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

**Efeito da densidade de marcadores e do tipo de matriz de parentesco
genômico na acurácia da seleção genômica em milho tropical**

Anna Rita Marcondes dos Santos

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

**Piracicaba
2016**

Anna Rita Marcondes dos Santos
Engenheira Agrônoma

**Efeito da densidade de marcadores e do tipo de matriz de parentesco genômico na
acurácia da seleção genômica em milho tropical**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **ROBERTO FRITSCHÉ NETO**

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2016

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

Santos, Anna Rita Marcondes dos

Efeito da densidade de marcadores e do tipo de matriz de parentesco genômico na acurácia da seleção genômica em milho tropical / Anna Rita Marcondes dos Santos. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016. 64 p. : il.

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

1. Desequilíbrio de ligação 2. SNP 3. Coeficiente de parentesco 4. Número demarcadores 5. Acurácia I. Título

CDD 633.15
S237e

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

AGRADECIMENTOS

Agradeço a Deus, pelo dom da vida.

A minha família e amigos pelo suporte e companheirismo.

Ao Professor Roberto Fritsche Neto pela oportunidade.

À Universidade de São Paulo, em especial ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas da Escola Superior de Agricultura “Luiz de Queiroz”, pela oportunidade de realização do doutorado.

Aos funcionários da USP/ESALQ pela colaboração na condução dos experimentos.

Aos colegas de trabalho do Laboratório de Melhoramento de Plantas Alógamas da USP/ESALQ pela agradável convivência e colaborações.

À Capes e ao CNPq pela concessão da bolsa de estudos.

À FAPESP (processo 2013/24135-2) pelo financiamento da pesquisa, à Pioneer Hi-Bred International Int. (Johnston, IA - USA) pela genotipagem das linhagens e à Helix Sementes® pelos dados dos híbridos.

SUMÁRIO

RESUMO	7
ABSTRACT	9
1 INTRODUÇÃO.....	11
2 REVISÃO BIBLIOGRÁFICA	15
2.1 O método REML/BLUP no melhoramento vegetal	15
2.2 Dados genômicos e identificação do parentesco	16
2.3 Matrizes de Parentesco genômico.....	18
2.3.1 Matriz de VanRaden (2008) (G_{vr}).....	18
2.3.2 Matriz de Astle e Balding (2009) (G_{ab})	20
2.3.3 Matriz de Yang (2010) baseada no parentesco unificado (G_{yu}).....	21
2.3.4 Matriz de Yang (2010) baseada no parentesco unificado ajustado (G_{ya})	22
2.4 Uso de Matrizes de Parentesco Genômico na Seleção Genômica.....	23
2.5 Redução na densidade de marcadores	25
3 MATERIAL E MÉTODOS.....	27
3.1 Painéis.....	Erro! Indicador não definido.
3.1.1 Linhagens.....	27
3.1.2 Híbridos	28
3.2 Genotipagem e controle de qualidade dos dados genômicos	29
3.2.1 Redução na densidade de marcadores	29
3.3 Análises genético-estatísticas	30
3.3.1 BLUP _I e G-BLUP	30
3.3.2 Matriz de VanRaden (2008) (G_{vr}).....	31
3.3.3 Matriz de Astle e Balding (2009) (G_{ab})	32
3.3.4 Matriz de parentesco unificado - Yang (2010) (G_{yu}).....	32
3.3.5 Matriz de parentesco unificado ajustado - Yang (2010) (G_{ya}).....	33
3.4 Validação cruzada e acurácia nos diferentes cenários de GS	33
4 RESULTADOS E DISCUSSÃO	37
4.1 Comparação dos diferentes cenários de Seleção Genômica (GS)	37
4.1.1 Densidade de marcadores	38
4.1.2 Tipo de matriz de parentesco	39
4.1.2.1 Parâmetros genéticos e herdabilidade (h^2)	40
4.1.2.2 <i>Heatmap</i>	47

4.1.2.3 Coincidência de seleção	50
5 CONCLUSÕES	55
REFERÊNCIAS	57

RESUMO

Efeito da densidade de marcadores e do tipo de matriz de parentesco genômico na acurácia da seleção genômica em milho tropical

A seleção genômica (*Genomic Selection* – GS) permite identificar indivíduos superiores com base no seu método genômico e apresenta a possibilidade de incorporação do melhoramento quantitativo à genética molecular. O melhor preditor linear não viesado (*Best Linear Unbiased Prediction* – BLUP) é um método de predição dos efeitos aleatórios do modelo com base nos componentes de variância obtidos por meio do método máxima verossimilhança restrita (*Restricted maximum likelihood* – REML). A eficiência do REML\BLUP pode ser incrementada por meio da incorporação de matrizes de parentesco genômico nos modelos de seleção genômica, uma vez que os efeitos genéticos genômicos aditivos dos indivíduos avaliados constituem os componentes aleatórios dos modelos mistos abordados. Dentre os algoritmos disponíveis para estimá-las, a partir de dados de polimorfismo de nucleotídeo único (*Single-nucleotide polymorphism* – SNP's) há as matrizes de Vanraden (2008) Astle e Balding (2009) e as de Yang *et al.* (2010) (matriz de parentesco unificado e matriz de parentesco unificado ajustado). Adicionalmente ao método de estimação da matriz de parentesco, a acurácia da GS depende da densidade de marcadores e da extensão e padrão de desequilíbrio de ligação (*Linkage disequilibrium* – LD) que existe no painel. Com isto, os objetivos foram: *i*) investigar o efeito da redução nas densidades de marcadores SNP's, por meio do LD, na estimativa do parentesco genético dos indivíduos e, consequentemente, na acurácia preditiva da GS; *ii*) estudar o efeito do uso de diferentes tipos de matrizes de parentesco genômico na acurácia da seleção genômica para linhagens e híbridos de milho tropical. Para isso, foram considerados dois painéis distintos de milho tropical: uma composta por 64 linhagens endogâmicas e a outra, por 452 híbridos. Estas foram avaliadas para o caráter produtividade de grãos, em oito e cinco ambientes, respectivamente. As linhagens e os híbridos foram genotipados com a plataforma Affymetrix® Axiom® Maize Genotyping Array, com cerca de 600 mil marcadores. Foram construídos diferentes cenários de GS quanto ao tipo de painel (linhagens e híbridos), densidade de marcadores (400k, 60k, 9586 e 1304 para linhagens e 50k, 4458 e 495, para híbridos) e tipos de matrizes de parentesco (citadas acima). Em cada um destes cenários foi estimada a herdabilidade, acurácia, capacidade preditiva e a coincidência de seleção. A partir dos resultados conclui-se que: tanto para híbridos como para linhagens, a densidade de marcas pode ser reduzida significativamente por meio do LD existente entre os SNP's, sem gerar prejuízos quanto à acurácia da GS, mas reduzindo os custos com genotipagem e a demanda computacional. Para predição genômica das linhagens, a matriz de Vanraden (2008) apresenta o melhor custo benefício, pois tem menor demanda computacional e proporciona resultados satisfatórios quanto à acurácia e coincidência de seleção. Para a predição genômica de híbridos, as matrizes de Yang *et al.* (2010) são superiores em relação às demais testadas, quanto à acurácia preditiva e coincidência de seleção.

Palavras-chave: Desequilíbrio de ligação; SNP; Coeficiente de parentesco; Número de marcadores; Acurácia

ABSTRACT

Effect of markers density and type of genomic relationship matrix in the accuracy of genomic selection in tropical maize

The efficiency of the Best Linear Unbiased Prediction (BLUP), along with the Restricted maximum likelihood (REML) methods, can be increased by incorporating genomic kinship matrices to Genomic Selection models (GS). Among the available algorithms to estimate these matrices from single-nucleotide polymorphism (SNP) data there are Vanraden (2008), Astle e Balding (2009) and Yang *et al.* (2010) (unified relationship matrix and adjusted unified relationship matrix). In addition to the estimation method of the relationship matrix, the accuracy of GS depends on the density of markers and the extent and pattern of linkage disequilibrium (LD) that there is in the panel. Thus, the aims were: *i*) assess the effect of marker density reduction, based on LD, on the estimative of the genetic relationship of individuals and, consequently, on the predictive accuracy of GS; *ii*) study the effects of types of genomic relationship matrices on the accuracy of genomic selection for tropical maize lines and hybrids. In order to achieve there, two distinct panel of tropical maize were used: the first consisting of 64 inbred lines and latter of 452 hybrids. These panels were evaluated in field trials for grain yield in eight and five environments, respectively. The lines and hybrids were genotyped by the Affymetrix® Axiom® Maize Genotyping Array platform with 612 thousand markers. Distinct scenarios of GS were comprised considering panel type (lines and hybrids), marker density (400k, 60k, 9586 and 1304 for lines and, 50k, 4458 and 495 for hybrids) and, types of kinship matrices cited aboved. In each of these scenarios heritability, accuracy, predictive capacity and the selection coincidence were estimated. From the results it was concluded that: for both, hybrids and lines, marker density can be significantly reduced based on LD between SNP's without affecting the accuracy. Furthermore, this reduction also decreases genotyping costs and computational requirements. For genomic prediction of lines, Vanraden (2008) matrix shows the best cost-benefit, because it demands less computer resources and provides satisfactory results in terms of accuracy and selection coincidence. On the other hand, for genomic prediction of hybrids, the matrices of Yang *et al.* (2010) yielded higher predictive accuracies and selection coincidences.

Keywords: Linkage disequilibrium; SNP; Kinship coefficient; Number of markers; Accuracy

1 INTRODUÇÃO

O uso de ferramentas flexíveis que permitam analisar dados desbalanceados, genética e estatisticamente, com a utilização de modelos mistos é essencial no melhoramento de plantas (FRITSCHÉ-NETO et al., 2010). Neste sentido, recomenda-se o uso do Melhor Preditor Linear Não Viesado (*Best Linear Unbiased Prediction* – BLUP), em conjunto com o Método Máxima Verossimilhança Restrita (*Restricted maximum likelihood* – REML) (HENDERSON, 1949; THOMPSON, 1975), para prever os valores genéticos dos indivíduos e a estimação de componentes de variância.

Tradicionalmente, a predição dos valores genéticos (*breeding value*) é feita a partir de *pedigree* (BERNARDO, 2010), onde é possível identificar os fundadores da população base (POWELL et al., 2010). Porém, se o *pedigree* não está completo e livre de erros, se há variação no coeficiente de endogamia (F) entre genitores e ou, se o desequilíbrio de ligação em fase gamética dentro da família não é levado em consideração, as predições de valores genéticos obtidas serão viesadas pelos efeitos da seleção (PURCELL et al., 2007). Neste caso, os componentes de variância da população e o parentesco entre os indivíduos serão pobremente estimados. Ademais, as suposições sobre a população base “ideal” raramente se verificam na prática (ENDELMAN; JANNINK, 2012).

Visando aumentar a eficiência do REML/BLUP, um procedimento bastante utilizado é a incorporação da matriz de parentesco realizado nos modelos de predição genômica. Tais matrizes têm potencial para explicar quase a totalidade da variância genética aditiva do painel, uma vez que o parentesco entre os genótipos avaliados atualmente é estimado a partir de dados de genotipagem em larga escala em que milhares de marcadores são distribuídos pelo genoma dos indivíduos (NEJATI-JAVAREMI et al., 1997).

Estimativas de parentesco genômico têm sido empregadas para melhorar a predição de indivíduos (VANRADEN, 2008), por meio da Seleção Genômica (*Genomic selection* – GS) (HAYES; GODDARD, 2001), explicar a estrutura de populações em estudos de genética de associação (*Genome Wide Association Studies* - GWAS) (STICH et al., 2008) e de estratificação populacional (AL ADHAMI et al., 2015; BERGFELDER-DRÜING et al., 2015). Neste contexto, além do BLUP fenotípico (BLUP_I), o qual representa os valores observados no melhoramento clássico para certa característica de interesse, novas matrizes de parentesco genômico têm sido desenvolvidas buscando captar intrinsecamente a identidade por estado (*identity by state* - IBS) e a identidade por descendência (*identity by descent* - IBD) entre os indivíduos.

Tem-se que a identidade por estado é uma medida de identidade entre dois indivíduos de acordo com o “estado” observado do alelo, sem considerar de quais genitores esses dois indivíduos herdaram este. Já a IBD, classicamente atrelada à proposta de Wright (WRIGHT, 1921, 1922), considera não apenas o estado atual dos alelos, mas também a probabilidade de que determinados alelos sejam herdados de ancestrais comuns. Com isto, se consegue estimar o parentesco genômico real que há entre dois indivíduos e não apenas o parentesco oriundo da esperança matemática (RESENDE et al., 2008).

Dessa forma, as matrizes de parentesco por estado têm como princípio usar a painel parental como a painel base (ENDELMAN; JANNINK, 2012). Assim, a covariância genética entre indivíduos é fundamentalmente uma propriedade de estado e não necessita de uma população base ideal composta por N indivíduos diploides, a qual é reconstituída a cada geração por uma amostra aleatória de $2N$ gametas (WRIGHT, 1950). Com isto, o parentesco surge da relação existente entre indivíduos que têm pelo menos um ancestral em comum e o coeficiente de endogamia (F) é a probabilidade de dois alelos em um determinado loco sejam idênticos por descendência (MALECOT, 1948).

Com base nestes dois conceitos, existem diferentes algoritmos para estimar matrizes de parentesco, a partir de dados de polimorfismo de nucleotídeo único (*Single-nucleotide polymorphism* – SNP's), em que a frequência alélica dos destes é estimada levando em consideração as similaridades IBS e IBD, ponderada pela frequência alélica do total de SNP's na amostra. Um destes foi proposto por Vanraden (2008), a qual está relacionada com a substituição alélica, cuja matriz de incidência molecular de indivíduos não genotipados provém da extensão das equações de modelos mistos. A matriz de Astle e Balding (2009), beneficia-se de teste de associação minimizar problemas de confundimento devido à estrutura da população e do parentesco críptico, mas indica apresentar falhas por não valorizar o desequilíbrio de ligação (*Linkage disequilibrium* – LD) entre os SNP's. Já a matriz proposta por Yang et al. (2010) (matriz de parentesco unificado) foi proposta para o estudo do caráter peso de humanos, em uso para associação de marcas significativamente associadas com o fenótipo (*Genome-wide association study* – GWAS) na ocasião. Este mesmo autor propôs um ajuste nesta matriz de parentesco unificado por meio de um coeficiente de regressão estabelecido empiricamente, objetivando explicar a herdabilidade total estimada (componentes aditivos e não aditivos) a partir da captura de marcadores de pequenos efeitos em desequilíbrio de ligação com alelos relacionados à variável causal estudada, peso em humanos, levando em consideração o MAF (*Minor allele frequency*) de ambos (YANG et al., 2010).

Sabe-se que a acurácia da GS depende da densidade de marcadores e da extensão e do padrão de desequilíbrio de ligação que existe no painel. Além disso, ela também depende da quantidade e precisão dos dados disponíveis para estimar os efeitos dos marcadores, além da eficiência da estratégia e dos métodos estatísticos utilizados na predição e da complexidade do controle genético da característica em estudo (BERNARDO, 2010).

Em relação ao efeito da densidade de marcadores sobre o estimador do parentesco genético genômico entre indivíduos sabe-se que esta é significativa, pois a variância residual de cada estimador depende do MAF do marcador e da LD entre os SNP's, o qual gera correlações entre eles. Como consequência, estimadores com baixa acurácia terão o mesmo peso que estimadores mais acurados (RINCENT, 2014). Assim, o parentesco entre dois indivíduos pode ser super ou subestimado, e modificar significativamente a acurácia da GS. Além disso, a redução na densidade de genotipagem pode reduzir os custos com genotipagem, a demanda computacional e os erros devidos a multicolinearidade. Neste contexto, indiferente do método, a matriz de parentesco mais adequada para a predição genômica é aquela que proporciona maior capacidade preditiva dos valores genéticos de indivíduos observados e não observados, mesmo sob número variável de marcadores (RESENDE, 2008).

