BRÁULIO FABIANO XAVIER DE MORAES

POPULAÇÃO DE ESTIMAÇÃO PARA PREDIÇÃO GENÔMICA EM *Eucalyptus spp*

LAVRAS – MG

2018

BRÁULIO FABIANO XAVIER DE MORAES

ESTRATÉGIAS PARA OTIMIZAR A POPULAÇÃO DE ESTIMAÇÃO PARA PREDIÇÃO GENÔMICA EM *Eucalyptus spp*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Profa. Dra. Flávia Maria Avelar Gonçalves

Orientadora

LAVRAS – MG

2018

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Moraes, Bráulio Fabiano Xavier de.
Estratégias para otimizar a população de estimação para predição genômica em *Eucalyptus spp* / Bráulio Fabiano Xavier de Moraes. – 2017.
49 p. : il.

Orientadora: Flávia Maria Avelar Gonçalves.
Tese (doutorado) - Universidade Federal de Lavras,2017.
Bibliografia.

1. Seleção genômica. 2. Estrutura de população. 3. Marcadores moleculares. I. Gonçalves, Flávia Maria Avelar. II. Título.

BRÁULIO FABIANO XAVIER DE MORAES

ESTRATÉGIAS PARA OTIMIZAR A POPULAÇÃO DE ESTIMAÇÃO PARA PREDIÇÃO GENÔMICA EM *Eucalyptus spp*

STRATEGIES TO OPTIMIZE THE ESTIMATION POPULATION FOR GENOMIC PREDICTION IN *Eucalyptus spp*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

APROVADA em 16 de Outubro de 2017.

Dr. Aurélio Mendes Aguiar	FIBRIA SA.
Dr. Bruno Marco de Lima	FIBRIA SA.
Dr. Heyder Diniz Silva	
Prof. Dr. Magno Antônio Patto Ramalho	UFLA

Profa. Dra. Flávia Maria Avelar Gonçalves
Orientadora

LAVRAS – MG

2018

Aos meus pais, Delcrécio e Márcia, ao meu irmão Breno

e à minha esposa Evelyn

DEDICO.

AGRADECIMENTOS

Agradeço a Deus pela vida, saúde, proteção e por me dar forças nas horas difíceis.

À Universidade Federal de Lavras, ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas e à FIBRIA pela oportunidade de realizar este trabalho.

Ao CNPq pela concessão da bolsa de estudos e à CAPES pela bolsa de doutorado sanduíche nos EUA.

À minha orientadora, professora Flávia Maria Avelar Gonçalves, pela orientação, confiança depositada, ensinamentos e principalmente pela grande amizade.

À minha família pelo carinho, apoio e incentivo.

À minha esposa, Evelyn, pelo companheirismo, felicidade e apoio em todos os momentos.

Aos colegas e amigos da Genética, pelo companheirismo.

MUITO OBRIGADO!

RESUMO

No melhoramento de plantas, a genotipagem em larga escala tem tido um grande interesse na busca por maiores ganhos genéticos com o uso da seleção genômica. A capacidade preditiva é o parâmetro que permite verificar o sucesso da seleção genômica nos programas de melhoramento. O tamanho e o relacionamento entre indivíduos da população em estudo, além da estrutura genética das populações de treinamento e validação, afetam as predições da seleção genômica. Diante do exposto, realizou-se este trabalho com o objetivo de otimizar o tamanho da população de treinamento e verificar impacto da estrutura de populacional entre as populações de treinamento e validação para obtenção das predições genômicas. Foi utilizada uma população de melhoramento de eucalipto composta por 860 indivíduos, avaliados fenotipicamente para 13 características aos 24 e 36 meses de idade e genotipados por meio do EUChip60k. A capacidade preditiva da seleção genômica foi obtida por meio da correlação entre os valores paramétricos e os valores estimados por RR-BLUP. O efeito do tamanho da população de treinamento foi verificado pela partição em subgrupos variando de 50 a 800 indivíduos. As populações de treinamento foram agrupadas pela análise de componentes principais e abordagem bayesiana (STRUCTURE) para verificar o efeito da estrutura populacional nas predições genômicas. O rápido decaimento do desequilíbrio de ligação com a correção para a estrutura populacional indicou forte efeito da estrutura na população em estudo. O aumento nas predições do modelo genômico foi pequeno e não significativo em subgrupos a partir de 300 indivíduos. As capacidades preditivas, ao remover o relacionamento entre indivíduos, foram drasticamente reduzidas, levando até obtenção de estimativas negativas. Foi verificado aumento na média das predições, para todas características avaliadas para o agrupamento de famílias pela análise de componentes principais e pelo agrupamento bayesiano em 3% e 9%, respectivamente. Um maior número de indivíduos na população de treinamento leva à maior capacidade de predição. A estrutura de população apresenta um papel importante, para se otimizar as populações de treinamento, proporcionando maior eficiência dos modelos de predição, quando há indivíduos relacionados nas populações de estimação e validação.

Palavras-chave: Seleção genômica. Estrutura de população. Marcadores moleculares. Capacidade preditiva.

ABSTRACT

Large-scale genotyping has been of great interest in the search for greater genetic gains by genomic selection use. The predictive capacity is the parameter that allows to verify the success of genomic selection in breeding programs. The size and relationship among individuals of the study population, as well as the genetic structure of the training and validation populations, affect the predictions of genomic selection. Based on the above, this work was carried out with the objective of optimizing the training population size and verify the impact of the population structure among the training and validation populations to obtain the genomic predictions. We evaluated an *Eucalyptus* breeding population, consisting of 860 individuals, phenotypically evaluated for 13 traits at 24 and 36 months old and genotyped by EUChip60k. The predictive capacity of genomic selection was obtained through the correlation between the parametric values and the values estimated by RR-BLUP. The effect of training population size was verified by partitioning into subgroups ranging from 50 to 800 individuals. The training populations were grouped by Principal Component Analysis (PCA) and Bayesian approach (STRUCTURE) to verify the effect of population structure on genomic predictions. The rapid decay of the linkage disequilibrium with the correction for the population structure indicated a strong effect of the structure in the study population. The increase in predictions of the genomic model was small and not significant in subgroups with more than 300 individuals. Predictive abilities in removing relationships between individuals were drastically reduced, leading to negative estimates. An increase in the predictions mean was verified for all evaluated characteristics for the grouping of families by analysis of main components and Bayesian grouping in 3% and 9%, respectively. A greater number of individuals in the training population leads to greater predictive capacity. Population structure has an important role to optimize training populations, providing greater efficiency of prediction models when there are related individuals in the estimation and validation populations.

Keywords: Genomic selection. Population structure. Molecular markers. Predictive Capability.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A genotipagem em larga escala do genoma de uma dada espécie tem se tornado mais acessível, uma vez que houve um grande progresso da nova geração de sequenciamento, e a redução dos custos tem levado a uma maior disponibilidade de plataformas existentes (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015).

As altas resoluções obtidas pela genotipagem de *Single Nucleotide Polimorfism* (Polimorfismo de único nucleotídeo – SNP), por meio das diversas técnicas existentes, facilitam as associações entre os SNPs e os fenótipos. Tanto no melhoramento de animais como no de plantas, a genotipagem em larga escala tem tido um grande interesse na busca por maiores ganhos genéticos com o uso da seleção genômica. A capacidade preditiva é o parâmetro que permite verificar o sucesso da seleção genômica nos programas de melhoramento (MEUWISSEN; HAYES; GODDARD, 2001).

