synthesis	-3.5766
GO:0008652	cellular amino acid biosynthetic process	-3.5766
GO:0006814	sodium ion transport	-3.4092
GO:0033615	mitochondrial proton-transporting ATP synthase complex assembly	-3.2096
GO:0015684	ferrous iron transport	-3.0653
GO:0006383	transcription from RNA polymerase III promoter	-3.0170
GO:0009058	biosynthetic process	-2.8292
GO:0006665	sphingolipid metabolic process	-2.8292
GO:0006568	tryptophan metabolic process	-2.7844
GO:0015992	proton transport	-2.7692
GO:0006821	chloride transport	-2.7537
GO:0006013	mannose metabolic process	-2.7319
GO:0015991	ATP hydrolysis coupled proton transport	-2.7227

Tabela 8 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMS

(conclusão)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0044237	cellular metabolic process	-2.7119
GO:0005985	sucrose metabolic process	-2.6896
GO:0016567	protein ubiquitination	-2.6045
GO:0009451	RNA modification	-2.5806
GO:0008380	RNA splicing	-2.5022
GO:0006275	regulation of DNA replication	-2.4239
GO:0006418	tRNA aminoacylation for protein translation	-2.1827
GO:0055114	oxidation-reduction process	-2.1615
GO:0051276	chromosome organization	-2.1615
GO:0006909	phagocytosis	-2.1594
GO:0006438	valyl-tRNA aminoacylation	-2.1465
GO:0009086	methionine biosynthetic process	-2.1465

Tabela 9 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMV (continua)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0008643	carbohydrate transport	-11.0588
GO:0006505	GPI anchor metabolic process	-9.0777
GO:0006545	glycine biosynthetic process	-8.6380
GO:0009165	nucleotide biosynthetic process	-8.6380
GO:0006231	dTMP biosynthetic process	-8.6380
GO:0008104	protein localization	-7.3591
GO:0019748	secondary metabolic process	-7.1718
GO:0016310	phosphorylation	-6.1164
GO:0009245	lipid A biosynthetic process	-6.0243
GO:0006511	ubiquitin-dependent protein catabolic process	-5.9185
GO:0045454	cell redox homeostasis	-5.7200
GO:0032012	regulation of ARF protein signal transduction	-5.1522
GO:0006857	oligopeptide transport	-5.1124
GO:0006520	cellular amino acid metabolic process	-4.3942
GO:0006450	regulation of translational fidelity	-4.3942
GO:0016567	protein ubiquitination	-4.3500
GO:0009086	methionine biosynthetic process	-3.8877

Tabela 9 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMV (continuação)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0006694	steroid biosynthetic process	-3.8020
GO:0006506	GPI anchor biosynthetic process	-3.7841
GO:0006468	protein phosphorylation	-3.7841
GO:0007275	multicellular organismal development	-3.3870
GO:0006810	transport	-3.3722
GO:0044267	cellular protein metabolic process	-3.3057
GO:0006275	regulation of DNA replication	-3.0370
GO:0006383	transcription from RNA polymerase III promoter	-3.0319
GO:0006821	chloride transport	-3.0302
GO:0007264	small GTPase mediated signal transduction	-3.0302
GO:0009116	nucleoside metabolic process	-3.0302
GO:0009058	biosynthetic process	-3.0277
GO:0006568	tryptophan metabolic process	-3.0056
GO:0044237	cellular metabolic process	-2.9261
GO:0035556	intracellular signal transduction	-2.7989
GO:0006013	mannose metabolic process	-2.7500
GO:0009451	RNA modification	-2.7164
GO:0008380	RNA splicing	-2.4038
GO:0006633	fatty acid biosynthetic process	-2.3809

Tabela 9 – Termos *gene ontology* (GO) enriquecidos para efeitos de maior magnitude, descrição e *p*-valores para o caráter PMV (conclusão)

Termo GO	Descrição	Log ₁₀ p-valor
GO:0001522	pseudouridine synthesis	-2.3708
GO:0005992	trehalose biosynthetic process	-2.3699
GO:0015991	ATP hydrolysis coupled proton transport	-2.3353
GO:0005985	sucrose metabolic process	-2.3140
GO:0007205	protein kinase C-activating G-protein coupled receptor signaling pathway	-2.2639
GO:0006909	phagocytosis	-2.2493
GO:0015992	proton transport	-2.1403
GO:0055114	oxidation-reduction process	-2.1391
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	-2.0015