Diante do exposto, os objetivos foram: *i)* investigar o efeito da redução nas densidades de marcadores SNP's, por meio do LD, na estimativa do parentesco genético dos indivíduos e, consequentemente, na acurácia preditiva da GS; *ii)* estudar o efeito do uso de diferentes tipos de matrizes de parentesco genômico na acurácia da seleção genômica para linhagens e híbridos de milho tropical.

2 REVISÃO BIBLIOGRÁFICA

2.1 O método REML/BLUP no melhoramento vegetal

Desde o seu desenvolvimento, o BLUP tem sido exaustivamente utilizado no melhoramento animal (VANRADEN, 2008) e só a partir das últimas duas décadas ganhou espaço no melhoramento de culturas perenes e florestais e, mais recentemente, em culturas anuais (PIEPHO et al., 2008).

O modelo misto proposto por (HENDERSON, 1949) pode ser expresso pela seguinte equação:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Em que $\mathbf{y}$ é o vetor de observações fenotípicas; $\mathbf{X}$ matriz de incidência de efeitos fixos ou de delineamento; $\mathbf{b}$ o vetor de efeitos fixos (média populacional, local, ano, por exemplo); $\mathbf{Z}$ a matriz de incidência de efeitos aleatórios; $\mathbf{u}$, o vetor de efeitos aleatórios (*breeding value*), sendo $\mathbf{u} \sim N(0, \mathbf{G})$ e $\mathbf{e}$, vetor de efeitos residuais, $\mathbf{e} \sim N(0, \mathbf{R})$. Tal modelo é solucionado pelas seguintes equações de Henderson (1949), em que a matriz $\mathbf{G}$ é a matriz de efeitos genéticos aditivos $\mathbf{A}$:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

O procedimento BLUP permite a utilização de todos os efeitos (fixos e aleatórios) do modelo estatístico, da covariância genética entre os indivíduos ou genitores em avaliação, ponderar aqueles genótipos com desigual número de observações na mesma ou em diferentes gerações e, assim, maximizar a acurácia, a qual depende basicamente da proporção entre a variação residual média e a variação genotípica, sendo que a variação residual média depende do número de repetições (RESENDE; DUARTE, 2007; CAVALCANTI; RESENDE, 2012). Dessa forma o BLUP maximiza a confiabilidade das comparações entre genótipos, ao considerar não apenas o desempenho *per se* do indivíduo, mas também sua ancestralidade, e permite lidar com o montante de dados disponíveis, sem comprometer o ordenamento de indivíduos quanto ao caráter observado, pelo fato de um ou outro genótipo ter sido avaliado mais extensivamente.

Tal procedimento assume que os componentes de variância são conhecidos, porém na prática não se tem acesso aos seus verdadeiros valores. Assim, é necessário um método que os estime de forma eficiente. Tal estimação pode ser feita pelo método da Máxima Verossimilhança Restrita (REML), em que os componentes de variância genéticos são preditos, levando em consideração a correlação que existe entre parentais e sua prole.

É importante ressaltar que a abordagem teórica dos princípios de genética quantitativa traz algumas pressuposições que, na maioria das vezes, não são totalmente satisfeitas (HOLLAND et al., 2003). Entre elas, a de que os genótipos considerados em programas de melhoramento podem reproduzir a painel base ideal e que, nesta, todos os genótipos são não correlacionados (PIEPHO et al., 2008). Ao passo que deveria ser considerada a possibilidade de parentesco entre os materiais a serem avaliados, sejam eles linhagens ou híbridos.

Neste sentido, o emprego da matriz de parentesco baseada na genealogia dos indivíduos pode aumentar a acurácia na predição dos valores genéticos e estimação dos componentes de variância, em detrimento de considerar os genitores como não relacionados, quando na verdade o são (HOLLAND et al., 2003). Para isso pode ser utilizada a matriz dos coeficientes de correlação entre essas linhagens obtida via *pedigree*. Neste caso, são avaliadas a covariância e a correlação genética existente entre o material experimental avaliado e o material que é passado aos próximos ciclos de melhoramento (CRUZ, 2012). Tal matriz, obtida pelo *pedigree* entre os indivíduos tem sido amplamente usada no melhoramento animal (MEYER, 1989; JOHNSON; THOMPSON, 1995; FALCÃO et al., 2009; MENEZES et al., 2013) e vegetal (BERNARDO, 1995; 1996; BAUER et al., 2006; NUNES et al., 2008). Estes se baseiam no fato de que quanto maior for a correlação genética de um genótipo de interesse com genótipos relacionados, mais informação pode ser adquirida a partir de registros de parentes (PIEPHO et al., 2008).

Mesmo usando a abordagem baseada em *pedigree* há possibilidades do comprometimento da qualidade e da veracidade das estimativas. Caso a coleção de germoplasma em um programa de melhoramento seja de origem desconhecida, registros de *pedigree* podem não estar indisponíveis ou quando disponíveis pode haver dificuldade em montá-lo (BERNARDO, 2010). Entretanto, na ausência deste, é possível o cálculo de identidade por descendência IBD a partir de dados de SNP's.

2.2 Dados genômicos e identificação do parentesco

Nas últimas décadas, as plataformas de genotipagem de alto rendimento tornaram-se disponíveis para várias espécies de plantas e animais, tornando o processo cada vez mais

rápido e barato. Marcadores moleculares do tipo SNP's (*Single-nucleotide polymorphism*) atualmente têm sido mais utilizados devido a sua capacidade de cobertura e confiabilidade em análises genômicas. Estes correspondem à forma de variação mais abundante no genoma, apresentam alta reprodutibilidade, possibilidade de automação, ampla cobertura do genoma e baixo custo (RESENDE, 2008). O desenvolvimento de grandes painéis de SNP's inicialmente foi feito para culturas importantes como o milho (GANAL et al., 2011) e a soja (SONG et al., 2013), e atualmente encontram-se disponíveis para as mais variadas espécies vegetais.

A combinação de dados de SNP's e fenótipos são usados para desenvolver e validar modelos de predição genômica para a ampla gama de características de interesse agrônomo. Além disto, são utilizados com sucesso em estudos de diversidade, estrutura genética de painéis e são amplamente empregados em seleção genômica (*Genomic Selection – GS*) e estudos de genética de associação (*Genome-wide association study – GWAS*). Na cultura do milho, esses marcadores vêm sendo amplamente empregados para a construção de mapas genéticos (GORE et al., 2009), verificação de parentesco e análises evolutivas (VAN HEERWAARDEN et al., 2011), análise de arquitetura genética de características complexas (COOK et al., 2012; PEIFFER et al., 2014), confirmação da identidade genética de linhagens (SEMAGN et al., 2012) e análise de associação (TIAN et al., 2011).

A utilização da matriz de parentesco baseada em SNP's tem incentivado trabalhos que buscam realizar o estudo de herança de características de importância agrônoma e prever valores genéticos, tanto de híbridos simples como de linhagens de milho (STICH, 2009; RIEDELSHEIMER et al., 2012; ZILA et al., 2014). Neste caso, a similaridade genética pode unificar dois conceitos fundamentais, identidade por estado (*identity by state – IBS*) e identidade por descendência (*identity by descent – IBD*). Efeitos de alelos IBD levam em conta sua capacidade de prever a probabilidade de os alelos em *loci* não observados, os quais podem estar ligados aos caracteres de interesse, serem IBS. Em contrapartida, a teoria da coalescência postula que dois segmentos cromossômicos homólogos são considerados idênticos por descendência somente se vierem de um ancestral comum sem que cada um deles tenha passado por recombinação ou mutação (ROSENBERG; FELDMAN, 2002). Assim, o parentesco entre dois indivíduos diploides pode ser definido em termos de que as probabilidades de cada subconjunto dos seus quatro alelos em um *locus* arbitrário é idêntico por descendência, o que significa que eles descendem de um ancestral comum, sem alelo que tenha sofrido mutação. A probabilidade de que os dois alelos homólogos dentro de um indivíduo *i* são IBD é conhecido como seu coeficiente de parentesco.

Em estudos de características quantitativas, como a produtividade de grãos, o componente poligênico é dependente do parentesco entre os indivíduos em avaliação e o componente do QTL (*Quantitative Trait Loci*) depende da proporção de alelos IBD entre pares de indivíduos que possuem o mesmo QTL. O componente poligênico está associado a muitos genes de pequenos efeitos, e é assumido que a média, incluindo vários poligenes, da proporção de alelos IBD entre dois indivíduos é equivalente ao parentesco genético médio entre dois irmãos. No caso do QTL, a proporção IBD varia entre pares de irmãos e é estimada por meio dos genótipos observados nos *loci* dos marcadores. Logicamente, a proporção de alelos IBD do QTL não é observada diretamente. Os IBDs do QTL são avaliados em cada segregação em nível individual e não em nível médio, permitindo conhecer exatamente cada segregação.

A perspectiva de que os métodos para estimar *loci* IBD devem, de fato, objetivar estimar a possibilidade de que os gametas sejam IBS em *loci* não observados elimina o conflito entre métodos IBD baseados em marcadores moleculares e métodos coalescentes (HARTL et al., 1997; NORDBORG, 2001). O reconhecimento de que é preciso estimar a probabilidade de que indivíduos possuam genes que são IBS em *loci* não observados leva a um método novo e imparcial para estimar variância genética e herdabilidade (MANOLIO et al., 2009)

Análise de ligação, métodos associação por marcas simples e métodos por haplótipos costumam utilizar o conceito de IBS em posições gênicas não observadas, mas frequentemente não estimam sua probabilidade formalmente. Quando isto é realizado, a verossimilhança dos dados fenotípicos pode ser maior (MEUWISSEN; GODDARD, 2001). Portanto, baseando-se nestes conceitos, é possível construir estruturas de matrizes de variância-covariância de parentesco que estejam de acordo com as pressuposições da genética quantitativa e do modelo estatístico-genético adotado (YANG et al., 2010).

2.3 Matrizes de Parentesco genômico

2.3.1 Matriz de Vanraden (2008) (G_{vr})

A análise e a seleção em um programa de melhoramento genético podem envolver simultaneamente indivíduos fenotipados e genotipados. Pode-se considerar o modelo linear misto geral ajustado para estimar os efeitos genéticos aditivos usando informações fenotípicas:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e},$$

em que $\mathbf{y}$ é o vetor de informações fenotípicas, $\mathbf{b}$ é o vetor de efeitos fixos individuais, $\mathbf{g}$ é o vetor dos efeitos genéticos aditivos e $\mathbf{e}$ o vetor de resíduo; $\mathbf{X}$ e $\mathbf{Z}$, as matrizes de incidência para os dois últimos vetores de efeitos fixo e aleatório, respectivamente.

Neste caso, este modelo pode ser substituído por $\mathbf{y} = \mathbf{Xb} + \mathbf{ZWm} + \mathbf{e}$, em que $\mathbf{m}$ é o vetor de efeitos dos marcadores e $\mathbf{W}$ é a matriz composta por valores 0, 1 e 2 para os genótipos MM, Mm, e mm, respectivamente, para o marcador bialélico e codominante i , em um indivíduo diploide. As equações de modelo misto para predição de g via método G-BLUP, baseiam-se na padronização e centralização da média em zero da matriz $\mathbf{W}$. Essa se revela a essência de estimação da matriz de parentesco genômico, baseada em efeitos aditivos, proposta por Yang et al. (2010), utilizada por Resende et al. (2012b) a qual substitui a matriz de parentesco baseada em *pedigree* e é função da matriz $\mathbf{WW'}$ presente nas próprias equações de modelos mistos. Após a padronização, a matriz de parentesco genômico $\mathbf{G}$ conterá a média ponderada das relações de parentesco estimadas de cada *locus* marcador, em que os pesos da ponderação são função da inversa da variância do erro de predição (*prediction error* - PEV) associada à variável indicadora $\mathbf{W}$ em cada marcador.

$$\mathbf{G}_{vr} = \frac{(\mathbf{W}^* \mathbf{W}^{*'})}{\sum_i^n [2p_i(1-p_i)]}$$

em que p_i é a frequência de um dos alelos do *locus* i e $\mathbf{W}^*$ é a matriz $\mathbf{W}$, a matriz de incidência de marcadores dos indivíduos, corrigida para suas médias em cada *locus* ($2p_i$).

A matriz $\mathbf{Gvr}$, proposta por Vanraden (2008), tem sido extensivamente utilizada (RESENDE et al., 2012a, 2012c; MUNOZ et al., 2014) por ser uma extensão das equações de modelos mistos que divide o valor fenótipo em genético (efeitos aleatórios) e ambientais. Ao mesmo tempo, quando indivíduos estão relacionados, este modelo necessita desta matriz de parentesco acurada para corrigir este efeito de covariância. Esta abordagem corrige adequadamente para qualquer efeito ambiental e gera estimativas de *breeding values*, pelo G-BLUP. Depois disto é feita uma regressão com base em um valor de parentesco esperado assumindo um modelo infinitesimal ($\mathbf{Gvr}$). Como consequência da utilização da $\mathbf{Gvr}$, maior precisão e menos viés na GS são esperados.

2.3.2 Matriz de Astle e Balding (2009) ($\mathbf{G}_{ab}$)

Sob muitos modelos de genética de painéis, se dois alelos não são IBD eles são considerados aleatórios caso sofrerem mutação ou pertencerem a um *pool* alélico (ROUSSET, 2002) e, com isto, considerados "não relacionados". Assim, o parentesco g_{ij} proposto por Astle e Balding (2009) é um coeficiente que indica se há correlação entre dados alelos provenientes dos indivíduos i e j . Se x_i e x_j forem convertidos em números, 0, 1 ou 2, então:

$$\text{Cov}(x_i, x_j) = 4p(i - p)g_{ij}$$

em que p é a fração de alelos de uma panel. Assim, g_{ij} pode ser estimada a partir de covariâncias, a partir da contagem de alelos do genoma. Especificamente, se escrevermos x como um vetor coluna dos indivíduos e deixarmos o indicador do *loci* L subscrito (ao invés de indivíduos), então:

$$\mathbf{G}_{ab} = \frac{1}{L} \sum_{l=1}^L \frac{(x_l - 2p_l)(x_l - 2p_l)^T}{4p_l(1 - p_l)}$$

é um estimador semi-definido imparcial e positivo para a matriz de parentesco $\mathbf{G}_{ab}$. $\hat{\mathbf{G}}_{ab}$ pode superestimar o compartilhamento de alelos esperado para indivíduos não aparentados. De acordo com Ritland (1996), estimadores considerados similares e generalizados em *locus* com mais de dois alelos foram propostos por Li e Horvitz (1953), mas apenas para coeficientes de endogamia. Na prática, não sabemos as frações de alelos p_l não oriundos de famílias endogâmicas e não relacionados. Este desvio pode subestimar $(\mathbf{G}_{ab})_{ij}$ e superestimar p_l a partir dos mesmos dados. Embora as correlações decorrentes da ancestralidade comum sejam, em princípio, positivas, devido à estimativa viesada de p_l , os elementos fora da diagonal de $\hat{\mathbf{G}}_{ab}$ podem ser negativos, propriedade que tem levado alguns autores a evitar tais estimadores de $\mathbf{G}_{ab}$.