Dentre os fatores que afetam as predições da seleção genômica (GS), alguns elencados estão o tipo e número de marcadores empregados na genotipagem (POLAND; RIFE, 2012; SCHAEFFER, 2006), o desequilíbrio de ligação (HABIER; FERNANDO; DEKKERS, 2007), o tamanho da população e o relacionamento entre indivíduos desta população (PSZCZOLA et al., 2012), além da estrutura genética das populações de treinamento e validação (ISIDRO et al., 2015).

Nos programas de melhoramento de *Eucalyptus* L'Her. (*Myrtaceae*), é esperado que a seleção genômica leve a maiores ganhos genéticos por unidade de tempo, principalmente, para características que se expressam tardiamente (RESENDE, M. F. et al., 2012). Recentemente, em diversos estudos foi explorado o potencial da seleção genômica em diferentes espécies de eucalipto (GRATTAPAGLIA et al., 2011; GRATTAPAGLIA; RESENDE, 2011; LIMA, 2014; RESENDE, M. D. et al., 2012; SANSALONI et al., 2010). No entanto, estudos mais aprofundados das populações que compõem o conjunto de estimação dos efeitos dos marcadores são escassos e o impacto do tamanho e estrutura genética das populações e o relacionamento entre os indivíduos podem apresentar grande impacto sobre as predições da seleção genômica.

Desse modo, o objetivo neste estudo foi otimizar o tamanho da população de treinamento da seleção genômica e verificar impacto da estrutura populacional entre as populações de treinamento e validação para seleção genômica de uma população de melhoramento de *Eucalyptus* spp.

2 REFERENCIAL TEÓRICO

2.1 Marcadores moleculares

Novas técnicas de identificação de polimorfismos de único nucleotídeo (*Single Nucleotyde Polimorfism* - SNPs) foram desenvolvidas na última década, o que fez com que fosse possível identificar um maior número de marcas no genoma das espécies de interesse. Com um número maior de marcadores moleculares disponíveis, tem sido possível mapear, cada vez com maior precisão, as populações segregantes, possibilitando assim ter aplicações mais complexas das informações oriundas dos marcadores (RESENDE et al., 2008).

Para que um marcador molecular seja considerado ideal a ser empregado, no melhoramento de plantas, levando em conta sua aplicação e a espécie envolvida, ele precisa apresentar distribuição uniforme em todo o genoma (não serem agrupados em certas regiões), características alélicas distintas (facilidade na identificação dos alelos), única cópia do marcador (sem efeito pleiotrópico), baixo custo (bom custo benefício no desenvolvimento da técnica e na genotipagem), fácil detecção e análise, alta disponibilidade, apresentar alta repetibilidade e reprodutibilidade, especificidade para o genoma em estudo, especialmente, para poliploides e não apresentar nenhum efeito prejudicial à informação fenotípica (JIANG, 2013).

As técnicas de genotipagem de alto rendimento (*high-throughput genotyping*) permitem a redução dos custos do uso de marcadores por unidade de informação e consequente por indivíduo avaliado (RESENDE, 2008). Isto permite que um grande número de indivíduos possa ser genotipado a um preço acessível. Desta maneira, populações inteiras oriundas de diferentes famílias podem ser genotipadas simultaneamente.

Outro fator que contribuiu para que a genotipagem dos SNPs se tornasse mais acessível foi o progresso da nova geração de sequenciamento (*Next Generation Sequencing – NGS*), e a redução dos custos tem levado à uma maior disponibilidade de plataformas existentes (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015).

O estabelecimento de um “chip” para genotipagem foi primeiramente descrito por Dong et al. (2001). Atualmente, os “chips” de genotipagem têm sido desenvolvidos para diversas espécies animais (HARRIS; JOHNSON, 2010; RAMOS et al., 2009) e plantas (GANAL et al., 2011; SONG et al., 2013; UNTERSEER et al., 2014). Em algumas espécies, os chips contendo maior número de SNPs estão sendo desenvolvidos, a fim de atender as

necessidades de aplicações de cada setor, oferecendo alto desempenho na genotipagem para aplicações da informação do genoma completo de cada espécie (KILIAN et al., 2003).

Dentre as vantagens do uso do SNP-Chip, estão o baixo custo de aplicação, alta densidade de marcas e de polimorfismo, velocidade da geração dos dados, alta acurácia e reprodutibilidade (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). No entanto, estas vantagens são observadas, após o estabelecimento do chip, uma vez que o seu processo de desenvolvimento é laborioso e oneroso, pois deve ser feito de maneira criteriosa, já que envolve diversas etapas e o tempo de desenvolvimento é elevado.

Em análises rotineiras, o uso de SNP-Chips tem grande vantagem sobre outros métodos de genotipagem, visto que os dados obtidos são de alta reprodutibilidade dentro e entre laboratórios e que os chips podem ser armazenados e reutilizados, que os marcadores serão sempre os mesmos a serem utilizados (GANAL et al., 2012). Desse modo, este método de genotipagem apresenta grande vantagem na aplicação da seleção genômica que preconiza o uso de plataformas robustas e com alta acurácia para genotipagem de um número elevado de marcas em muitos indivíduos.

2.2 Seleção genômica

A genotipagem de alto rendimento tem revolucionado as análises genéticas em diversas espécies nos últimos anos (UNTERSEER et al., 2014). As altas resoluções obtidas pela genotipagem de SNPs, por meio das técnicas de genotipagem em larga escala, proporcionam as associações mais acuradas entre os SNPs e os fenótipos dos indivíduos avaliados.

A GS é um tipo de seleção, baseada nos marcadores moleculares, proposta por Meuwissen, Hayes e Goddard (2001), por meio da qual se podem selecionar indivíduos com características de interesse com base na predição, por intermédio de milhares de marcas que cobrem todo o genoma de maneira densa, pois se espera que todos os genes estejam em desequilíbrio de ligação (LD) com pelo menos um marcador (RESENDE et al., 2008). Uma vez que o LD é relativamente pequeno em muitas espécies, a seleção genômica requer um grande número de marcas, já que quanto menor o DL maior será a chance de haver recombinantes e, assim, quando se tem uso de poucas marcas moleculares, podem não estar ligadas ao gene de interesse (HAMBLIN; BUCKLER; JANNINK, 2011). Desse modo, o uso de marcadores com alta densidade é uma das características fundamentas da GS.

A seleção genômica descarta a necessidade de se ter associações entre o QTL e o marcador. A seleção, nesse caso, é realizada com base na estimação dos valores genômicos, o GEBV (*genomic estimated breeding value*). Primeiramente faz-se necessário o desenvolvimento de modelos preditivos a fim de predizer os GEBVs (NAKAYA; ISOBE, 2012). O processo de desenvolvimento do modelo é chamado de fase de treinamento, em que indivíduos de uma população de treinamento são fenotipados e genotipados, para predizer relações entre a informação fenotípica e genotípica, por meio de modelos estatísticos. Uma vez realizado o treinamento e elaboração do modelo genômico, ele será empregado para a seleção dos indivíduos que foram somente genotipados os quais terão os GEBVs estimados, para seleção dos indivíduos desejáveis na fase de melhoramento, utilizando apenas o genótipo e sua predição fenotípica (MEUWISSEN; HAYES; GODDARD, 2001).