Um problema que pode surgir a partir do estimador de parentesco proposto por Astle e Balding (2009) e outros estimadores clássicos como a IBD, ou o estimador de Vanraden (2008) é que o desequilíbrio de ligação entre SNP's não é levado em consideração. Como resultado regiões do genoma que apresentam várias marcas com maior LD recebem, indevidamente, peso maior na estimativa de parentesco (RINCENT, 2014).

2.3.3 Matriz de Yang (2010) baseada no parentesco unificado (G_{yu})

Outra forma de estimação da matriz de parentesco quando não se dispõe do *pedigree* da painel em estudo, ou se trata de uma painel não estruturada é o parentesco unificado aditivo bruto (*raw unified additive relationship - raw uAr*, $\hat{A}$ ou G_{yu}), proposta por Yang et al. (2010) que, assim como a anterior, pode ser incorporada nas equações do BLUP tradicional. No caso de caracteres quantitativos, como a produtividade de grãos, nos quais vários alelos contribuem para manifestação do fenótipo, atribui-se o valor 0, 1 até o n-ésimo valor a cada um dos alelos. Assim, à semelhança do parentesco entre gametas, o parentesco entre cada indivíduo com ele mesmo e com os demais pode ser estimado, constituindo o parentesco unificado bruto.

O parentesco entre indivíduos diferentes é dado por:

$$G_{yu} = A_{ij} = \frac{1}{m} \sum_k \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1 - p_k)}, \text{ para } i \neq j$$

em que x é um indicador do SNP k nos indivíduos i e j , p_k é a frequência alélica (frequência de cada alelo) do SNP k , e m é o número de SNP's na amostra.

O parentesco de um indivíduo com ele mesmo é predito por meio da seguinte equação:

$$G_{yu} = A_{ij} = 1 + \frac{1}{m} \sum_k \frac{x_{ik}^2 - (1 + 2p_k)(x_{ik} + 2p_k^2)}{2p_k(1 - p_k)}, \text{ para } i = j$$

em que x é o indicador do SNP k nos indivíduos i , p_k é a frequência alélica (frequência de cada alelo) do SNP k , e m é o número de SNP's na amostra.

Em estudos de caracteres quantitativos têm sido encontrados SNP's significativamente associados a caracteres complexos de interesse, mas que correspondem a uma pequena fração da variação genética. A variância explicada pelos SNP's é geralmente muito menor do que a proporção de variância fenotípica, levando a subestimação da herdabilidade no sentido restrito.

As possíveis causas incluem a existência de interação intra-alélica, interação alelo-ambiente, herança epigenética, ou a de que a σ_a^2 não é detectada porque os efeitos individuais são muito pequenos para testes de significância rigorosos. G_{yu} pode apresentar elevado erro amostral, caso o número de marcas utilizadas seja baixo (POWELL et al., 2010).

2.3.4 Matriz de Yang (2010) baseada no parentesco unificado ajustado (G_{ya})

É possível realizar um ajuste na G_{yu} e tornar sua estimativa menos viesada, gerando a G_{ya} , baseada no parentesco unificado ajustado como explanado no trabalho de (POWELL et al., 2010). Para isto considera-se um ajuste do coeficiente de endogamia (F), por meio da $E(F / \hat{F}) = \tilde{F}$, em que $\hat{F}$ corresponde a uma estimativa média de F , para todos os *loci* (YANG et al., 2010).

Na matriz G_{ya} , as estimativas de parentesco foram corrigidas pelo erro amostral do método anterior ($\sigma_e^2 = 1/m$) por meio da regressão de A em $\hat{A}(1/m)$. Sendo o coeficiente de regressão obtido por $\frac{\text{var}(A)}{\text{var}(A) + \frac{1}{m}}$ e $\text{var}(\hat{A}) = \text{var}(A) + \frac{1}{m}$, então a G_{yu} ajustada é dada por:

$$G_{ya} = A_{ij}^* = \left(1 - \frac{\frac{1}{m}}{\text{var}(\hat{A})} \right) \hat{A}, \text{ para } i \neq j;$$

e

$$G_{ya} = A_{ij}^{*'} = 1 + \left(1 - \frac{\frac{1}{m}}{\text{var}(\hat{A})} \right) (\hat{A} - 1), \text{ para } i = j.$$

Como as variáveis resposta são desconhecidas, não é possível estimar seu desequilíbrio de ligação (*Linkage Disequilibrium* – LD) observando os SNP's diretamente, mas pode-se mimetiza-las considerando o LD entre os SNP's genotipados de modo que o LD entre SNP's é apenas um guia para o LD entre variáveis resposta e SNP's.

As expressões acima dependem do MAF das variáveis causais (c). Se os *loci* de interesse tiverem o mesmo espectro de frequência de alelos que os SNP's genotipados ($\theta = 0,5$), em seguida, $c = 0$, e $1/N$ pode ser interpretado como o erro de amostragem para estimar o parentesco a partir dos N SNP's aleatórios ao longo de todo o genoma. Se $c > 0$ consequentemente $\theta < 0,5$, pois o parentesco em variáveis resposta com baixo MAF (*Minor allele frequency*) é tipicamente menor do que o parentesco médio ao longo de todo o genoma.

Portanto, dado o número de SNP's usado, pode-se corrigir a estimativa da variância explicada pelo SNP's para LD incompleto com variáveis resposta, caso elas tenham a mesma frequência alélica que os SNP's genotipados.

A desvantagem do estimador **Gyu** é que, mesmo que possibilite a correção da variância para o LD incompleto ($c = 0$) e estimação da proporção de variância explicada por variantes causais, esta estimativa assume que o LD entre SNP's e variantes causais é tão forte quanto aquela entre os SNP's genotipados. Isso foi utilizado como pressuposição para elaboração de **Gya**, no entanto, se os SNP's causais tendem a ser menor do que a MAF dos SNP's que foram testados, o LD entre SNP's e variantes causais deve ser reduzido e desvio padrão elevado (HAYES; GODDARD, 2001).

Um ponto favorável ao uso da **Gyu** é que resultados encontrados a partir dessa matriz são consistentes com dados fenotípicos a uma frequência mais baixa do que os SNP's usados em outras matrizes e, por conseguinte, menor LD com estes SNP's que o LD dos entre SNP's e poderiam explicar a herdabilidade total estimada, importante na análise de híbridos (POWELL et al., 2010).

2.4 Uso de Matrizes de Parentesco Genômico na Seleção Genômica

A ideia de usar grandes painéis de marcadores moleculares para predições genômicas não é recente (HALEY; VISSCHER, 1998; HAYES; GODDARD, 2001). Tal abordagem coloca em prática a velha ideia de seleção assistida por marcadores para caracteres quantitativos. No entanto, a sua aplicação efetiva só se tornou viável com o advento da seleção genômica, que têm proporcionado ganhos genéticos consideráveis por unidade de tempo, reduzindo a duração dos ciclos de melhoramento (GRATTAPAGLIA; RESENDE, 2011).

Com o BLUP empregando apenas dados fenotípicos, os estimadores encontrados podem não ser superiores aos de abordagem clássica, como Análise de Variância (ANOVA). Mas com o incremento da informação oriunda de dados genômicos dos materiais avaliados em um programa de melhoramento vegetal, há a possibilidade aprimorar os estudos de seleção e associação genômica, caso considerada a ploidia, a detecção de alelos idênticos por descendência e a incorporação da matriz de parentesco genômico (LIMA). Assim, a predição de valores genéticos possibilita selecionar indivíduos para os quais os fenótipos não estão disponíveis (SCHAEFFER, 2006) com precisão ainda mais elevada em painéis para os quais informações de *pedigree* estão disponíveis.

Ao fazer uso da matriz genômica (**G**), em vez da matriz aditiva (**A**), baseada em *pedigree*, torna a análise do BLUP tradicional diferenciada, pois a matriz de incidência que relaciona os indivíduos aos alelos de cada *locus* é estimada com base em informação de marcadores. Então, a forma como os marcadores são usados para estimar a matriz de parentesco genômico e como são definidos os efeitos alélicos para cada *locus* resultará nos diversos modelos de análise para predição genômica, entre eles os modelos G-BLUP (CLARK; VAN DER WERF, 2013).

O G-BLUP apresenta outro uso importante que é a estimação da herdabilidade total explicada por todos os marcadores simultaneamente. Dada a matriz **G**, são estimados por REML, pelas equações de modelo misto, os parâmetros genéticos e componentes de variância com maiores precisão e coincidência de seleção (POWELL et al., 2010).

É preferível utilizar estas estimativas em associação com as estimativas baseadas em *pedigree*, raramente disponíveis na sua totalidade. Isto porque há uma sutil diferença entre estimativas computadas a partir de um *pedigree* completo e o realizado com base em material genômico partilhado. Por exemplo, se duas linhagens de indivíduos distintos reúnem-se em um ancestral comum muitas gerações no passado, então este antepassado irá contribuir, ligeiramente, para o parentesco à base do *pedigree* indivíduos, mas pode ou não ter passado qualquer genético material a ambos. Da mesma forma, dois pares de irmãos em um *pedigree* não oriundo de endogamia podem ter o mesmo parentesco, pelo *pedigree*, mas empiricamente diferente (WEIR et al., 2006).

Com isto, a análise de caracteres quantitativos em um programa de melhoramento genético pode envolver simultaneamente indivíduos fenotipados e genotipados. Para isto, pode-se considerar o modelo linear misto geral ajustado para estimar os efeitos genéticos aditivos usando informações fenotípicas:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \mathbf{e},$$

em que, **y** é o vetor de informações fenotípicas, **b** é o vetor de efeitos fixos individuais, **g** é o vetor dos efeitos genéticos aditivos, sendo $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ e **e** o vetor de resíduo, em que $\mathbf{e} \sim N(0, \sigma^2)$; **X** e **Z**, as matrizes de incidência para os dois últimos vetores de efeitos fixo e aleatório, respectivamente; **G** é a matriz de parentesco genômico, a qual pode ser estimada por vários métodos.

2.5 Redução na densidade de marcadores

Sabe-se que a seleção genômica (*Genomic selection* – GS) pode resultar em redução de custos e no aumento das taxas de ganho genético (ZHONG et al., 2009). Para isso, os pontos fortes dos métodos de predição da GS são definidos sob diferentes condições de desequilíbrio de ligação e de densidade de marcadores, entre outros fatores.

Neste contexto, os níveis e estruturas de LD para determinada painel podem constituir o ponto de partida para a seleção genômica. Existe clara interação entre a extensão de LD e densidade de marcadores: em geral, quanto maior a densidade de marcadores utilizados, mais acuradas são as predições dos valores genéticos genômicos, efeito que se acentua em condição de baixa LD (ZHONG et al., 2009). O nível ideal de colinearidade ou LD pode depender da magnitude do efeito de marcadores. Teoricamente, a predição dentro das painéis bi-parentais é a situação mais favorável para a GS devido ao máximo desequilíbrio ligação entre os *loci* de características quantitativas e marcadores (BERNARDO; YU, 2007). Estas observações reforçam a importância de conhecer a extensão da LD para determinar densidades de marcadores necessárias e de investigar se um número menor de marcadores informativos pode ser utilizado e ainda proporcionar uma predição acurada.

A acurácia devido ao LD pode ser confundida com aquela devido ao parentesco genético. Uma grande diversidade, com muitos eventos de recombinação, leva a um baixo grau de desequilíbrio de ligação, que é favorável ao mapeamento de QTL's e GS. No entanto, o LD não é apenas devido à ligação genética, mas também pode ser causado pela estrutura do painel de indivíduos, parentesco, deriva e seleção (JANNINK; WALSH, 2002). O componente de LD devido à estrutura e parentesco do painel pode gerar falsos positivos e, portanto, tem de ser corretamente modelados, por meio de inferências sobre a estruturação usando dados genotípicos e estimadores de parentesco entre os indivíduos. Para estimar o número mínimo de marcadores necessários para cobrir o genoma e controlar falsos positivos, é preciso tomar em consideração para estimar o número de marcadores necessário para cobrir o genoma corrigir para parentesco e estimar a parte do LD apenas devido à ligação, estimando-se a LD intra cromossômica utilizando todos os marcadores (RINCENT, 2014).

Um declínio na LD devido a cruzamentos entre linhagens contrastantes, por exemplo, pode gerar um gradiente útil para avaliar a eficácia de métodos de predição, como o G-BLUP, cuja precisão é devida ao LD entre marcadores e QTL, numa fração substancial (HABIER et al., 2007). Em painéis com menores LD e, ou, menor variância genética (painel de linhagens),

a acurácia de predição tende a ser mais afetada em baixa densidade de SNPs do que em condições opostas (ZHANG et al., 2012).

Com isto, no processo de genotipagem, a seleção de marcadores informativos pode ser uma alternativa para diminuir o custo de novas genotipagens para aplicação da GS em etapas subsequentes. Existe a seleção com base no efeito das marcas, (*Ridge Regression best linear unbiased predictor* – RR-BLUP); a redução via teste de associação pela abordagem *genome wide association studies* (GWAS) em que os marcadores significativos são selecionados e o modelo ajustado somente com estes marcadores (SUBEDI et al., 2013) e, redução via desequilíbrio de ligação. Com essa redução, um número maior de indivíduos pode ser genotipado, mais haplótipos podem ser amostrados, o tempo de análise computacional e a multicolinearidade dos dados podem ser reduzidos.

Neste sentido, alguns estudos indicaram que moderada a alta acurácia de predição pode ser obtida em painéis mesmo com baixa densidade de marcadores (aproximadamente 150) e um painel relativamente pequeno (LIAN et al., 2014). Assim, as perspectivas futuras da seleção genômica são de que as densidades de marcadores serão baixas em relação ao ponto de desequilíbrio de ligação e que, sob essas condições, o LD entre marcador e QTL será o componente mais importante. Desse modo, grandes tamanhos de amostra irão compensar a alta variação de erro nas avaliações fenotípicas (HAMBLIN et al., 2011).

3 MATERIAL E MÉTODOS

3.1 Painéis

Foram considerados dois painéis distintos: um composto por 64 linhagens endogâmicas da ESALQ - USP (SANT'ANA, 2015) e a outro por 452 híbridos, disponibilizado pela Helix Sementes®.

As linhagens são oriundas do Programa de Melhoramento de Alógamas, haviam sido previamente avaliadas a campo em relação à EUN, e apresentaram elevado grau de diversidade genética, comparável a de painéis maiores analisados em outros estudos, indicando a possibilidade de inclusão dessas em painéis úteis para estudos de seleção genômica. Foram conduzidas em Piracicaba – SP e Anhumas – SP, nas safrinha de 2014 e 2015, sem a ocorrência de desbalanceamento, sendo a colheita dos grãos realizada manualmente, parcela por parcela.

Os 452 híbridos obtidos pelo delineamento genético do tipo dialelo parcial com o uso de 111 linhagens endogâmicas. A produtividade de grãos foi obtida com a colheita mecanizada de toda a parcela. Esta variável foi ajustada considerando a padronização da umidade na colheita para 13% e para estande, quando este não apresentou efeito genético significativo. Os experimentos foram conduzidos no ano agrícola de 2015 delineados em blocos casualizados com duas repetições, avaliados em Ipiacú-MG, Patos de Minas-MG, Sertanópolis-PR, Nova Mutum-MT, e Sorriso-MT. Houve desbalanceamento estatístico e genético, pois nem todos os híbridos foram avaliados em todos os locais e nem todas as combinações híbridas entre linhagens genitoras foram analisadas. A condução experimental e obtenção dos dados genômicos das linhagens genitoras são creditados a empresa Helix Sementes Ltda.