As capacidades preditivas dos modelos de seleção genômica têm sido determinadas, por meio da correlação entre as estimativas dos valores genéticos genômicos estimados (GEBVs) e os valores genéticos verdadeiros (TBVs – *true breeding values*), obtidos pela avaliação fenotípica dos indivíduos testados (MEUWISSEN; HAYES; GODDARD, 2001). A validação cruzada tem sido empregada, para obter os valores de capacidade de predição e/ou acurácia dos modelos preditivos, utilizando os dados fenotípicos e genotípicos de uma população de treinamento e os dados genotípicos da população de validação.

Na validação cruzada, é empregado um conjunto de dados que é subdividido aleatoriamente em grupos para formar a população de treinamento e a população de validação (CROSSA et al., 2010; HEFFNER; SORRELLS; JANNINK, 2009; JANNINK; LORENZ; IWATA, 2010; MASSMAN et al., 2013). Assim, um subconjunto da população de validação é retirado do modelo de seleção genômica (LEE et al., 2008).

A capacidade preditiva é o parâmetro que permite verificar o sucesso da seleção genômica nos programas de melhoramento. Podemos, então, citar como fatores importantes que afetam as predições da seleção genômica: o número de marcas e o tipo de marcadores empregados na genotipagem (HESLOT et al., 2012; POLAND; RIFE, 2012; SCHAEFFER, 2006), a herdabilidade da característica avaliada (HEFFNER; SORRELLS; JANNINK, 2009; LORENZ, 2013), o desequilíbrio de ligação (HABIER; FERNANDO; DEKKERS, 2007), o modelo estatístico para predição (HESLOT et al., 2012), tamanho efetivo da população (DAETWYLER; VILLANUEVA; WOOLLIAMS, 2008), a relação genética entre as populações de treinamento e validação (ALBRECHT et al., 2011; PSZCZOLA et al., 2012) e a estrutura genética das populações em estudo (GUO et al., 2013; ISIDRO et al., 2015).

Por sua vez, a capacidade da estimação dos efeitos dos SNPs podem sofrer influência do tamanho da população de referência (GODDARD, 2009). Já a variância genética explicada pelos SNPs sofre influência do tamanho efetivo da população e da densidade de marcadores utilizados para genotipagem (PSZCZOLA et al., 2012).

Com o intuito de obter informações acuradas, na predição dos GEBVs em populações de validação, é importante que o LD entre as marcas e os genes ou QTLs sejam presentes e persistentes entre as populações de treinamento e validação (GRAPES et al., 2004, 2006; YU et al., 2005). Dessa maneira, quanto mais distante geneticamente for a população de treinamento da população de validação menor será a confiabilidade dos valores genéticos genômicos e menor a acurácia da predição para os indivíduos da população de validação.

Mesmo em situações nas quais diversas famílias compõem a população de estimação, as predições de novos cruzamentos podem gerar predições extremamente baixas ou até negativas, por falta de relacionamento genético das famílias envolvidas (LEHERMEIER et al., 2014). No entanto, predições mais acuradas podem ser obtidas, quando os indivíduos de uma família específica são avaliados fenotipicamente e genotipados para predizer os GEBVs de indivíduos da mesma família.

Muitos são os desafios, para implementação da seleção genômica, podem-se destacar o aumento das acurácias de seleção e aumento do tamanho da população de referência (HAYES et al., 2009). Deste modo, é importante um estudo detalhado das populações de interesse, para aplicação da seleção genômica, a fim de obter acurácias das predições elevadas para que sejam selecionados os indivíduos superiores.

2.3 Estrutura populacional na predição genômica

Grande parte dos estudos com objetivo de verificar a eficiência da seleção genômica leva em consideração apenas uma população, uma vez que os modelos preditivos são usados, principalmente, em populações e ambientes específicos (RESENDE, M. D. et al., 2012).

Para aumentar a capacidade preditivas dos modelos de seleção genômica, deve-se procurar minimizar o relacionamento entre os indivíduos da população de treinamento e os da população de validação, consequentemente, os genótipos da população de treinamento não devem ser relacionados entre si, mas devem representar toda a população (ISIDRO et al., 2015). Portanto o emprego de populações de estimação e seleção relacionadas propicia a obtenção de perdições mais acuradas.

Diferentes estruturas de populações podem ter grande efeito na escolha dos indivíduos para compor a população de validação. Isidro et al. (2015) testaram métodos, para otimização da população de treinamento, em dois painéis distintos de germoplasma com diferentes origens, diferentes estruturas de populações e diferentes características e relataram que a estrutura de população teve um importante papel na otimização da população de treinamento. Esses autores destacam que a estrutura populacional deve ser avaliada sempre antes de realizar as predições genômicas e que a estratificação da população com base na estrutura se comporta melhor, quando os alelos que controlam os caracteres são distribuídos, de acordo com a estrutura.

O estudo da estrutura genética populacional tem apresentado interesse crescente em diferentes áreas do conhecimento. Tecnologias como painéis de SNP de alta densidade e sequenciamento de nova geração têm facilitado a produção de um grande número de dados que permitiram a investigação das relações genéticas em humanos e em outros organismos.

Dentre os objetivos da genética de populações, o uso de dados moleculares permite resumir e informar sobre as relações entres os indivíduos em estudo. Uma das abordagens mais comuns, para analisar a estrutura de população, a partir de dados genômicos é a análise de componentes principais (PCA). A abordagem de PCA baseia-se na análise de uma matriz de covariância/correlação entre marcadores, que pode ser definida de diferentes maneiras, cujas entradas quantificam a semelhança genética entre pares de indivíduos. Os componentes principais dessa matriz representam as direções do espaço máximo da amostra que explicam o padrão observado de similaridade genética. A visualização dos padrões principais da estrutura nos dados pode ser alcançada, por representação gráfica de sucessivos componentes principais, em que grupos de indivíduos podem ser interpretados como populações genéticas, enquanto a mistura de duas populações resulta em grupos de indivíduos, ao longo de um linha, apesar de que outros eventos históricos, também, podem produzir sinais de componentes principais idênticos, além de outras questões que podem dificultar a interpretação dos componentes principais (LAWSON et al., 2012).

Um software que tem sido amplamente utilizado, para estudo da estrutura de população, é o STRUCTURE via abordagem bayesiana, introduzido por Pritchard, Stephens e Donnelly (2000), a fim de descrever populações cruzadas e detectar a estrutura genética de amostras, avaliação de mistura de populações, análise de hibridação e em estudos de migração e análise da proporção alélica.

2.4 Otimização da população de treinamento

A população de treinamento é aquela em que se obtêm as informações fenotípicas e genotípicas, para criar o modelo de predição, que posteriormente será empregado em uma população de indivíduos não testados para se obter os GEBVs para seleção (MEUWISSEN; HAYES; GODDARD, 2001). A otimização das populações de treinamento tem tido muito enfoque nos trabalhos recentes, visto ter impacto direto nas capacidades preditivas e nas acurácias dos modelos de predição por ser altamente influenciada pela população de calibração do modelo (HABIER; FERNANDO; DEKKERS, 2007; ISIDRO et al., 2015; PSZCZOLA et al., 2012).

Maiores populações de treinamento tendem a propiciar a obtenção de acurácias mais elevadas, principalmente, quando se trabalham com características controladas por maior número de genes com menores efeitos (GODDARD; HAYES, 2009).