3.1.1 Linhagens

Para avaliação fenotípica do Paine 1, as linhagens foram distribuídas no campo de acordo com o delineamento experimental látice simples (8x8), com duas repetições. Foi utilizado o cultivo em faixas, constituído por 64 linhagens x dois níveis de N aplicado (ideal e baixo). As parcelas foram constituídas de uma fileira com 4m de comprimento, utilizando-se o espaçamento de 0,80m entre fileiras e 0,2m entre plantas, totalizando 20 plantas por fileira, após desbaste. Este esquema foi repetido por dois anos agrícolas, nas segundas safras de 2014 e 2015. Os dados de produtividade de grãos (ton.ha^{-1}) foram corrigidos para teor de umidade dos grãos (13%).

Os ensaios foram conduzidos em dois locais. O primeiro situa-se no Departamento de Genética da ESALQ (22°42'39"S; 47°38'09"W, 540m), Piracicaba-SP, Brasil. Esta área possui solo do tipo nitossolo vermelho eutrófico (terra roxa estruturada). O segundo local situa-se na Estação Experimental Anhumas (22°84'30"S; 48°02'20"W, 530 m) e possui solo do tipo latossolo amarelo distrófico. Trata-se de um local semelhante aos encontrados nos Cerrados do Brasil Central (DA MOTTA et al., 2002).

Nos experimentos, com dose ideal de N (AN) foram aplicados 150 kg.ha⁻¹ de N (27 kg.ha⁻¹ no plantio e 133 kg.ha⁻¹ em cobertura) e no baixo N (BN), 60 kg.ha⁻¹ de N (27 kg.ha⁻¹ no plantio e 33 kg.ha⁻¹ em cobertura). O manejo e os demais tratos culturais foram realizados de acordo com os procedimentos recomendados para a cultura do milho (BORÉM et al., 2015). Apesar de ter sido observada correlação positiva entre os níveis de N e a produtividade, ou seja, os genótipos apresentarem comportamento diferencial ao variar o nível de N, essa interação foi do tipo simples para todos os casos e, assim, pôde-se dar prosseguimento às análises estatístico-genéticas utilizando a média dos valores observados em ambos os níveis de N.

Cada ambiente foi constituído pela combinação Ano x Local x Nível de N. A painel 1 foi avaliada, portanto, em seis ambientes. Isto porque no ano agrícola de 2014 foi feita a avaliação em apenas um local, na Estação Experimental Anhumas, sob os dois níveis de N.

3.1.2 Híbridos

Os 452 híbridos utilizados foram obtidos a partir de cruzamentos entre 128 linhagens endogâmicas da empresa Helix Sementes®, em esquema de dialelo completo sem recíprocos, (Painel 2). Estes foram avaliados quanto a produtividade de grãos (PG) em cinco locais no Brasil (Tabela 1)

Para a avaliação dos híbridos cada parcela foi constituída de duas fileiras com 5 m de comprimento. O espaçamento entre linhas foi de 0,70 m e 22 covas por fileira, totalizando 44 plantas por parcela. A produtividade de grãos (ton.ha⁻¹) proveniente da colheita de toda a parcela, foi corrigida para umidade de grãos (13%) e extrapolada para um ha. O delineamento experimental utilizado foi o de blocos completos casualizados, com duas repetições. Os experimentos foram conduzidos e colhidos de forma mecanizada no ano agrícola de 2014/15.

Tabela 1 - Descrição dos ambientes de avaliação das linhagens (Painel 1) e híbridos simples de milho tropical (Painel 2)

Ambiente	Linhagens			Ambiente	Híbridos	
	Local	Ano agrícola	Sistema de cultivo		Local	Ano agrícola
1	Anhembi - SP	2013/14	BN	1	Ipiacú – MG	2014/15
2	Anhembi - SP	2013/14	AN	2	Nova Mutum – MT	2014/15
3	Piracicaba - SP	2014/15	BN	3	Patos de Minas - MG	2014/15
4	Piracicaba - SP	2014/15	AN	4	Sertãoópolis – PR	2014/15
5	Anhembi - SP	2014/15	BN	5	Sorriso – MT	2014/15
6	Anhembi - SP	2014/15	AN			

BN – Baixo Nitrogênio; AN – Dose ideal de Nitrogênio

3.2 Genotipagem e controle de qualidade dos dados genômicos

A coleta do material genético foi realizada no estádio de três folhas completamente expandidas (V3). O DNA genômico das linhagens e dos híbridos foi coletado e isolado utilizando o protocolo adaptado de Murray e Thompson (1980).

As amostras das linhagens e dos híbridos foram genotipados com a plataforma Affymetrix® Axiom® Maize Genotyping Array (UNTERSEER et al., 2014) com aproximadamente 660 mil marcadores do tipo SNP, em 2015. Para realizar o controle de qualidade dos marcadores SNP's obtidos, foi empregado o *Call Rate* ($CR \geq 0,95$), que consiste em eliminar SNP's com grande quantidade de valores perdidos e a *Minor Allele Frequency* ($MAF \leq 0,05$), esta relacionada com o polimorfismo dos *loci* marcadores no painel. As análises foram realizadas por meio do pacote HapEstXXR (KNUEPPEL et al., 2015) do software R (R CORE TEAM, 2014). Após o controle de qualidade foram realizadas novas reduções no número de marcas para compor os cenários de GS descritos abaixo.

3.2.1 Redução na densidade de marcadores

Para linhagens foram inicialmente retirados todos aqueles marcadores que apresentavam em pelo menos um indivíduo o genótipo heterozigoto. Posteriormente, tanto para linhagens como para híbridos, foram realizadas reduções utilizando informações fornecidas pela curva de decaimento do desequilíbrio de ligação (*LD decay*) fornecida pela regressão não linear proposta por Hill e Weir (1988):

$$E(r^2) = \left[\frac{10 + C}{(2 + C)(11 + C)} \right] \left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)} \right]$$

em que, N corresponde ao tamanho efetivo da população, c à fração de recombinação entre os locos, C é dado por $4Nc$ e n representa o tamanho da amostra. Estas análises foram efetuadas no software *R* (R CORE TEAM, 2014) utilizando o pacote *synbreed* (WIMMER et al., 2012).

A curva de LD *decay* fornece a tendência de redução do desequilíbrio de ligação em função do aumento da distância entre os marcadores. Neste trabalho foi estimada uma curva de decaimento do desequilíbrio de ligação para cada cromossomo do milho. A partir de cada curva estimada foram obtidas duas janelas (intervalos entre marcadores em kb) nas quais o LD observado correspondia a 0,2 ou a 0,02. Após determinar tais intervalos, foram selecionadas marcas, a partir do primeiro SNP do cromossomo, visando representar segmentos cromossômicos que segreguem de forma independente, conforme proposto por Rincent (2014). Com isto, foram formados quatro cenários de densidade de marcadores para as linhagens (400k, 60k, 9586 e 1304 SNP's) e três para os híbridos (50k, 4458 e 495 SNP's).

3.3 Análises genético-estatísticas

As análises foram realizadas por meio de equações de modelos mistos, do melhor preditor linear não viesado (*Best Linear Unbiased Prediction* – BLUP), em conjunto com o método máxima verossimilhança restrita (*Restricted maximum likelihood* – REML) (HENDERSON, 1949; THOMPSON, 1975). No primeiro modelo foi desconsiderado o parentesco entre os indivíduos (BLUP_I), assim a matriz de variância-covariância genética aditiva entre os indivíduos foi uma matriz identidade. Este modelo incorpora os valores observados, obtidos por meio da fenotipagem, para produtividade de grãos (ton. ha⁻¹), corrigida para umidade de grãos a 13%, mimetizando a conjuntura do melhoramento clássico.

Nos modelos seguintes foram considerados quatro diferentes tipos de matrizes de parentesco genômico (**G**), as quais serão descritas adiante.

3.3.1 BLUP_I e G-BLUP

A predição de parâmetros genéticos e dos valores genéticos foi realizada considerando o modelo genético-estatístico:

$$\mathbf{y} = \mathbf{Ht} + \mathbf{Xr} + \mathbf{Zs} + \mathbf{Wb} + \mathbf{Kc} + \mathbf{e}$$

em que, y é o vetor de observações dos indivíduos para produtividade de grãos; t é o vetor de efeito de genótipos, em que $t \sim N(0, T)$, sendo $T = G\sigma_t^2$; r é o vetor do efeito de ambiente somados à média geral, considerado como fixo; s é o vetor do efeito de repetição dentro de ambiente, considerado como fixo; b é o vetor do efeito de blocos dentro de repetição alocados dentro de local, em que $b \sim N(0, B)$, sendo que $B = I\sigma_b^2$; c é o vetor dos efeitos de interação genótipo x ambiente, em que $c \sim N(0, C)$, sendo que $C = I\sigma_c^2$; e é o erro experimental médio, em que $e \sim N(0, \epsilon)$, sendo $\epsilon = I\sigma^2$. X , Z , W são matrizes de incidência que relacionam os efeitos dos vetores independentes de cada matriz ao vetor dependente. G representa as matrizes de parentesco genômico entre os indivíduos, as quais foram computadas à parte e então incorporadas ao modelo. Estas foram obtidas por meio do pacote `snprReady` (em preparo) (GRANATO; FRITSCHÉ-NETO, 2016). Todas foram desenvolvidas por meio de *scripts* próprios desenvolvidos na linguagem S4 do R (R CORE TEAM, 2014). Já as análises de modelos mistos, tanto de BLUP_I como os de G-BLUP, foram realizadas utilizando o pacote do R, ASREML/R (BUTLER et al., 2009).

3.3.2 Matriz de VanRaden (2008) (G_{vr})

A matriz G_{vr} proposta por Vanraden (2008) é função da própria matriz $W'W$ presente nas equações de modelos mistos. A matriz de variância-covariância de parentesco genômico aditivo (G_{vr}) foi incorporada nas equações do BLUP tradicional, conforme apresentado por Resende et al. (2012b):

$$G_{vr} = \frac{(W^* W^{*'})}{\sum_i^n [2p_i(1-p_i)]},$$

em que, p_i é a frequência de um dos alelos do *locus* i e W^* é a matriz de incidência de marcadores dos indivíduos, corrigida para suas médias em cada *locus* ($2p_i$). Para garantir G como uma matriz positiva definida, somou-se uma constante à diagonal principal da matriz formada, obtida por meio de $G = G^p + 10^{-6}I$, em que I é uma matriz identidade.

3.3.3 Matriz de Astle e Balding (2009) ($\mathbf{G}_{ab}$)

A matriz de Astle e Balding (2009), $\mathbf{G}_{ab}$, não leva em consideração o LD entre os SNP's e atribui maior peso nos *loci* com baixa diversidade genética, sendo obtida por:

$$\mathbf{G}_{ab} = \frac{1}{L} \sum_{l=1}^L \frac{(x_l - 2p_l)(x_l - 2p_l)^T}{4p_l(1 - p_l)}$$

em que, p_l é a frequência do alelo no loco 1, x_l é o genótipo dos indivíduos e L é o número de marcadores considerado. Nota-se que este é um estimador semi-definido imparcial e positivo para a matriz de parentesco $\mathbf{G}_{ab}$. Assim, $\hat{\mathbf{G}}_{ab}$ pode superestimar o compartilhamento de alelos esperado para indivíduos não aparentados. Na prática, não sabemos as frações de alelos p_l não oriundos de famílias endogâmicas e não relacionados. Este desvio pode subestimar $(\mathbf{G}_{ab})_{ij}$ e superestimar p_l a partir dos mesmos dados. Embora as correlações decorrentes da ancestralidade comum sejam, em princípio, positivas, devido à estimativa viesada de p_l , os elementos fora da diagonal de $\hat{\mathbf{G}}_{ab}$ podem ser negativos, propriedade que tem levado alguns autores a evitar tais estimadores de $\mathbf{G}_{ab}$ (RINCENT, 2014).

3.3.4 Matriz de parentesco unificado - Yang (2010) ($\mathbf{G}_{yu}$)

As matrizes de Yang et al. (2010) são a matriz de parentesco unificado $\mathbf{G}_{yu}$, a qual pode gerar valores subestimados de herdabilidade devido ao desequilíbrio de ligação incompleto entre SNP's e variantes causais e a matriz de parentesco unificado ajustado $\mathbf{G}_{ya}$ (variantes causais têm MAF's mais baixos do que os SNP's e LD menor do que entre marcadores e variantes causais). Todavia, o coeficiente de parentesco ($\mathbf{G}_{yu}$) entre indivíduos diferentes é dado por:

$$\mathbf{G}_{yu} = \mathbf{A}_{ij} = \frac{1}{m} \sum_k \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1 - p_k)}, \text{ para } i \neq j$$

em que, x é um indicador do SNP k nos indivíduos i e j , p_k é a frequência alélica (frequência de cada alelo) do SNP k , e m é o número de SNP's na amostra.

Já o parentesco de um indivíduo com ele mesmo é estimado:

$$\mathbf{G}_{yu'} = \mathbf{A}_{ij} = 1 + \frac{1}{m} \sum_k \frac{x_{ik}^2 - (1 + 2p_k)(x_{ik} + 2p_k^2)}{2p_k(1 - p_k)}, \text{ para } i = j$$

em que, x é o indicador do SNP k nos indivíduos i , p_k é a frequência alélica (frequência de cada alelo) do SNP k , e m é o número de SNP's na amostra.

3.3.5 Matriz de parentesco unificado ajustado - Yang (2010) ($\mathbf{G}_{ya}$)

A matriz de parentesco unificado ajustado é obtida pela regressão da descrita anteriormente ($\mathbf{G}_{yu}$) em função da matriz de identidade, tornando-a imparcial. Com isto, na matriz $\mathbf{G}_{ya}$, as estimativas de parentesco foram corrigidas pelo erro amostral do método anterior ($\sigma_e^2 = 1/m$) por meio da regressão de $\mathbf{A}$ em $\hat{\mathbf{A}}(1/m)$. O coeficiente de regressão, obtido por $\frac{\text{var}(\mathbf{A})}{\text{var}(\mathbf{A}) + \frac{1}{m}}$ e $\text{var}(\hat{\mathbf{A}}) = \text{var}(\mathbf{A}) + \frac{1}{m}$. Assim, a $\mathbf{G}_{yu}$ ajustada ($\mathbf{G}_{ya}$) é dada por:

$$\mathbf{G}_{ya} = \mathbf{A}_{ij}^* = \left(1 - \frac{\frac{1}{m}}{\text{var}(\hat{\mathbf{A}})} \right) \hat{\mathbf{A}}, \text{ para } i \neq j;$$

e

$$\mathbf{G}_{ya} = \mathbf{A}_{ij}^{*'} = 1 + \left(1 - \frac{\frac{1}{m}}{\text{var}(\hat{\mathbf{A}})} \right) (\hat{\mathbf{A}} - \mathbf{I}), \text{ para } i = j.$$

3.4 Validação cruzada e acurácia nos diferentes cenários de GS

Para as linhagens, utilizando o método (*k-fold*), a validação cruzada foi feita considerando dois grupos de 32 indivíduos ($k=2$). Já para os híbridos, foi utilizado um sistema de validação cruzada na qual eles foram divididos em cinco grupos ($k=5$), sendo que em ambas as situações cada grupo de indivíduos foi predito pelos demais. O processo de reamostragem dos grupos foi realizado 500 vezes. Para cada um dos cenários testados (tipo de painel x tipo de matriz de parentesco genômico x densidade de marcadores), a cada

reamostragem a capacidade preditiva (r_{yy}), estimada pela correlação entre o valor genotípico predito e o valor fenotípico observado nos indivíduos, foi estimada.

A partir destas reamostragens foi estimada para cada cenário a herdabilidade no sentido amplo (h^2) a nível de indivíduo:

$$h^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \sigma_{ge}^2 + \sigma_e^2)}$$

em que, σ_g^2 é a variância genotípica; σ_{ge}^2 , a variância da interação genótipo x local e σ_e^2 , a variância residual.

Outro parâmetro estimado foi a acurácia de predição via marcadores: $\hat{r}_{gg} = r_{yy} / \sqrt{h^2}$, em que r_{yy} é a capacidade preditiva.

A coincidência da seleção baseada na predição de valores genotípicos por meio de matrizes de parentesco genômico para a PG (ton.ha⁻¹) com aqueles indivíduos selecionados a partir do BLUP sem informação de parentesco foi estimada a partir da expressão proposta por Hamblin e Zimmermann (1986):

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

em que, C : é a intensidade de seleção desconsiderando o parentesco entre os (i) genótipos selecionados; A : é o número de genótipos selecionados com base nas matrizes de parentesco genômico que coincidiram com os selecionados desconsiderando tais matrizes; M : é o número total de genótipos selecionados.

Tal estimativa quantifica a percentagem de genótipos selecionados com base nas matrizes de parentesco genômico ($\mathbf{G}$) que coincidiram com aqueles selecionados desconsiderando a informação de parentesco. O valor estimado da coincidência de seleção é máximo quando os genótipos selecionados com base nas matrizes $\mathbf{G}$ coincidirem totalmente com os indivíduos selecionados apenas desconsiderando o parentesco.

Por fim, usando as acurácias preditivas de cada cenário como variável resposta foi realizada uma análise de variância, considerando o seguinte modelo:

$$y_{isp} = \mu + t_i + d_{s(i)} + m_{p(i)} + e_{isp}$$

em que μ é a média geral das acurácias preditivas; t_i é o efeito fixo de tipo de painel; $d_{s(i)}$ é o efeito fixo de densidade de marcadores dentro de tipo de painel; $m_{p(i)}$ é o efeito fixo de tipo

de matriz de parentesco genômico dentro de tipo de painel e e_{isp} é o efeito residual do modelo, em que $e_{isp} \sim (0, \sigma^2)$.

4 RESULTADOS E DISCUSSÃO

4.1 Comparação dos diferentes cenários de Seleção Genômica (GS)

Para a comparação dos cenários, em que variaram o método de obtenção da matriz de parentesco genômico, a densidade de marcadores SNP's e o tipo de painel foi realizada uma análise de variância das acurácias preditivas, estimadas para cada uma das 500 reamostragens do processo de validação cruzada.

Tabela 2 - Análise de variância conjunta das acurácias preditivas obtidas, considerando tipo de painel, densidade de marcadores e método de obtenção da matriz de parentesco genômico

Fator de variação	GL	QM
Painel	1	453,61 ^{**}
Densidade/Painel	5	3,48 ^{**}
Matriz/Painel	6	0,006 ^{ns}
Resíduo	13987	0,008

ns - não significativo; ** - significativo a 1% pelo teste F

Como era esperado, foi observado efeito significativo para tipo de painel. Com isto, decidiu-se efetuar análises individuais para cada painel, uma vez que as densidades de marcadores empregadas dentro de cada uma eram diferentes (Tabela 3).

Tabela 3 - Análise de variância individual das acurácias preditivas obtidas para linhagens e híbridos considerando diferentes densidades de marcadores e método de obtenção da matriz de parentesco genômico

Fator de variação	Linhagens		Híbridos	
	GL	QM	GL	QM
Densidade	3	5,80 ^{**}	2	0,004 ^{**}
Matriz	3	0,0002 ^{ns}	3	0,012 ^{**}
Resíduo	7993	0,015	5994	0,0003

ns - não significativo; ** - significativo a 1% pelo teste F

A densidade de marcas utilizada e o tipo de matriz de obtenção de parentesco genômico apresentaram efeito significativo, quando analisadas dentro de tipo de painel (grau de heterozigosidade). A partir dessa observação, foram realizados testes de *Tukey* para cada fonte de variação do modelo (Tabela 4 e Tabela 5). Por meio destas comparações foi possível identificar as matrizes de parentesco genômico mais indicadas para cada painel, assim como para densidade de marcadores.

Tabela 4 – Teste de *Tukey* para acurácias preditivas médias para método de obtenção da matriz de parentesco genômico na painel de linhagens

Matriz de Parentesco Genômico	$\hat{r}_{gg}$	Densidade de marcadores	$\hat{r}_{gg}$
$\mathbf{G}_{vr}$	0,271 A ⁱ	60k	0,303 A
$\mathbf{G}_{yu}$	0,271 A	400k	0,296 A
$\mathbf{G}_{ya}$	0,270 A	9586 [§]	0,294 A
$\mathbf{G}_{ab}$	0,270 A	1304 ^{§§}	0,190 B

ⁱLetras diferentes na coluna indicam diferença significativa a 0,05 de probabilidade pelo teste de *Tukey*; [§] - redução do número de SNP's com base no LD $\cong 0,2$; ^{§§} - redução do número de SNP's com base no LD $\cong 0,02$; $\hat{r}_{gg}$ é a acurácia preditiva; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

Tabela 5 - Teste de *Tukey* para acurácias preditivas médias para método de obtenção da matriz de parentesco genômico e densidade de marcadores na painel de híbridos

Matriz de Parentesco Genômico	$\hat{r}_{gg}$	Densidade de marcadores	$\hat{r}_{gg}$
$\mathbf{G}_{ya}$	0,637 A ⁱ	50k	0,635 A
$\mathbf{G}_{yu}$	0,636 A	495 ^{§§}	0,636 A
$\mathbf{G}_{vr}$	0,632 B	4458 [§]	0,633 B
$\mathbf{G}_{ab}$	0,631 B	-	-

ⁱLetras diferentes na coluna indicam diferença significativa a 0,05 de probabilidade pelo teste de *Tukey*; [§] - redução do número de SNP's com base no LD $\cong 0,2$; ^{§§} - redução do número de SNP's com base no LD $\cong 0,02$; $\hat{r}_{gg}$ é a acurácia preditiva; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

4.1.1 Densidade de marcadores

Para linhagens não houve diferença entre o tipo de matriz, mas sim, entre as diferentes densidades de marcas usadas, sendo a mais indicada aquela considerando o LD $\cong 0,2$, isto porque a acurácia preditiva média é a mesma ao se utilizarem 400k e 9586 marcadores SNP's na genotipagem deste painel. Por outro lado, quando realizada a segunda redução, com base no desequilíbrio de ligação, LD $\cong 0,02$, notou-se queda acentuada na acurácia preditiva ($\hat{r}_{gg}$), ou seja, para todas as matrizes isto foi prejudicial, pois provavelmente foram eliminados marcadores que eram informativos para indicar presença ou não de parentesco e também em LD com os QTL's que controlam a característica.

No painel de híbridos, a redução considerando $LD \cong 0,02$ foi prejudicial. Entretanto a redução considerando $LD \cong 0,2$ não afetou a $\hat{r}_{gg}$ em termos de acurácia e apresenta-se como o cenário mais adequado, por proporcionar a melhor relação custo benefício, já que os 495 SNP's foram suficientes para produzir resultados satisfatórios e semelhantes ao cenário de densidade.

Os *chips* de genotipagem de alta densidade tornaram possível a alta cobertura do genoma com marcadores (um marcador a cada 50.000 pares de bases), viabilizando particionar 100% da variabilidade genética aditiva de uma característica por todo genoma e prever o *breeding value* sem a coleta de dados fenotípicos do indivíduo (MEUWISSEN; GODDARD, 2001). As tecnologias de genotipagem de SNP's foram aplicadas primeiramente para desenvolvimento de ensaios diagnósticos para doenças genéticas (HOLLOWAY et al., 1999), e a demanda atual é testar um número baixo de polimorfismos (SNP's) em um grande número de amostras, com rapidez e alta acurácia.

Assim, o fato de poucas marcas terem sido suficientes para produzir resultados satisfatórios para predição de valores genéticos genômicos de forma altamente acurada, traz grandes vantagens para a implementação de variadas plataformas e painéis (*chips*) de menor densidade, levando a melhorias de tecnologias e redução de custos (HEATON et al., 2005; CHESSA et al., 2007), assim como da demanda computacional.

4.1.2 Tipo de matriz de parentesco

No painel de híbridos, os métodos de estimação da matriz de parentesco genômico que proporcionaram maiores acurácias preditivas foram os propostos por Yang *et al.* (2010). Isso pode estar associado ao fato de a similaridade em $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$ ser maior se dois indivíduos compartilharem alelos raros do que alelos comuns (RINCENT, 2014), apresentando, assim, maior acurácia preditiva. Assim como a herdabilidade, a arquitetura genética de caracteres de interesse afeta o desempenho da predição de modelos de GS, tanto para caracteres complexos (PG) quanto para caracteres qualitativos (ZHANG et al., 2015). Além do mais, as matrizes de Vanraden (2008) e Astle e Balding (2009) não levam em conta o desequilíbrio de ligação entre os SNP's, quando há presença de alto LD ou menor diversidade em determinados *loci*. Estes tendem a contribuir mais na estimação do parentesco, o qual é superestimado e não provê a acurácia preditiva desejável para ser incorporada em modelos mistos de GS. Nota-se

também que estes são mais estáveis para diferentes cenários de LD entre marcadores e densidade de genotipagem.

Por outro lado, no painel de linhagens não houve diferença entre o tipo de matriz quanto à acurácia preditiva. Entretanto há evidências de que entre elas há alguma que pode ser mais adequada para predição de *breeding values* via Seleção genômica (*Genomic Selection* – GS).

Devido ao elevado grau de homozigose e, portanto, maior atuação dos efeitos genéticos aditivos sobre a expressão do caráter PG em linhagens endogâmicas de milho, as matrizes de parentesco genômico que, aparentemente, subestimaram o parentesco entre indivíduos podem ser as mais realistas, por terem captado maior parte da variância devido aos efeitos genéticos aditivos. Subestimativas do parentesco foram observadas, indiferente do tipo de matriz de parentesco utilizada, por Oliehoek et al. (2006), os quais afirmaram que quando alelos idênticos de uma painel-base desconhecida são erroneamente estimados, o parentesco médio no painel atual é viesado.

Segundo Weir et al. (2006) o parentesco entre indivíduos pode ser subestimado se o efeito de estrutura do painel for ignorado. Além disso, como o parentesco médio observado e a probabilidade média de os alelos serem idênticos por estado definem um painel, se esta probabilidade é superestimada isso também pode resultar na subestimação do parentesco genético genômico (TORO et al., 2003). Dessa forma é válida a análise da herdabilidade (h^2), entre outros parâmetros genéticos e estatísticos, para inferir sobre qual metodologia de obtenção da matriz de parentesco genômico é mais fidedigna à realidade em painéis de linhagens.

4.1.2.1 Parâmetros genéticos e herdabilidade (h^2)

Foi realizada a análise de deviance (ANADEV) para o caráter produtividade de grãos (PG), depois de verificado que havia homogeneidade das variâncias entre os experimentos, por meio do método proposto por (PIMENTEL-GOMES, 2000). Foram constatadas diferenças significativas por meio do teste de razão de verossimilhança (*Likelihood-Ratio Test* - LRT) ($p < 0,01$) para o efeito de genótipo, tanto nas linhagens como nos híbridos. Isto indica a existência de variabilidade genética nos painéis estudados para o caráter produtividade de grãos. A interação genótipo x ambiente também foi significativa, indicando que os materiais

apresentaram comportamento diferencial em função dos ambientes em que foram avaliados (Tabela 6 e 7).

Tabela 6 – Componentes de variância e herdabilidade no sentido amplo em nível de indivíduos (h^2) sob diferentes densidades de SNP's e matrizes de parentesco genômico para produtividade de grãos (PG) em linhagens de milho tropical avaliado em oito ambientes

	Densidade	I	G_{vr}	G_{ab}	G_{yu}	G_{ya}
σ_g^2	400k	342000,71**	370094,70**	1609399,93**	199768,88**	199775,18**
	60k		371587,29**	1470394,54**	181632,09**	181670,68**
	9586§		382030,79**	1507549,51**	186094,41**	186231,74**
	1304§§		452614,55**	1654967,59**	203667,45**	204181,61**
σ_{ge}^2	400k	227247,02**	229724,10**	229722,71**	229275,92**	229274,15**
	60k		230100,08**	230073,23**	230669,94**	230661,87**
	9586§		230951,86**	230932,88**	230710,85**	230698,07**
	1304§§		232532,78**	231902,12**	231800,72**	231625,39**
σ_e^2	400k	311365,97	311365,97	311365,97	311365,97	311365,97
	60k		311365,97	311365,97	311365,97	311365,97
	9586§		311365,97	311365,97	311365,97	311365,97
	1304§§		311365,97	311365,97	231800,72	311365,97
h^2	400k	0,84	0,85	0,96	0,76	0,76
	60k		0,85	0,96	0,74	0,74
	9586§		0,85	0,96	0,74	0,74
	1304§§		0,87	0,96	0,76	0,76

LRT - Likelihood-Ratio Test; ns - não significativo; * - significativo a 5% pela LRT ou F Wald; ** - significativo a 1% pela LRT ou F Wald. § - redução do número de SNP's com base no LD $\cong 0,2$; §§ - redução do número de SNP's com base no LD $\cong 0,02$; σ_g^2 - variância dos efeitos genéticos; σ_{ge}^2 - variância da interação dos efeitos genéticos x ambiente; σ_e^2 - variância dos efeitos ambientais; **G_{vr}** - matriz de Vanraden; **G_{ab}** - matriz de Astle e Balding; **G_{yu}** - matriz de parentesco unificado de Yang; **G_{ya}** - matriz de parentesco unificado ajustado de Yang

Ao negligenciar o parentesco entre os indivíduos, matriz **I**, tem-se a condição de maior variabilidade genética, pois os indivíduos são considerados como não relacionados e o tamanho efetivo é igual ao número de indivíduos (WRIGHT, 1922). Neste contexto, as estimativas dos componentes de variância na painel de linhagens indicaram superestimação dos parâmetros genéticos quando obtidos pela matriz **G_{ab}**. Já aqueles obtidos pelo uso das matrizes **G_{yu}** e **G_{ya}** apresentaram valores menores em relação à matriz identidade (**I**),

devido ao ajuste do tamanho efetivo (N_e) da painel pelo parentesco (Tabela 6). O cenário com a matriz $\mathbf{G}_{vr}$ foi a mais semelhante à situação de negligência do possível parentesco entre as linhagens avaliadas.

As herdabilidades estimadas (h^2) (Tabela 7) foram superiores com o uso das matrizes $\mathbf{G}_{ab}$. Outro resultado interessante foi que as herdabilidades estimadas foram as mesmas para $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$, independente da densidade. Isso não era esperado, pois no primeiro caso, valores de herdabilidade e capacidade preditiva deveriam ser menores como consequência da inadequada estimação do valor genético quando não se corrige para efeito de estrutura de painel (POWELL et al., 2010). Por outro lado, indica que estes métodos são consistentes mesmo sob diferentes densidades de marcadores e níveis de LD.

Em relação ao número de SNP's utilizados, em geral, quando se diminuiu a densidade de marcadores aumentaram-se as estimativas dos componentes de variância genética e de interação. Isto era esperado, uma vez que modificações na densidade de marcadores tendem a acentuar as diferenças entre os modelos (CROSSA et al., 2013). Entretanto, as herdabilidades apresentaram pouca variação dentro de cada tipo de parentesco utilizado, indicando a proporcionalidade de aumento de ambos os componentes. Considerando todas as densidades de marcas, a matriz $\mathbf{G}_{ab}$ proporcionou a maior herdabilidade, o que consequentemente proporcionaria ainda maior relação entre genótipo e fenótipo.