Em situações nas quais temos uma população com indivíduos com elevada diversidade genética, uma dúvida que surge é como selecionar os melhores indivíduos, para compor a população de treinamento e construir o modelo estatístico, a fim de predizer os fenótipos dos indivíduos da população de validação (ISIDRO et al., 2015). Alguns estudos foram conduzidos com o intuito de sanar esta dúvida (GUO et al., 2013; PSZCZOLA et al., 2012).

As populações de treinamento tem sido foco de estudos em que estão envolvidos cruzamentos bi parentais e cruzamentos múltiplos, a fim de otimizar a composição dessas populações, para obtenção de predições mais acuradas (ASORO et al., 2011; ISIDRO et al., 2015; LORENZ; SMITH; JANNINK, 2012).

De acordo com Technow et al. (2014), o estabelecimento de grupos de treinamento com tamanho suficiente, para predição genômica em animais ou com germoplasma de diferentes grupos, no melhoramento vegetal, é extremamente oneroso, portanto agrupar conjuntos de indivíduos de diferentes grupos de germoplasma pode favorecer as predições dos indivíduos superiores e, assim, elevar o poder e ganhos obtidos pela seleção genômica. Para isso, é importante ter populações de treinamento grandes o suficiente para representar a população de melhoramento. Desse modo, o efeito do tamanho da população de treinamento utilizada, para construção do modelo genômico, tem importante papel à seleção de indivíduos superiores.

3 MATERIAL E MÉTODOS

3.1 População de melhoramento e dados fenotípicos

Para este estudo foram utilizados dados de uma população de melhoramento de *Eucalyptus* composta por 69 famílias híbridas, derivadas dos cruzamentos entre *Eucalyptus grandis*, *Eucalyptus urophylla*, *Eucalyptus camaldulensis* e *Eucalyptus saligna*. Quarenta e um genitores foram cruzados em um esquema de dialelo incompleto. As progênies obtidas dos cruzamentos foram plantadas a campo em Aracruz, Brasil (19°49'S/40°05'W), seguindo um delineamento experimental de alfalátice com 40 repetições e uma árvore por parcela. As informações detalhadas da população em estudo estão descritas na Tabela 1. Um total de 860 árvores foram fenotipadas para características de crescimento e qualidade de madeira aos 24 e 36 meses de idade.

Tabela 1 - Atributos gerais da população de melhoramento de *Eucalyptus* spp. em estudo.

Atributos	População
Número total de árvores	9400
Número total de famílias	232
Número médio de árvores por família	40
Número de parentais cruzados	41
Espécies dos parentais	*Eucalyptus grandis*, *E. urophylla*, *E. camaldulensis*, *E. saligna* e híbridos F_1 dessas espécies
Delineamento de cruzamento	Dialelo incompleto
Delineamento experimental	Alfalátice
Número de famílias amostradas para seleção genômica selection	69
Tamanho efetivo populacional (Ne)	27
Número médio de indivíduos amostrados por família	12
Número total de amostras para GS	860

Fonte: Do autor (2017).

Os caracteres de crescimento avaliados foram circunferência à altura do peito (CAP-2 e CAP-3), altura total (Alt-3) mensurada pelo hipsômetro Haglof Vertex (Haglof Inc, Madison, Mississippi) e espessura de casca (ESC) medida, a partir de um fragmento coletado à altura do peito, por meio de paquímetro digital.

A densidade de madeira (DEN-3) foi medida, após remoção da casca, com base na penetração de Pylodin. Para predição da qualidade de madeira, as amostras de serragem foram retiradas à altura do peito, utilizando um perfurador (10 mm de diâmetro), na direção norte-sul e cada amostra foi seca ao ar e utilizada, para medir indiretamente a qualidade da madeira por espectroscopia de refletância no infravermelho próximo (NIRS), para estimar as características químicas e físicas, utilizando um NirSystem 5000 (Foss Inc., Hillerød, Dinamarca). As curvas de calibração foram desenvolvidas internamente pela Fibria Celulose S.A. usando métodos descritos anteriormente por Raymond e Schimleck (2002). As características químicas da madeira estimadas foram o teor de pentosanas (NIR-PENT), a razão siringil/guaiacil (NIR-SG), rendimento da polpa (NIR-REN) e o teor de lignina (NIR-LIG). Os caracteres físicos da madeira foram densidade básica de madeira (NIR-DEN), comprimento da fibra (NIR-FIB), coarseness (NIR-COA) e número de fibras por grama (NIR-FGR).

Foi realizada análise de variância e estimados os componentes de variância genéticos e residuais para as 860 árvores de eucalipto avaliadas na população de melhoramento. Foram também estimadas as herdabilidades para todas as características por meio do software ASReml (GILMOUR et al., 2009).

3.2 Dados genotípicos

O DNA genômico foi extraído das folhas de 860 árvores, utilizando o método CTAB (*Cationic hexadecyl trimethyl ammonium bromide*), modificado (DOYLE; DOYLE, 1987), quantificado e diluído em concentrações entre 20-40ng.µL-1. As amostras de DNA foram genotipadas pela Geneseek (Lincoln, NE, EUA), utilizando a tecnologia Infinium (Illumina, San Diego), realizada com o uso de chip de DNA para eucalipto contendo 60904 SNPs (EUChip60k.br) (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015). Foi realizada a remoção de marcadores que falharam em mais de 10% (*Callrate* = 10%) das amostras para realização das análises subsequentes.

3.3 Análise do desequilíbrio de ligação, relacionamento e estrutura de população

As estimativas do desequilíbrio de ligação (LD) foram calculadas por meio da correlação das frequências alélicas de diferentes locos (r^2) (ROBERTSON; HILL, 1984).

As análises foram realizadas por meio do pacote do programa R LDcorSV (MANGIN et al., 2012). As estimativas de r^2 também foram obtidas por meio da correção do viés causado pela estrutura populacional (r^2_S). As análises foram realizadas com os 40902 SNPs polimórficos considerando a frequência do alelo menos frequente (MAF - *minor allele frequency*) maior que 0 genotipados para a população.

A curva de decaimento do LD foi ajustada por meio de modelo de regressão não linear dos valores de correlações de locos adjacentes (MARRONI et al., 2011) tanto para r^2 quanto para r^2_S. O decaimento foi plotado em gráfico até distância de 50 mil pares de base (kpb).

Para descrever a estrutura da população, foram utilizadas a primeira, empregando análises exploratória de componentes principais e o segundo, por meio de um método heurístico baseado em algoritmos de agrupamento bayesiano.

As análises de componentes principais (PCA) foram realizadas com os SNPs obtidos para a população. A partir da análise de componentes principais, foi construído o gráfico bidimensional. A análise de componentes principais foi realizada, por meio do software R, utilizando o pacote SNPRelate (ZHENG et al., 2012) e, com base nas estimativas dos pesos dos componentes principais, foi empregada a análise discriminante de componentes principais (JOMBART et al., 2010). Após a obtenção dos componentes principais, foi estimando o número de *clusters* (k) e utilizado o algoritmo *K-means* para identificar o melhor número de clusters, a partir do melhor valor de BIC *(Bayesian Information Criterion).*

A estrutura genética da população de melhoramento, baseada no método de agrupamento bayesiano, foi estimada com uso do software STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000). Dez interações foram realizadas para cada número de grupos testados (K), o qual variou de 1 a 10 populações, com dez repetições para cada K, com período de *burn-in* de 10 mil e 100 mil repetições da cadeia de Markov (MCMC). O número de grupos genéticos, de acordo com o indicado por Evanno, Regnaut e Goudet (2005), foi determinado por meio da plataforma web STRUCTURE HARVESTER (EARL; VONHOLDT, 2012). A matriz do K mais provável foi analisada pelo software CLUMP (JAKOBSSON; ROSENBERG, 2007).