Tabela 7 – Componentes de variância e herdabilidade no sentido amplo em nível de indivíduos (h^2) sob diferentes densidades de SNP's e matrizes de parentesco genômico para produtividade de grãos (PG) em híbridos de milho tropical avaliado em cinco ambientes

	Densidade	I	G_{vr}	G_{ab}	G_{yu}	G_{ya}
σ_g^2	50k		652899,69**	1191158,86**	5311440,67**	1190724,43**
	4458 [§]	856018,27**	666364,78**	1263973,27**	5386283,21**	1256339,58**
	495 ^{§§}		670363,84**	1355434,07**	5377376,41**	1243609,96**
σ_{ge}^2	50k		1020539,97**	852763,08**	1020185,08**	852362,73**
	4458 [§]	757709,94**	1022726,57**	859661,14**	1021631,13**	856025,27**
	495 ^{§§}		1027747,74**	853319,03**	1026842,53**	827586,04**
σ_e^2	50k		937417,54	937417,59	937417,54	937417,58
	4458 [§]	937417,54	937417,54	937417,54	937417,54	937417,54
	495 ^{§§}		937417,54	937417,54	937417,54	937417,54
h^2	50k		0,69	0,95	0,82	0,82
	4458 [§]	0,78	0,69	0,95	0,83	0,82
	495 ^{§§}		0,69	0,95	0,82	0,82

LRT - Likelihood-Ratio Test; ns - não significativo; * - significativo a 5% pela LRT ou F Wald; ** - significativo a 1% pela LRT ou F Wald. § - redução do número de SNP's com base no LD $\cong 0,2$; §§ - redução do número de SNP's com base no LD $\cong 0,02$; σ_g^2 - variância dos efeitos genéticos; σ_{ge}^2 - variância da interação dos efeitos genéticos x ambiente; σ_e^2 - variância dos efeitos ambientais; G_{vr} - matriz de Vanraden; G_{ab} - matriz de Astle e Balding; G_{yu} - matriz de parentesco unificado de Yang; G_{ya} - matriz de parentesco unificado ajustado de Yang

Assim como na painel de linhagens houve uma tendência de aumento dos componentes de variância conforme a densidade de marcadores diminuiu, na painel de híbridos. Em relação à herdabilidade, a matriz G_{ab} novamente mostrou-se superior às demais.

A matriz G_{vr} foi a que apresentou menor herdabilidade. Como no modelo misto, foi considerado apenas o efeito aditivo dos SNP's, a variância genética não aditiva e variância devido às interações gene-ambiente podem ter existido, mas não contribuíram diretamente para a herdabilidade (POWELL et al., 2010).

Deve-se chamar atenção para o caso da matriz G_{ab} (ASTLE; BALDING, 2009), cujos resultados indicaram a superestimação dos parâmetros do modelo. Os componentes genéticos superaram o limite teórico dado pelo cenário do BLUP_I. Portanto tal matriz não é recomendável nessas condições de análise, pois parâmetros genéticos importantes, como a

herdabilidade, são superestimados, o que influencia diretamente na predição dos valores genéticos genômicos e na resposta a seleção (Tabela 7).

Tabela 8 – Indicadores de eficiência do REML/BLUP considerando a matriz identidade e as matrizes $\mathbf{G}_{vr}$, $\mathbf{G}_{ab}$, $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$, respectivamente, para produtividade de grãos (PG), com diferentes densidades de marcadores em linhagens de milho tropical

Matriz de Parentesco Genômico	Densidade de marcadores	r_{yy}	$\hat{r}_{gg}$
$\mathbf{G}_{vr}$	400k	0,279 ^(0,109)	0,30
	60k	0,285 ^(0,112)	0,31
	9586 [§]	0,271 ^(0,105)	0,30
	1304 ^{§§}	0,163 ^(0,116)	0,18
$\mathbf{G}_{ab}$	400k	0,270 ^(0,109)	0,29
	60k	0,275 ^(0,112)	0,30
	9586 [§]	0,267 ^(0,105)	0,29
	1304 ^{§§}	0,178 ^(0,114)	0,19
$\mathbf{G}_{yu}$	400k	0,270 ^(0,109)	0,29
	60k	0,274 ^(0,112)	0,30
	9586 [§]	0,269 ^(0,105)	0,29
	1304 ^{§§}	0,177 ^(0,113)	0,19
$\mathbf{G}_{ya}$	400k	0,270 ^(0,109)	0,29
	60k	0,274 ^(0,112)	0,30
	9586 [§]	0,268 ^(0,105)	0,29
	1304 ^{§§}	0,177 ^(0,113)	0,19

§ - redução do número de SNP's com base no LD $\cong 0,2$; §§ - redução do número de SNP's com base no LD $\cong 0,02$; r_{yy} é a capacidade preditiva; $\hat{r}_{gg}$ é a acurácia preditiva; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

As médias das acurácias preditivas e capacidades preditivas em linhagens (Tabela 8) das matrizes de parentesco para cada densidade de marcas foram maiores quando incorporadas a $\mathbf{G}_{vr}$ ou $\mathbf{G}_{yu}$ no modelo misto. Assim, para a comparação de genótipos, os valores genéticos genômicos preditos por elas são os mais acurados, assim como já demonstrado por outros autores (QUAAS, 1988; PIEPHO et al., 2008; CHEN et al., 2011).

Tanto r_{yy} quanto $\hat{r}_{gg}$ foram similares ao considerar matrizes distintas, e ligeiramente diferentes para densidades de marcadores em cada matriz. Dentre as matrizes, os maiores valores de r_{yy} foram observados nas maiores densidades de marcadores, apresentando queda brusca quando se reduziu sensivelmente a densidade (LD $\cong 0,02$). Isto ocorreu, pois o ajuste

do modelo com a maior densidade de marcadores possui atributos que garante uma boa capacidade preditiva. Isto porque o número de marcas (1.304) considerado pode não ter sido suficiente para explicar o parentesco (YU et al., 2009; ZHONG et al., 2009). Já o modelo que incorpora a matriz identidade não considera as relações existentes entre os indivíduos, sendo menos realista pela incapacidade de os preditores usarem a informação dos indivíduos por meio da correlação genética que surge da genealogia, a qual foi baseada na correlação entre SNP's dos indivíduos do painel.

Neste contexto, é importante ressaltar que o cenário utilizado para predição dos valores genéticos contou com a correção para LD incompleta entre SNP's e variantes causais. Se este procedimento estiver correto, a variância explicada pelas variantes causais não deve depender do número de SNP's usados em apenas 64 linhagens, o que pode ser considerado um número baixo de indivíduos. Para as estimativas de parentesco, a proporção da variância explicada aumenta com o número de SNP's, pois o erro de predição é reduzido por meio da inclusão de mais de SNP's (YANG et al., 2010). No estudo de caracteres complexos, é necessário encontrar SNP's individuais que estejam significativamente associados com eles, pois a variância tipicamente explicada por cada SNP é pequena e tamanhos grandes de amostras são necessários para demonstrar que os seus efeitos individuais são estatisticamente significativos, capazes de fornecer alta capacidade preditiva.

Embora não haja comprovação de que os valores de capacidade preditiva e acurácia preditiva difiram estaticamente, podem-se eleger as melhores combinações de matrizes de parentesco e densidade de marcadores seriam mais adequadas para emprego nos modelos de GS. Notou-se que os maiores valores de r_{yy} e $\hat{r}_{gg}$ decorreram da combinação da matriz de Vanraden (2008) com a densidade de 60k, e sofreram grande queda quando realizada a redução para $LD \cong 0,2$ (Tabela 8). Isto indica que matrizes consagradas, as quais podem ser consideradas mais elementares por exigirem menor tempo computacional foram suficientes para analisar dados de similaridade genética e fornecer resultados não-viesados. Além disso, uma densidade de marcadores nem sempre é significado de maior acurácia preditiva. Resultado semelhante foi encontrado por Zhang et al. (2015), em que baixa densidade de SNP's (~ 200 marcadores) mostrou-se suficiente para uma boa predição em painéis biparentais de milho, tanto para características simples quanto para as mais complexas.

Tabela 9 – Indicadores de eficiência do REML/BLUP considerando a matriz identidade e as matrizes $\mathbf{G}_{vr}$, $\mathbf{G}_{ab}$, $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$, respectivamente, para produtividade de grãos (PG), com diferentes densidades de marcadores em híbridos de milho tropical

Matriz de Parentesco Genômico	Densidade de marcadores	r_{yy}	$\hat{r}_{gg}$
$\mathbf{G}_{vr}$	50k	0,559 ^(0,0151)	0,63
	4458 [§]	0,556 ^(0,0152)	0,63
	495 ^{§§}	0,558 ^(0,0147)	0,63
$\mathbf{G}_{ab}$	50k	0,558 ^(0,0150)	0,63
	4458 [§]	0,556 ^(0,0151)	0,63
	495 ^{§§}	0,558 ^(0,0147)	0,63
$\mathbf{G}_{yu}$	50k	0,562 ^(0,0154)	0,64
	4458 [§]	0,560 ^(0,0154)	0,63
	495 ^{§§}	0,563 ^(0,0148)	0,64
$\mathbf{G}_{ya}$	50k	0,562 ^(0,0154)	0,64
	4458 [§]	0,561 ^(0,0153)	0,64
	495 ^{§§}	0,564 ^(0,0145)	0,64

§ - redução do número de SNP's com base no $LD \cong 0,2$; §§ - redução do número de SNP's com base no $LD \cong 0,02$; r_{yy} é a capacidade preditiva; $\hat{r}_{gg}$ é a acurácia preditiva; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

Comparados às 64 linhagens, os 452 híbridos constituem um grupo de indivíduos de maior tamanho efetivo (N_e), heterozigosidade, e estruturação populacional. As acurácias preditivas e capacidades preditivas foram similares ao considerar as quatro matrizes e consideravelmente superiores aos das linhagens. Neste caso, o número de marcas não influenciou, pois os híbridos constituem um grupo estruturado, com maior número de indivíduos altamente aparentados entre si, o que gerou maior capacidade preditiva, independentemente da densidade de marcadores. Para níveis coincidentes de herdabilidade (a herdabilidade fenotípica) a capacidade preditiva atinge um máximo e tende a se estabilizar a partir da inclusão de maior captura da variância genética devida à estrutura de painel. Este resultado concorda com o reportado por Janss et al. (2012), cujo estudo incorporando o ajuste dos autovetores de $\mathbf{G}$ como covariáveis do modelo fornece uma correção para essa estruturação. Tais autovetores, no contexto deste estudo, eram informativos sobre os indivíduos que dominam as relações de parentesco e os agrupam em subgrupos estruturados.

A partir dos resultados de redução da densidade via LD nota-se que os SNP's informativos não foram perdidos, por isso, este processo não influenciou tanto na capacidade

preditiva e acurácia quanto o observado nas linhagens. Assim, o número de marcadores utilizados na elaboração de matrizes de parentesco genômico aditivo não apresentou relação linear com os parâmetros estatísticos acurácia preditiva e capacidade preditiva. Isto se torna regra a menos que a escolha dos marcadores não seja coerente com o número de segmentos cromossômicos segregantes ou se for desconsiderado o desequilíbrio de ligação entre esses (RINCENT, 2014).

4.1.2.2 *Heatmap*

Com o uso de *heatmaps* é possível identificar a estrutura hierárquica de agrupamentos de linha e coluna em uma matriz de dados. Chamados de *cluster heatmaps*, tais figuras funcionam como um *display* em que as linhas e colunas representam os indivíduos das painéis, em que cada célula é colorida com base no nível de parentesco entre indivíduos, variando do azul (indivíduos pouco aparentados) ao vermelho (indivíduos altamente aparentados) (HIGA et al., 2010).

No painel de linhagens, em geral foi observada formação de pequenos *clusters*, indiferente do tipo de matriz de parentesco usada (Figura 1), ou seja, a existência de pequenos grupos de indivíduos mais aparentados.

Observou-se em Figura 1(a), que a matriz $\mathbf{G}_{vr}$ foi a que apresentou maior número de indivíduos pouco aparentados (em azul), e também apontou para *clusters* claramente observáveis. Esta matriz apresentou, em média, maior capacidade preditiva (Tabela 8) e seu algoritmo requer menor tempo de processamento computacional. Tendo em vista que o teste de *Tukey* não apontou diferença significativa entre as matrizes, a $\mathbf{G}_{vr}$ apresenta-se como a melhor opção neste contexto. Tal matriz tem como fundamento, desde os primeiros estudos que a envolvem, a captura dos efeitos de substituição alélica, ou seja, prima pelos efeitos genéticos aditivos, os quais se pronunciam com pujança em linhagens endogâmicas de milho. Já a matriz $\mathbf{G}_{ab}$, Figura 1 (b), potencializou o parentesco entre as linhagens, assim como a formação de grupos de indivíduos com elevado grau de parentesco, o que reafirma as características indesejáveis deste tipo de matriz, a superestimação do parentesco e a baixa acurácia preditiva, já apontadas anteriormente. O *heatmap* construído a partir da matriz $\mathbf{G}_{yu}$, (c), e da $\mathbf{G}_{ya}$, (d), semelhantes entre si, indicaram menor parentesco entre as 64 linhagens, exceto nos casos de um indivíduo com ele mesmo. Tais matrizes, como já explanado, tendem

a serem superiores por levarem em consideração o LD. Porém, não apresentaram diferenças significativas em relação as demais.

Figura 1 – Cluster *heatmaps* das linhagens, formado a partir das matrizes de parentesco de ordem 64x64. (a) – matriz $\mathbf{G}_{vr}$; (b) – matriz $\mathbf{G}_{ab}$; (c) – matriz $\mathbf{G}_{yu}$; (d) – matriz $\mathbf{G}_{ya}$; os indivíduos estão representados nas linha e colunas de cada *heatmap*, e cada célula colorida representa o parentesco entre eles; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

No painel de linhagens os indivíduos se mostraram pouco aparentados, com matrizes de parentesco genômico esparsas, gerando modelos altamente correlacionados com o modelo do BLUP fenotípico. Já os híbridos constituíram um grupo mais estruturado, com maior número de indivíduos altamente aparentados entre si. Pôde-se observar em (a), *heatmap* gerado a partir da matriz $\mathbf{G}_{vr}$, o qual apresentou maior número de *clusters* de indivíduos, em relação aos *heatmaps* desenvolvidos a partir das matrizes $\mathbf{G}_{yu}$, (c) e $\mathbf{G}_{ya}$, (d). Em (b) fica

claro que a matriz $\mathbf{G}_{ab}$ gerou um padrão elevado de parentesco (superestimação de parentesco) entre os 452 híbridos, com a presença de dois subgrupos de menos aparentados que os demais. (Figura 2). Estes resultados reiteram que as matrizes $\mathbf{G}_{vr}$ e $\mathbf{G}_{ab}$ não levam em consideração o LD entre os SNP's e superestimam o parentesco entre os indivíduos. Com isto, pode-se concluir que as matrizes $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$ são as mais adequadas, reafirmando as suas características de maiores capacidades preditivas (Tabela 9). Considerando a natureza dessas duas matrizes, uma explicação de natureza reflexiva pode ser dada pelo fato de essas terem sido empregadas com sucesso em estudos de populações humanas, realidade na qual populações pequenas apresentam elevada heterozigose para a maioria dos locos, assim como no caso de híbridos simples de milho.

Figura 2 – Cluster *heatmaps* dos híbridos formado a partir das matrizes de parentesco de ordem 452 x 452. (a) – matriz $\mathbf{G}_{vr}$; (b) – matriz $\mathbf{G}_{ab}$; (c) – matriz $\mathbf{G}_{yu}$; (d) – matriz $\mathbf{G}_{ya}$; os indivíduos estão representados nas linha e colunas de cada *heatmap*, e cada célula colorida representa o parentesco entre eles; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

4.1.2.3 Coincidência de seleção

Verificou-se que o modelo misto adotado proporcionou uma adequada abordagem para análise conjunta dos grupos de experimentos. Para análise de dados, tanto balanceados quanto desbalanceados, pode-se empregar o modelo mais simples, assumindo a identidade para a matriz de variância e covariância entre os genótipos. Pode-se também ajustar modelos mais realistas, que admitam a presença de parentesco entre estes. Neste caso é possível realizar a análise conjunta dos experimentos sem descartar indivíduos e estabelecer um ordenamento em função de valores genéticos genômicos preditos (BERNARDO, 2010). Assim, a exploração da informação de matrizes de parentesco genômico pode ser essencial para assegurar o correto ordenamento dos indivíduos (MUNOZ et al., 2014).

Neste sentido, visando determinar a coincidência da seleção das melhores linhagens selecionadas desconsiderando o parentesco entre elas (matriz **I**) e da seleção utilizando as matrizes, $\mathbf{G}_{vr}$, $\mathbf{G}_{yu}$, $\mathbf{G}_{ya}$ e $\mathbf{G}_{ab}$ foi estimada a coincidência de Hamblin e Zimmermann (1986) para os 18,75% (seleção das 12 melhores linhagens) com melhor desempenho para produtividade de grãos. Os resultados indicaram coincidência de seleção de 100%, portanto os mesmos indivíduos foram selecionados em todos os cenários (Figura 3). Esses resultados, combinados aos da correlação de *Spearman* (ρ) entre os valores genéticos preditos pelas matrizes de parentesco genômico e as médias fenotípicas ajustadas indicam a importância de métodos mais simples de obtenção da matriz de parentesco realizado, pela correlação quase perfeita apresentada com os valores estimados pelo BLUP fenotípico (Tabela 10).

Quanto ao comportamento das 12 melhores linhagens selecionadas com base na matriz **I** nas matrizes de parentesco $\mathbf{G}_{vr}$, $\mathbf{G}_{yu}$, $\mathbf{G}_{ya}$ e $\mathbf{G}_{ab}$, em geral o ordenamento dos indivíduos foi o mesmo, independentemente da matriz utilizada no modelo no painel de linhagens, sendo dois dos indivíduos mais produtivos os mesmos para todas as matrizes nas diferentes densidades. Notou-se também que a seleção baseada na matriz de Astle e Balding (2009) seria a mais divergente daquela baseada no BLUP fenotípico. Ao passo que o comportamento deste painel predito aquelas propostas por Yang et al. (2010), quanto à produtividade de grãos, foram similares entre si e com as médias ajustadas, desconsiderando parentesco. Este comportamento persistiu ao longo das densidades de marcadores testadas. Houve grande anuência entre o comportamento das linhagens segundo a matriz identidade e a de Vanraden (2008). Com base no que vem sendo discutido a partir dos parâmetros apresentados acima, a

matriz $\mathbf{G}_{vr}$ proporcionaria valores genéticos genômicos preditos mais próximos aos valores de médias ajustadas com base em avaliação fenotípica.

Valores tão elevados podem ser explicados pela característica de predominância de efeitos genéticos aditivos no painel de linhagens, sendo esses efeitos os captados pelas matrizes de parentesco realizado.

Tabela 10 - Correlação de *Spearman* (ρ) e coincidência de seleção para as linhagens nos diferentes cenários de matrizes de parentesco e densidades de marcadores

Matriz de Parentesco Genômico	Densidade de marcadores	Coincidência de seleção	ρ
$\mathbf{G}_{vr}$	400k	100,00	0,99
	60k	100,00	0,99
	9586 [§]	100,00	0,99
	1304 ^{§§}	100,00	0,99
$\mathbf{G}_{ab}$	400k	100,00	0,99
	60k	100,00	0,99
	9586 [§]	100,00	0,99
	1304 ^{§§}	100,00	0,99
$\mathbf{G}_{yu}$	400k	100,00	0,99
	60k	100,00	0,99
	9586 [§]	100,00	0,99
	1304 ^{§§}	100,00	0,99
$\mathbf{G}_{ya}$	400k	100,00	0,99
	60k	100,00	0,99
	9586 [§]	100,00	0,99
	1304 ^{§§}	100,00	0,99

[§] - redução do número de SNP's com base no LD $\cong 0,2$; ^{§§} - redução do número de SNP's com base no LD $\cong 0,02$; ρ = correlação de Spearman; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

Figura 3 – Comportamento das 12 linhagens selecionadas com base na matriz **I** e nas matrizes de parentesco **G_{vr}**, **G_{ab}**, **G_{yu}** e **G_{ya}**. a) 400k; b) 60k; c) LD $\cong$ 0,2; d) LD $\cong$ 0,02; **I** – matriz identidade; **Astle** – matriz de Astle e Balding; **GAWW** – matriz de Vanraden; **GAU** – matriz de parentesco unificado de Yang; **GAUad** – matriz de parentesco unificado ajustado de Yang

A coincidência de Hamblin e Zimmermann (1986) para 2,5% dos híbridos com melhor desempenho por avaliação efetuada não variou entre densidades de marcadores. Os maiores valores de coincidência foram observados para as matrizes propostas por (YANG et al., 2010). O maior valor foi observado para **G_{ya}** quando usados 495 marcadores (LD $\cong$ 0,02), reiterando que maior quantidade de marcadores não significam maiores acurácias. Os valores de correlação de *Spearman* foram inferiores aos observados no painel de linhagens, mas ainda sim, maiores em **G_{yu}** e **G_{ya}**.

Tabela 11 - Correlação de *Spearman* (ρ) e coincidência de seleção para os híbridos nos diferentes cenários de matrizes de parentesco e densidades de marcadores

Matriz de Parentesco Genômico	Densidade de marcadores	Coincidência de seleção	ρ
$\mathbf{G}_{vr}$	50k	65,76	0,95
	4458 [§]	65,76	0,95
	495 ^{§§}	65,76	0,95
$\mathbf{G}_{ab}$	50k	65,76	0,95
	4458 [§]	65,76	0,95
	495 ^{§§}	65,76	0,95
$\mathbf{G}_{yu}$	50k	74,32	0,96
	4458 [§]	74,32	0,96
	495 ^{§§}	74,32	0,95
$\mathbf{G}_{ya}$	50k	74,32	0,96
	4458 [§]	74,32	0,96
	495 ^{§§}	82,88	0,96

[§] - redução do número de SNP's com base no $LD \cong 0,2$; ^{§§} - redução do número de SNP's com base no $LD \cong 0,02$; ρ = correlação de Spearman; $\mathbf{G}_{vr}$ – matriz de Vanraden; $\mathbf{G}_{ab}$ – matriz de Astle e Balding; $\mathbf{G}_{yu}$ – matriz de parentesco unificado de Yang; $\mathbf{G}_{ya}$ – matriz de parentesco unificado ajustado de Yang

Em relação à coincidência de seleção entre os híbridos via matrizes $\mathbf{I}$, $\mathbf{G}_{ab}$ e $\mathbf{G}_{vr}$, esta variou sensivelmente, havendo até mesmo mudança no ordenamento para alguns indivíduos (Figura 4). Para as matrizes $\mathbf{G}_{yu}$ e $\mathbf{G}_{ya}$ a modificação do posicionamento dos indivíduos no *ranking* de produtividade média de grãos foi constantemente observado, em relação à matriz $\mathbf{I}$, porém coincidentes entre si na maioria dos casos, o que nos leva a considera-las mais consistentes e capazes de gerar valores genéticos genômicos preditos fidedignos à realidade (BLUP's fenotípicos) e passíveis de serem incorporados em modelos de seleção genômica. As diferentes densidades de marcas pouco influenciaram no comportamento dos 12 melhores híbridos selecionados com base na matriz $\mathbf{I}$ e demais matrizes.

Alteração no ordenamento de acordo com a matriz foi observada tanto nas linhagens (Figura 3) quanto nos híbridos, sendo mais intensa nesses (Figura 4). Isto reitera a informações já discutidas anteriormente, inclusive a observada nos *cluster heatmaps* (Figura 1 e Figura 2).

Figura 4 – Comportamento dos 12 híbridos selecionados com base na matriz **I** e nas matrizes de parentesco **G_{vr}**, **G_{ab}**, **G_{yu}** e **G_{ya}**. a) 50k; b) $LD \cong 0,2$; c) $LD \cong 0,02$; **I** – matriz identidade; **Astle** – matriz de Astle e Balding; **GAWW** – matriz de Vanraden; **GAU** – matriz de parentesco unificado de Yang; **GAUad** – matriz de parentesco unificado ajustado de Yang

5 CONCLUSÕES

Tanto para híbridos como para linhagens, a densidade de marcas pode ser reduzida significativamente por meio do LD existente entre os SNP's, sem gerar prejuízos quanto a acurácia da GS, mas reduzindo os custos com genotipagem e a demanda computacional.

Para predição genômica das linhagens, a matriz de Vanraden (2008) apresenta o melhor custo benefício, pois tem menor demanda computacional e proporciona resultados satisfatórios quanto à acurácia e coincidência de seleção.

Para a predição genômica de híbridos, as matrizes de Yang *et al.* (2010) são superiores em relação às demais testadas, quanto à acurácia preditiva e coincidência de seleção.

REFERÊNCIAS

- AL ADHAMI, H.; EVANO, B.; LE DIGARCHER, A.; GUEYDAN, C.; DUBOIS, E.; PARRINELLO, H.; DANTEC, C.; BOUSCHET, T.; VARRAULT, A.; JOURNOT, L. A systems-level approach to parental genomic imprinting: the imprinted gene network includes extracellular matrix genes and regulates cell cycle exit and differentiation. **Genome Research**, New York, v. 25, n. 3, p. 353-367, 2015.
- ASTLE, W.; BALDING, D.J. Population structure and cryptic relatedness in genetic association studies. **Statistical Science**, Hayward, v. 24, n. 4, p. 451-471, 2009.
- BAUER, A.M.; REETZ, T.C.; LÉON, J. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. **Crop Science**, Madison, v. 46, n. 6, p. 2685-2691, 2006.
- BERGFELDER-DRÜING, S.; GROSSE-BRINKHAUS, C.; LIND, B.; ERBE, M.; SCHELLANDER, K.; SIMIANER, H.; THOLEN, E. A genome-wide association study in large white and landrace pig populations for number piglets born alive. **PLoS One**, San Francisco, v. 10, n. 3, p. 1-19, 2015.
- BERNARDO, R. Genetic models for predicting maize single-cross performance in unbalanced yield trial data. **Crop Science**, Madison, v. 35, n. 1, p. 141-147, 1995.
- _____. Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. **Crop Science**, Madison, v. 36, n. 4, p. 862-866, 1996.
- _____. **Breeding for quantitative traits in plants**. 2nd ed. Woodbury: Stemma Press, Woodbury, 2010. 369 p.
- BERNARDO, R.; YU, J. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, Madison, v. 47, n. 3, p. 1082-1090, 2007.
- BORÉM, A.; GALVÃO, J.C.C.; PIMENTEL, M.A. **Milho: do plantio à colheita**. Viçosa: Ed. UFV, 2015. 351 p.
- BUTLER, D.; CULLIS, B.; GILMOUR, A.; GOGEL, B. **ASReml-R: reference manual**. Brisbane: Queensland Department of Primary Industries, 2009. 160 p.
- CAVALCANTI, J.J.V.; RESENDE, M. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, Cruz das Almas, v. 34, p. 840-846, 2012.
- CHEN, C.-Y.; MISZTAL, I.; AGUILAR, I.; LEGARRA, A.; MUIR, W. Effect of different genomic relationship matrices on accuracy and scale. **Journal of Animal Science**, Champaign, v. 89, n. 9, p. 2673-2679, 2011.
- CHESSA, S.; CHIATTI, F.; CERIOTTI, G.; CAROLI, A.; CONSOLANDI, C.; PAGNACCO, G.; CASTIGLIONI, B. Development of a single nucleotide polymorphism genotyping microarray platform for the identification of bovine milk protein genetic polymorphisms. **Journal of Dairy Science**, Champaign, v. 90, n. 1, p. 451-464, 2007.

CLARK, S.A.; VAN DER WERF, J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: GONDRO, C.; VAN DER WERF, J.; HAYES, B. **Genome-wide association studies and genomic prediction**. New York: Humana Press, 2013. p. 321-330.

COOK, J.P.; MCMULLEN, M.D.; HOLLAND, J.B.; TIAN, F.; BRADBURY, P.; ROSS-IBARRA, J.; BUCKLER, E.S.; FLINT-GARCIA, S.A. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. **Plant Physiology**, Bethesda, v. 158, n. 2, p. 824-834, 2012.

CROSSA, J.; BEYENE, Y.; KASSA, S.; PÉREZ, P.; HICKEY, J. M.; CHEN, C.; DE LOS CAMPOS, G.; BURGUEÑO, J.; WINDHAUSEN, V. S.; BUCKLER, E. Genomic prediction in maize breeding populations with genotyping-by-sequencing. **G3: Genes| Genomes| Genetics**, Chicago, v. 3, n. 11, p. 1903-1926, 2013.

CRUZ, C.D. **Modelos biométricos aplicados ao melhoramento genético**. 4. ed. Viçosa: Universidade Federal de Viçosa, 2012. Disponível em: <<http://books.google.com.br/books?id=4uryygAACAAJ>>. Acesso em: 15 mar. 2016.

DA MOTTA, P.E.F.; CARVALHO FILHO, A. de; KER, J.C.; PEREIRA, N.R.; CARVALHO JUNIOR, W. de; BLANCANEAU, P. Relações solo-superfície geomórfica e evolução da paisagem em uma área do Planalto Central Brasileiro. **Pesquisa Agropecuária Brasileira**, Brasília, v. 37, n. 6, p. 869-878, 2002.

ENDELMAN, J.B.; JANNINK, J.-L. Shrinkage estimation of the realized relationship matrix. **G3: Genes| Genomes| Genetics**, Chicago, v. 2, n. 11, p. 1405-1413, 2012.

FALCÃO, A.J.S.; MARTINS, E.N.; COSTA, C.N.; MAZUCHELI, J. Efeitos do número de animais na matriz de parentesco sobre as estimativas de componentes de variância para produção de leite usando os métodos de máxima verossimilhança restrita e bayesiano. **Revista Brasileira de Zootecnia**, Viçosa, v. 38, p. 1468-1477, 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-35982009000800011&nrm=iso>. Acesso em: 15 abr. 2016.

FRITSCHÉ-NETO, R.; GONÇALVES, M.C.; VENCOSKY, R.; SOUZA-JÚNIOR, C. Prediction of genotypic values of maize hybrids in unbalanced experiments. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 10, n. 1, p. 32-39, 2010.

GANAL, M.W.; DURSTEWITZ, G.; POLLEY, A.; BÉRARD, A.; BUCKLER, E.S.; CHARCOSSET, A.; CLARKE, J.D.; GRANER, E.-M.; HANSEN, M.; JOETS, J. A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. **PloS One**, San Francisco, v. 6, n. 12, p. e28334, 2011.

GORE, M.A.; CHIA, J.-M.; ELSHIRE, R.J.; SUN, Q.; ERSOZ, E.S.; HURWITZ, B.L.; PEIFFER, J.A.; MCMULLEN, M.D.; GRILLS, G.S.; ROSS-IBARRA, J. A first-generation haplotype map of maize. **Science**, Washington, v. 326, n. 5956, p. 1115-1117, 2009.

GRANATO, Í.S.C.; FRITSCHÉ-NETO, R.F. **Package 'snpReady'**. Piracicaba: ESALQ, Departamento de Genética, 2016. 160 p.