A seleção dos SNPs, para análises da estrutura de populações, foi feita utilizando o pacote SNPRelate no software R (ZHENG et al., 2012), buscando utilização de um conjunto de SNPs que estejam em equilíbrio de ligação um com o outro. Esta abordagem conhecida como LD *Prunning* (PURCELL et al., 2007) tem como objetivo analisar o desequilíbrio de ligação entre dois marcadores adjacentes e retirar um dos SNPs se esse desequilíbrio estiver

acima do valor determinado a priori. Ambas as abordagens minimizam possíveis associações espúrias que possam ocorrer, além de reduzir a complexidade dos dados.

3.3.1 Predições genômicas

Os marcadores tiveram os efeitos estimados, utilizando o modelo RR-Blup – *"random regression best linear unbiased predictor"* (MEUWISSEN; HAYES; GODDARD, 2001), seguindo o seguinte modelo:

$$y = X\beta + Za + e,$$

em que: *y* é o vetor de dados fenotípicos, β é um vetor dos efeitos fixos, *a* é o vetor dos efeitos aleatórios dos indivíduos (Marcadores SNPs), *e* é o efeito aleatório residual; e X e Z são as matrizes de incidência para β e *a*, respectivamente. A equação do modelo misto para o BLUP (*best linear unbiased predictor* – melhor preditor linear não viezado) está apresentada a seguir:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\ \frac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix},$$

em que: Z é a matriz de SNP parametrizada, σ_g^2 é a variância genética total do caráter, σ_e^2 é a variância residual e *n* é o número de marcadores. A análise do modelo RR-BLUP foi realizada no software R (R CORE TEAM, 2015) utilizando o pacote rrBLUP (ENDELMAN, 2011).

Para validar a predição dos valores genômicos, foi utilizada a validação cruzada *ten-fold*, com a amostragem dentro das subpopulações criadas pela estratificação da população. Com a formação dos estratos de cada subpopulação, a população de treinamento foi criada, selecionando um número de indivíduos de cada estrato proporcional ao seu tamanho. Dessa maneira, estratos com maior número de indivíduos terão maior representação na população de treinamento que os menores estratos.

Foi, também, realizada a validação cruzada entre diferentes estratos de subpopulações criados. Nessa etapa, quando a população de treinamento era de um estrado, a população de

validação era de um estrato diferente, dessa maneira, a calibração do modelo foi realizada em uma subpopulação e a validação foi realizada em outra subpopulação.

Os efeitos dos marcadores estimados na população de treinamento e aplicados na população de validação para estimar os GEBVs. A capacidade preditiva da seleção genômica ($r_{y\hat{y}}$) foi estimada utilizando a correlação de Pearson entre os valores estimados dos GEBVs e os fenótipos da população de validação. Esse processo foi repetido 100 vezes a fim de estimar a média das capacidades preditivas.

3.3.1.1 Predições genômicas usando diferentes tamanhos de população de treinamento

O modelo RR-BLUP foi ajustado utilizando diferentes tamanhos de população de treinamento composta por diferentes números de indivíduos. Estes foram previamente amostrados aleatoriamente dentro da população de melhoramento.

Para cada tamanho de população de treinamento, foi estimada a capacidade preditiva e a herdabilidade genômica. Foram utilizados agrupamentos com 50, 100, 150, 200, 300, 400 e 800 indivíduos na população de treinamento obtidos aleatoriamente e os indivíduos restantes da população foram utilizados para a população de validação. Cada amostragem para cada número de indivíduos foi repetida por 1000 vezes. Assim foi possível obter as capacidades preditivas, para os tamanhos da população de treinamento em estudo, para cada característica avaliada.

3.3.1.2 Predições genômicas utilizando subgrupos de indivíduos e famílias

Para verificar o impacto da estrutura populacional e o relacionamento de indivíduos, na predição dos modelos de seleção genômica, os indivíduos foram agrupados de modo que minimizassem o relacionamento entre as populações de treinamento e de validação.

Foram utilizadas duas abordagens para estratificação; a primeira foi o estudo da estrutura de população, por meio das análises com modelo de agrupamento por estrutura populacional obtida pelo método bayesiano, e o segundo, o uso da análise de componentes principais (PCA) para agrupamento dos indivíduos e das famílias.

A definição das populações de treinamento e validação foi efetuada de três maneiras distintas, porém ambas levando em consideração a estrutura populacional obtida pelos grupos definidos com base nos componentes principais (PCA) e análises de estrutura genética pelo método bayesiano. A primeira estratificação foi realizada, considerando todos os indivíduos

da população conjuntamente os quais foram subdivididos, por meio da análise de PCA, denominado, neste estudo, como Indivíduos-PCA. Já o segundo agrupamento foi feito entre as 69 famílias que compõem a população, tido como Famílias-PCA. E, por último, foram utilizados os grupos genéticos obtidos pela abordagem bayesiana, tratado como Indivíduos-Structure.

O método de estratificação por Indivíduos-PCA foi realizado com base nas informações dos marcadores moleculares dos 860 indivíduos que compunham a população de melhoramento. Para isso, foram estimados os valores dos PCA e plotados os gráficos com os pesos de cada indivíduo e, assim, foram subdivididas as populações.

A estratificação das famílias (Famílias-PCA) foi realizada com base nos dados genotípicos médios das 69 famílias que compunham a população. Foi obtida a média da informação genotípica de cada SNP para cada família. Dessa maneira, os indivíduos oriundos de um mesmo cruzamento não seriam alocados aleatoriamente em grupos distintos, não desconsiderando o relacionamento por descendência desses indivíduos.

Com base na informação média de cada um dos SNPs, para as 69 famílias em estudo, foi realizada a PCA e realizado o agrupamento dos indivíduos dentro dos grupos das famílias mais relacionadas. Para obtenção desse valor médio, foi feito o somatório da leitura de cada SNP, para os indivíduos da referida família e dividido pelo número de indivíduos que compunha esta família. Após agrupamento das famílias e identificação dos indivíduos de cada grupo, foi realizado o processo de validação cruzada para verificação da capacidade preditiva por meio da estratificação de famílias.

Foram, também, obtidas as matrizes de parentesco *Kinship*, para evidenciar a relação de parentesco entre os indivíduos que compõem os diferentes estratos, a fim de verificar o relacionamento entre eles.

Para obtenção das capacidades preditivas para ambos os métodos de estratificação, foi realizada a validação cruzada em cada direção e foi feita a predição dentro dos estratos das subpopulações (indivíduos relacionados) e entre estratos (não relacionados). Para isso, foi utilizada a validação tipo “10-fold” em cada direção, considerando o mesmo número de indivíduos sendo amostrados na população de treinamento e na população de validação. Este procedimento foi repetido por 10 vezes e, ao final a capacidade preditiva, foi obtida pela média dos 100 valores estimados pela validação cruzada.

4 RESULTADOS

4.1 Dados fenotípicos e genotipagem

Foi realizada análise de variância e estimados os componentes de variância genéticos e residuais para as 860 árvores de eucalipto avaliadas na população de melhoramento. Foram também estimadas as herdabilidades para todas as características. Todas as características foram padronizadas e obtida a soma dos valores padronizados (PAD) a fim de facilitar as comparações futuras (TABELA 2).

Tabela 2 - Resumo das características fenotípicas avaliadas para a população de melhoramento de eucalipto composta por 860 indivíduos.

Característica	Unidade	Média	Var. genética	Var. residual	Herdabilidade
CAP-2	cm	30,01	7,9003	12,0281	0,3663
CAP-3	cm	36,95	31,4845	36,4235	0,4446
ESC	mm	5,41	1,0037	0,4274	0,6962
ALT-3	m	17,47	2,3419	4,4849	0,3378
NIR-COA	mm	6,63	0,0363	0,5338	0,0576
NIR-DEN	Kg/m^3	414,65	127,8520	748,4950	0,1446
NIR-FGR	Milhões/grama	25,44	3,8381	6,5128	0,3314
NIR-FIB	%	0,70	0,0001	0,0015	0,0716
NIR-LIG	mg/100m	30,09	0,1295	1,0703	0,1068
NIR-PENT	%	16,74	0,5125	0,4354	0,5174
NIR-REN	%	52,66	0,1767	1,3503	0,1133
NIR-SG	-	0,70	0,0007	0,0134	0,0476
DEN-3	mm	14,46	5,3959	2,4745	0,6557
PAD	-	0,00	13,6837	11,5378	0,5289

Fonte: Do autor (2017).

Quanto à genotipagem, dos 60904 SNPs selecionados a priori e estabelecidos no Chip de DNA (SILVA-JUNIOR; FARIA; GRATTAPAGLIA, 2015), 49201 (80,7%) foram genotipados para a população de melhoramento de eucalipto em estudo, usando o arquivo de agrupamento SNP apropriado para chamadas SNP e filtros para SNPs com "call rate" $\geq$ 90%. Após a seleção de SNP polimórficos (MAF>0), 40902 SNPs foram utilizados para análises, com uma taxa final de dados faltantes de 1,2%. Esses marcadores que apresentaram dados faltantes foram substituídos pelo alelo mais frequente.

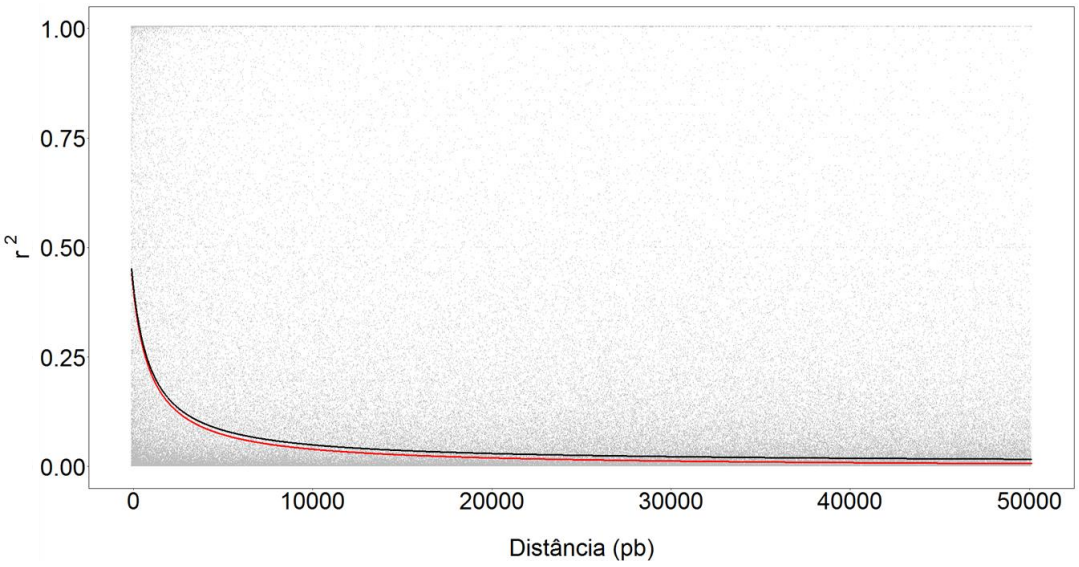
4.2 Desequilíbrio de Ligação (LD)

O desequilíbrio de ligação foi calculado para todas as distâncias, comparados aos pares, entre todos os 40932 SNPs polimórficos (MAF>0) em cada cromossomo separadamente (média de 3721 SNPs por cromossomo), com distância variando de 30 pares de base até milhares de pares de base. A média do LD do genoma, para um par de SNPs a uma distância de 100 Kbp um do outro, foi de r^2=0,0766. Ao corrigir o LD para a estrutura populacional (r^2_S), as estimativas médias foram reduzidas para 0,0465 (Figura 1).

O LD do genoma apresentou decaimento até um r^2 abaixo de 0,2 dentro de 15,6 Kb e 70,6 Kb (linha preta), enquanto o r^2_S (corrigindo a estrutura da população) mostrou um decaimento ligeiramente mais rápido dentro de 12,7 e 56,5 Kb (linha vermelha). O decaimento do LD mais rápido para o r^2_S confirma o efeito da estrutura populacional para esta população de melhoramento.

Observaram-se diferenças no LD médio por cromossomos (Tabela 3), como mostrado pelo gráfico de decaimento de LD. Podem-se tomar como exemplo os cromossomos 5 e 9, o cromossomo 5 mostrou uma taxa um pouco mais baixa de decaimento do LD, quando comparado com o cromossomo 9, sugerindo existência de diferença na taxa de recombinação. Tomando-se por base um r^2 <0,2 como limiar, o LD apresentou decaimento < 1Mb, consistente com o reduzido tamanho efetivo da população (Ne = 27) e origem híbrida dessa população, principalmente, quando se trata de uma população a qual envolve multiespécies.

Figura 1 - Padrão de decaimento do LD em população de melhoramento de *Eucalyptus* spp. até distância de 500 kpb nas comparações par a par de todos os 11 cromossomos. Curvas de decaimento sem a correção para estrutura de população (linha preta) e decaimento ajustado para estrutura de população (linha vermelha).



Fonte: Do autor (2017).

Tabela 3 - Estatísticas do número de SNPs genotipados e utilizados nas análises e estimativas cromossômicas individuais do desequilíbrio de ligação r2 considerando todos alelos (MAF>0) na população de melhoramento de eucalipto.

Cromossomos	#SNPs	#SNPs MAF>0	r^2 médio	r^2_S médio
1	3616	3067	0,0800	0,0654
2	5388	4505	0,0765	0,0643
3	4869	3895	0,0710	0,0508
4	3477	2913	0,0793	0,0678
5	4546	3462	0,0658	0,0567
6	5483	4752	0,0778	0,0610
7	3994	3257	0,0766	0,0467
8	5967	4936	0,0729	0,0569
9	3541	3021	0,0848	0,0789
10	3975	3390	0,0814	0,0545
11	4345	3734	0,0765	0,0487
Todos	49201	40932	0,0766	0,0465

Fonte: Do autor (2017).