- GRATTAPAGLIA, D.; RESENDE, M.D.V. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, Berlin, v. 7, n. 2, p. 241-255, 2011.
- HABIER, D.; FERNANDO, R.; DEKKERS, J. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, Austin, v. 177, n. 4, p. 2389-2397, 2007.
- HALEY, C.; VISSCHER, P. Strategies to utilize marker-quantitative trait loci associations. **Journal of Dairy Science**, Champaign, v. 81, p. 85-97, 1998.
- HAMBLIN, J.; ZIMMERMANN, M. J.D.O. Breeding common bean for yield in mixtures. **Plant Breeding Reviews**, Westport, v. 4, p. 245-272, 1986.
- HAMBLIN, M.T.; BUCKLER, E.S.; JANNINK, J.-L. Population genetics of genomics-based crop improvement methods. **Trends in Genetics**, London, v. 27, n. 3, p. 98-106, 2011.
- HARTL, D.L.; CLARK, A.G.; CLARK, A.G. **Principles of population genetics**. Sunderland: Sinauer Associates, 1997. 565 p.
- HAYES, B.; GODDARD, M. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, n. 4, p. 1819-1829, 2001.
- HEATON, M.P.; KEEN, J.E.; CLAWSON, M.L.; HARHAY, G.P.; BAUER, N.; SHULTZ, C.; GREEN, B.T.; DURSO, L.; CHITKO-MCKOWN, C.G.; LAEGREID, W.W. Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. **Journal of the American Veterinary Medical Association**, Schaumburg, v. 226, n. 8, p. 1311-1314, 2005.
- HENDERSON, C. Estimation of changes in herd environment. **Journal of Dairy Science**, Champaign, v. 32, n. 8, p. 706, 1949.
- HIGA, R.H.; ALMEIDA REGITANO, L.C. de; IBELLI, A.M.G.; SANTOS, I.L.N. dos. **Estudo de algoritmos de biclustering para a análise de expressão gênica utilizando a tecnologia de microarranjo**. Campinas: Embrapa Informática Agropecuária, 2010. 45 p.
- HILL, W.; WEIR, B. Variances and covariances of squared linkage disequilibria in finite populations. **Theoretical Population Biology**, New York, v. 33, n. 1, p. 54-78, 1988. I
- HOLLAND, J.B.; NYQUIST, W.E.; CERVANTES-MARTÍNEZ, C.T. Estimating and interpreting heritability for plant breeding: an update. **Plant Breeding Reviews**, Westport, v. 22, p. 9-112, 2003.
- HOLLOWAY, J.W.; BEGHÉ, B.; TURNER, S.; HINKS, L.J.; DAY, I.N.; HOWELL, W.M. Comparison of three methods for single nucleotide polymorphism typing for DNA bank studies: sequence-specific oligonucleotide probe hybridisation, TaqMan liquid phase hybridisation, and microplate array diagonal gel electrophoresis (MADGE). **Human Mutation**, New York, v. 14, n. 4, p. 340, 1999.

JANNINK, J.-L.; WALSH, B. Association mapping in plant populations. In: KANG, M.S.C. (Ed.). **Quantitative genetics, genomics and plant breeding**. Wallingford: CABI Publ., 2002. p. 59-68.

JANSS, L.; DE LOS CAMPOS, G.; SHEEHAN, N.; SORENSEN, D. Inferences from genomic models in stratified populations. **Genetics**, Austin, v. 192, n. 2, p. 693-704, 2012.

JOHNSON, D.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of Dairy Science**, Champaign, v. 78, n. 2, p. 449-456, 1995.

KNUEPPEL, S.; ROHDE, K.; KNUEPPEL, M.S.; LAZYLOAD, T. **Package 'HapEstXXR'-Multi-Locus Stepwise Regression (MSR)**. 2015. 23 p.

LI, C.; HORVITZ, D. Some methods of estimating the inbreeding coefficient. **American Journal of Human Genetics**, Chicago, v. 5, n. 2, p. 107, 1953.

LIAN, L.; JACOBSON, A.; ZHONG, S.; BERNARDO, R. Genomewide prediction accuracy within 969 maize biparental populations. **Crop Science**, Madison, v. 54, n. 4, p. 1514-1522, 2014.

LIMA, B.M.D. **Bridging genomics and quantitative genetics of *Eucalyptus*: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data**. 2014. 92 p. Thesis (Doctor in Plant Genetics and Breeding) - "Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba, 2014.

MALECOT, G. **Mathematics of heredity**. Paris : Masson etCie., 1948. 63 p.

MANOLIO, T.A.; COLLINS, F.S.; COX, N.J.; GOLDSTEIN, D.B.; HINDORFF, L.A.; HUNTER, D.J.; MCCARTHY, M.I.; RAMOS, E.M.; CARDON, L.R.; CHAKRAVARTI, A. Finding the missing heritability of complex diseases. **Nature**, London, v. 461, n. 7265, p. 747-753, 2009.

MENEZES, G.R.D.O.; TORRES, R.D.A.; TORRES JÚNIOR, R.A.D.A.; SILVA, L.O.C.D.; GONDO, A.; EUCLYDES, R.F. Estimation of genetic parameters for growth traits in Tabapuã cattle using a multi-trait model. **Revista Brasileira de Zootecnia**, Viçosa, v. 42, n. 8, p. 570-574, 2013.

MEUWISSEN, T.; GODDARD, M.E. Prediction of identity by descent probabilities from marker-haplotypes. **Genetics Selection Evolution**, Paris, v. 33, n. 6, p. 605-634, 2001.

MEYER, K. Estimation of genetic parameters. In: HILL, W.G.; MACKAY, T.F.C. **Reviews on molecular and quantitative genetic approaches in honor of Alan Robertson**. Wallingford: CAB International, 1989. p. 159-167.

MUNOZ, P.R.; RESENDE, M.F.; HUBER, D.A.; QUESADA, T.; RESENDE, M.D.; NEALE, D.B.; WEGRZYN, J.L.; KIRST, M.; PETER, G.F. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. **Crop Science**, Madison, v. 54, n. 3, p. 1115-1123, 2014.

MURRAY, M.; THOMPSON, W.F. Rapid isolation of high molecular weight plant DNA. **Nucleic Acids Research**, Oxford, v. 8, n. 19, p. 4321-4326, 1980.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, Champaign, v. 75, n. 7, p. 1738-1745, 1997.

NORDBORG, M. Coalescent theory. In: BALDING, D.J.; BISHOP, M.; CANNING, C. (Ed.). **Handbook of statistical genetics**. New Jersey: Wiley, 2001. p. 769-774.

NUNES, J.A.R.; MORETO, A.L.; RAMALHO, M.A P. Using genealogy to improve selection efficiency of pedigree method. **Scientia Agricola**, Piracicaba, v. 65, n. 1, p. 25-30, 2008.

OLIEHOEK, P.A.; WINDIG, J.J.; VAN ARENDONK, J.A.; BIJMA, P. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. **Genetics**, Austin, v. 173, n. 1, p. 483-496, 2006.

PEIFFER, J.A.; ROMAY, M.C.; GORE, M.A.; FLINT-GARCIA, S.A.; ZHANG, Z.; MILLARD, M.J.; GARDNER, C.A.; MCMULLEN, M.D.; HOLLAND, J B.; BRADBURY, P.J. The genetic architecture of maize height. **Genetics**, Austin, v. 196, n. 4, p. 1337-1356, 2014.

PIEPHO, H.-P.; MÖHRING, J.; MELCHINGER, A.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, Wageningen, v. 161, n. 1/2, p. 209-228, 2008.

PIMENTEL-GOMES, F. **Course of experimental statistics**. Piracicaba: Nobel, 2000.

POWELL, J.E.; VISSCHER, P.M.; GODDARD, M.E. Reconciling the analysis of IBD and IBS in complex trait studies. **Nature Reviews Genetics**, London, v. 11, n. 11, p. 800-805, 2010.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M.A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P.I.; DALY, M.J. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, Chicago, v. 81, n. 3, p. 559-575, 2007.

QUAAS, R. Additive genetic model with groups and relationships. **Journal of Dairy Science**, Champaign, v. 71, n. 5, p. 1338-1345, 1988.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2014.

RESENDE, M.D.V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 2008. 101 p.

RESENDE, M.D.V.; DUARTE, J.B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, Goiânia, v. 37, n. 3, p. 182-194, 2007.

RESENDE, M.D.V.; LOPES, P.S.; SILVA, R.L. da; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n. 56, p. 63, 2008.

RESENDE, M.D.V. SILVA, F.; LOPES, P.; AZEVEDO, C. Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial. Viçosa: UFV, 2012a. 291 p.

RESENDE, M.F.; MUÑOZ, P.; RESENDE, M.D.V.; GARRICK, D.J.; FERNANDO, R.L.; DAVIS, J.M.; JOKELA, E.J.; MARTIN, T.A.; PETER, G.F.; KIRST, M. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). **Genetics**, Austin, v. 190, n. 4, p. 1503-1510, 2012b.

RESENDE, M.D.V.; RESENDE, M.F.; SANSALONI, C.P.; PETROLI, C.D.; MISSIAGGIA, A.A.; AGUIAR, A.M.; ABAD, J.M.; TAKAHASHI, E.K.; ROSADO, A.M.; FARIA, D.A. Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, Cambridge, v. 194, n. 1, p. 116-128, 2012c.

RIEDELSEIMER, C.; CZEDIK-EYSENBERG, A.; GRIEDER, C.; LISEC, J.; TECHNOW, F.; SULPICE, R.; ALTMANN, T.; STITT, M.; WILLMITZER, L.; MELCHINGER, A.E. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. **Nature Genetics**, New York, v. 44, n. 2, p. 217-220, 2012.

RINCENT, R. **Optimization of association genetics and genomic selection strategies for populations of different diversity levels: application in maize (*Zea mays* L.)**. Paris: AgroParisTech, 2014. 164 p.

RITLAND, K. Estimators for pairwise relatedness and individual inbreeding coefficients. **Genetical Research**, London, v. 67, n. 2, p. 175-185, 1996.

ROSENBERG, N.A.; FELDMAN, M.W. **The relationship between coalescence times and population divergence times**. Oxford: Oxford University Press, 2002. 46 p.

ROUSSET, F. Inbreeding and relatedness coefficients: what do they measure? **Heredity**, London, v. 88, n. 5, p. 371-380, 2002.

SANT'ANA, G.C. **Relação da estrutura genética com a eficiência no uso de nitrogênio e com a acurácia de GWAS em milho tropical**. 2015. 66 p. Tese (D.S. em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa, 2015.

SCHAEFFER, L. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, Berlin, v. 123, n. 4, p. 218-223, 2006.

SEMAGN, K.; MAGOROKOSHO, C.; VIVEK, B.S.; MAKUMBI, D.; BEYENE, Y.; MUGO, S.; PRASANNA, B.; WARBURTON, M.L. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. **BMC Genomics**, New York, v. 13, n. 1, p. 113, 2012.

SONG, Q.; HYTEN, D.L.; JIA, G.; QUIGLEY, C.V.; FICKUS, E.W.; NELSON, R.L.; CREGAN, P.B. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. **PLoS One**, San Francisco, v. 8, n. 1, p. e54985, 2013.

STICH, B. Comparison of mixed-model approaches for association mapping. **Genetics**, Austin, v. 178, n. 3, p. 1745-1754, 2008.

_____. Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. **Genetics**, Austin, v. 183, n. 4, p. 1525-1534, 2009.

SUBEDI, S.; FENG, Z.; DEARDON, R.; SCHENKEL, F. SNP selection for predicting a quantitative trait. **Journal of Applied Statistics**, Abingdon, v. 40, n. 3, p. 600-613, 2013.

THOMPSON, E A. **Human evolutionary trees**. Cambridge: CUP Archive, 1975. 158 p.

TIAN, F.; BRADBURY, P.J.; BROWN, P.J.; HUNG, H.; SUN, Q.; FLINT-GARCIA, S.; ROCHEFORD, T.R.; MCMULLEN, M.D.; HOLLAND, J.B.; BUCKLER, E.S. Genome-wide association study of leaf architecture in the maize nested association mapping population. **Nature Genetics**, New York, v. 43, n. 2, p. 159-162, 2011.

TORO, M.; BARRAGAN, C.; OVILO, C. Estimation of genetic variability of the founder population in a conservation scheme using microsatellites. **Animal Genetics**, Oxford, v. 34, n. 3, p. 226-228, 2003.

UNTERSEER, S.; BAUER, E.; HABERER, G.; SEIDEL, M.; KNAAK, C.; OUZUNOVA, M.; MEITINGER, T.; STROM, T.M.; FRIES, R.; PAUSCH, H. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. **BMC Genomics**, New York, v. 15, n. 1, p. 1, 2014.

VAN HEERWAARDEN, J.; DOEBLEY, J.; BRIGGS, W.H.; GLAUBITZ, J.C.; GOODMAN, M.M.; GONZALEZ, J.D.J.S.; ROSS-IBARRA, J. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. **Proceedings of the National Academy of Sciences of the USA**, Washington, v. 108, n. 3, p. 1088-1092, 2011.

VANRADEN, P. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, Champaign, v. 91, n. 11, p. 4414-4423, 2008.

WEIR, B.S.; ANDERSON, A.D.; HEPLER, A.B. Genetic relatedness analysis: modern data and new challenges. **Nature Reviews Genetics**, London, v. 7, n. 10, p. 771-780, 2006.

WIMMER, V.; ALBRECHT, T.; AUINGER, H.-J.; SCHÖN, C.-C. Synbreed: a framework for the analysis of genomic prediction data using R. **Bioinformatics**, Cambridge, v. 28, n. 15, p. 2086-2087, 2012.

WRIGHT, S. Systems of mating. I. The biometric relations between parent and offspring. **Genetics**, Austin, v. 6, n. 2, p. 111, 1921.

_____. Coefficients of inbreeding and relationship. **American Naturalist**, Chicago, v. 56, n. 645, p. 330-338, 1922.

_____. Genetical structure of populations. **Nature**, London, v. 166, p. 247-249, 1950.

YANG, J.; BENYAMIN, B.; MCEVOY, B.P.; GORDON, S.; HENDERS, A.K.; NYHOLT, D.R.; MADDEN, P.A.; HEATH, A.C.; MARTIN, N.G.; MONTGOMERY, G.W. Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics**, New York, v. 42, n. 7, p. 565-569, 2010.

YU, J.; ZHANG, Z.; ZHU, C.; TABANAO, D.A.; PRESSOIR, G.; TUINSTRA, M.R.; KRESOVICH, S.; TODHUNTER, R.J.; BUCKLER, E.S. Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. **The Plant Genome**, Madison, v. 2, n. 1, p. 63-77, 2009.

ZHANG, X.; PÉREZ-RODRÍGUEZ, P.; SEMAGN, K.; BEYENE, Y.; BABU, R.; LÓPEZ-CRUZ, M.; SAN VICENTE, F.; OLSEN, M.; BUCKLER, E.; JANNINK, J.-L. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. **Heredity**, London, v. 114, n. 3, p. 291-299, 2015.

ZHANG, Z.; GUILLAUME, F.; SARTELET, A.; CHARLIER, C.; GEORGES, M.; FARNIR, F.; DRUET, T. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. **Bioinformatics**, Cambridge, v. 28, n. 19, p. 2467-2473, 2012.

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

ZILA, C.T.; OGUT, F.; ROMAY, M.C.; GARDNER, C.A.; BUCKLER, E.S.; HOLLAND, J.B. Genome-wide association study of *Fusarium* ear rot disease in the USA maize inbred line collection. **BMC Plant Biology**, New York, v. 14, n. 1, p. 372, 2014.