4.3 Estrutura de população

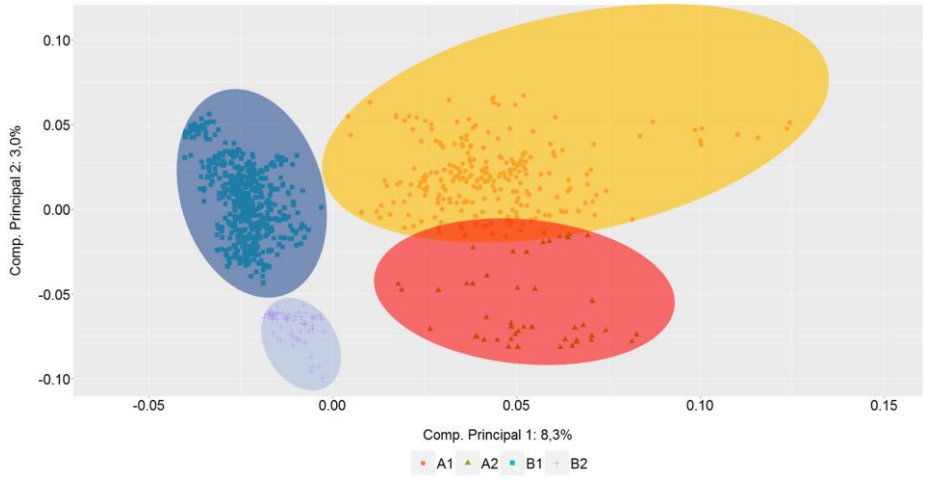
Foram realizadas as análises de componentes principais, para apresentação e sumarização da variância genética explicada pelos marcadores SNP's, para todos os 860 indivíduos da população assim como para as 69 famílias (Figuras 2 e 3). Foi utilizado um subgrupo de 4126 SNPs extraídos, com base no LD *prunning* e foram amostrados, em média, 375 SNPs por cromossomo.

A análise de agrupamento revelou que os 860 indivíduos que compõem a população de melhoramento de eucalipto foram particionados pelos dois primeiros componentes principais, contabilizando 8,3% e 3,0% da variância genética para o primeiro e segundo componentes, respectivamente (Figura 2). A análise de agrupamento dividiu a população em quatro subpopulações, pois elas continham de 233 (A1), 57 (A2), 506 (B1) e 64 (B2) indivíduos, baseando na maximização da distância genética entre subpopulações.

Os dois primeiros componentes principais do agrupamento por famílias explicaram 8,2% e 5,4% da variação genética, respectivamente (Figura 3). A análise de agrupamento possibilitou a divisão em duas subpopulações; uma foi composta por 42 famílias e a outra por 27. Dentre elas, a primeira continha 305 (A) indivíduos e a segunda 555 (B) indivíduos.

A análise da estrutura genética com uma abordagem bayesiana (Indivíduos-Structure) mostrou que k=3 foi o número ideal de grupos genéticos (K) que melhor ajustou, uma vez que apresentou o maior valor de ΔK (Figura 4), de acordo com as inferências apresentadas por Evanno, Regnaut e Goudet (2005).

Figura 2 - Gráfico dos dois primeiros componentes principais e análise de agrupamento com 40932 marcadores SNPs para a população de melhoramento de *Eucalyptus* spp. com 860 indivíduos originados da hibridação de multiespécies.

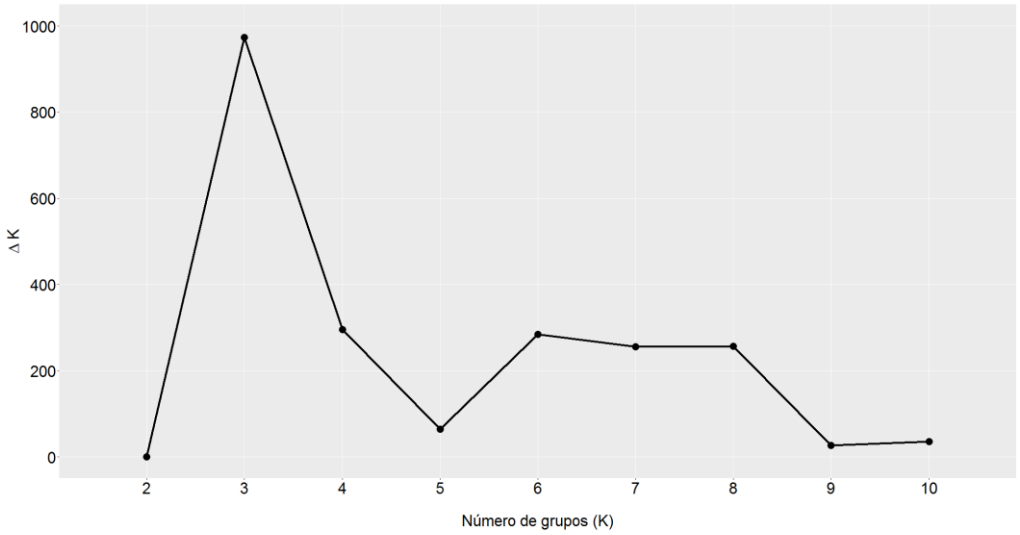


Fonte: Do autor (2017).

Figura 3 - Gráfico dos dois primeiros componentes principais e análise de agrupamento com 40902 marcadores SNPs das 69 famílias da população de melhoramento de *Eucalyptus* spp.

Fonte: Do autor (2017).

Figura 4 - Valores de delta K (ΔK) para os respectivos números de grupos (K) obtidos, por meio da análise de agrupamento bayesiano, via software STRUCTURE.



Fonte: Do autor (2017).

Com uma probabilidade de adesão superior a 0,60, o agrupamento bayesiano indicou que os 860 indivíduos foram agrupados da seguinte forma: 419 indivíduos para o grupo 1, 163 indivíduos para o grupo 2 e 278 para o grupo 3 (Figura 5). Dos 860 indivíduos, 128 apresentaram probabilidades mistas; 117 ficaram entre os grupos 1 e grupo 3, 3 indivíduos

entre os grupos 1 e 2 e o 8 ficou entre os grupos 2 e 3. Dessa maneira, os indivíduos foram atribuídos aos grupos de maior confiança.

Figura 5 - Agrupamento pela inferência bayesiana da população de melhoramento de Eucalyptus (k=3).

Fonte: Do autor (2017).

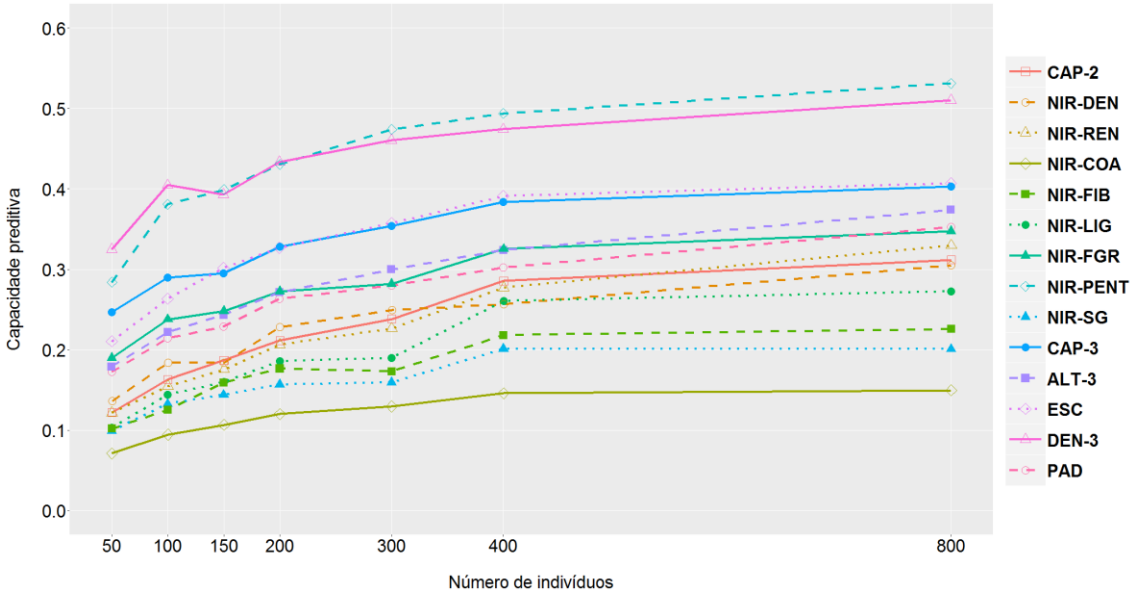
4.4 Impacto do tamanho da população de treinamento nas predições genômicas

As análises de predição genômica foram realizadas utilizando diferentes tamanhos de população de treinamento, para obtenção das capacidades preditivas do modelo, dentre os diferentes tamanhos de população utilizados (50, 100, 150, 200, 300, 400 e 800 indivíduos) (Figura 6). A distribuição das capacidades preditivas, para cada tamanho de população, foi obtida por meio da amostragem aleatória das 860 árvores.

As capacidades preditivas médias variaram de 0,09 para NIR-SG (50 indivíduos) e 0,53 para NIR-PENT (800 indivíduos). Para todas características avaliadas, a capacidade preditiva apresentou um aumento de acordo com aumento da população de treinamento. No entanto, pode-se notar que, para algumas características, este aumento foi maior que para outras, (por exemplo, DEN-3 e NIR-PENT), indicando que há um aumento de diferentes intensidades, para características específicas, quando se tem maior número de indivíduos na população de treinamento.

As capacidades preditivas, para a variável padronizada, apresentaram comportamento intermediário, quando comparadas a todas as características. Os valores de capacidade preditiva variaram de 0,18 até 0,35, quando se avaliaram 50 e 800 indivíduos. Para a maioria das variáveis, o aumento na predição foi pequeno e não significativo, quando aumentou de 300 para 800 indivíduos.

Figura 6 - Capacidades preditivas dos genótipos da população de validação de *Eucalyptus* spp. Sete tamanhos diferentes de população (50, 100, 150, 200, 300, 400 e 800 indivíduos) foram utilizados, para otimização do tamanho da população de treinamento, para as 13 características avaliadas e para a variável padronizada.



Fonte: Do autor (2017).

4.5 Impacto do relacionamento de indivíduos na população de treinamento

Com o intuito de verificar o impacto do relacionamento entre indivíduos para aplicação da seleção genômica, primeiramente, foi obtida a média do relacionamento entre os indivíduos de cada população de referência ou treinamento estabalecida neste estudo.

A população de referência, composta por todos os selecionados aleatoriamente, apresentou coeficiente de relacionamento genômico de, aproximandamente, 0,05 (Tabela 4). Os agrupamentos que apresentaram maior coeficiente de relacionamento foram aqueles obtidos, por meio do agrupamento dos indivíduos (0,09 à 0,15), enquanto o agrupamento por famílias apresentou valores intermediários (0,08 e 0,10).

Quando a população de melhoramento foi fracionada por meio do agrupamento bayesiano, verificou-se que foram obtidos coeficientes de relacionamento superiores aos obtidos pelo agrupamento por famílias, porém inferiores ao agrupamento realizado pelos indivíduos, variando, portanto, de 0,09 a 0,11.

Tabela 4 - Médias dos coeficientes de relacionamento genômico por estado entre os indivíduos que compõem as populações de referência entre os diferentes grupos avaliados e número de indivíduos em cada um dos grupos.

Grupos	População de referência	Coeficiente de relacionamento	Número de indivíduos
Amostragem Aleatória	Todos	0,0588	860
Famílias-PCA	A	0,1084	305
	B	0,0883	555
Indivíduos-PCA	A1	0,1197	233
	A2	0,1491	57
	B1	0,0916	506
	B2	0,1589	64
Indivíduos-Structure	1	0,0909	419
	2	0,1116	163
	3	0,0901	278

Fonte: Do autor (2017).

Para estimar as capacidades preditivas para as populações de referência, os modelos de seleção genômica foram obtidos, visando reduzir o relacionamento entre os indivíduos da população de treinamento e de validação, baseados na diferenciação da estrutura genética obtida por meio das PCA dos inivíduos e das famílias, além do agrupamento bayesiano gerado pelo software STRUCTURE.

Como esperado, a capacidade preditiva foi baixa, quando a população de treinamento era composta por uma subpopulação utilizada para predizer uma outra subpopulação. Isso aconteceu, em todos os casos, tanto entre quanto dentro dos métodos de estratificação. Era esperado que, para todas as situações, as predições fossem mais altas entre indivíduos relacionados, quando comparados à combinação de todas as subpopulações, ou seja, de todos os indivíduos. No entanto esse comportamento foi variável entre os métodos de estratificação e entre caraterísticas fenotípicas.

Um ponto que pode explicar, parcialmente, esta situação é o número de indivíduos que compuseram as subpopulações. Nos casos em que mais de 300 indivíduos compunham as subpopulações, este tamanho de população de estimação já atingiu um ponto o qual reduziu o acréscimo na predição, quando se considera o tamanho da população de treinamento. Em todas situações avaliadas, quando as populações de validação não eram relacionadas às populações de estimação, as capacidades preditivas foram extremamente baixas (Figura 7a, 7b e 7c), apresentando até estimativas negativas.

No primeiro método de estratificação (Famílias-PCA), para seis características, foram obtidos valores negativos de capacidade preditiva. Isso foi verificado para as caracteristicas NIR-REN, NIR-COA, NIR-LIG, NIR-SG, NIR-ESC e DEN-3. O comportamento das capacidades preditivas obtidas, por meio da estratificação dos indivíduos por PCA, foi variável entre as características avaliadas nas situações em que os grupos de validação e treinamento eram relacionados (Figura 7a). Para as variáveis NIR-REN, NIR-LIG, NIR-SG e NIR-ESC, as predições foram superiores, apresentando diferenças significativas às da amostragem aleatória de todos os indivíduos da população. Já para as características NIR-DEN, NIR-FGR, NIR-PENT, CAP-3 e DEN-3, a amostragem aleatória foi superior à do método de estratificação em questão. Nos demais, não foram detectadas diferenças significativas entre as capacidades preditivas dos indivíduos relacionados e dos amostrados aleatoriamente. Na média de todas as características avaliadas, houve redução de 2% nas capacidades preditivas e, para a varíavel padronizada, houve redução de 15%, quando comparado com a análise de todos os indivíduos.

Para o segundo método de estratificação, por meio dos PCA das familias, não foram detectadas estimativas negativas para as capacidades preditivas, para nenhuma das 14 características avaliadas (Figura 7b), no entanto as estimativas das capacidades preditivas, para as subpopulações que não são relacionadas, apresentaram valores muito baixos para todas as características. Para este método de estratificação, seis características avalidas foram, significativamente superiores, incluindo a variável padronizada. Para cinco características, a predição obtida pelo método aleatório foi superior às obtidas pela estratificação das famílias. Para as demais, não houve diferença significativa ($P<0,05$). Foi verificado aumento médio de 3% comparando com a análise de todos os indivíduos e, para a variável padronizada, o aumento foi da ordem de 8%.

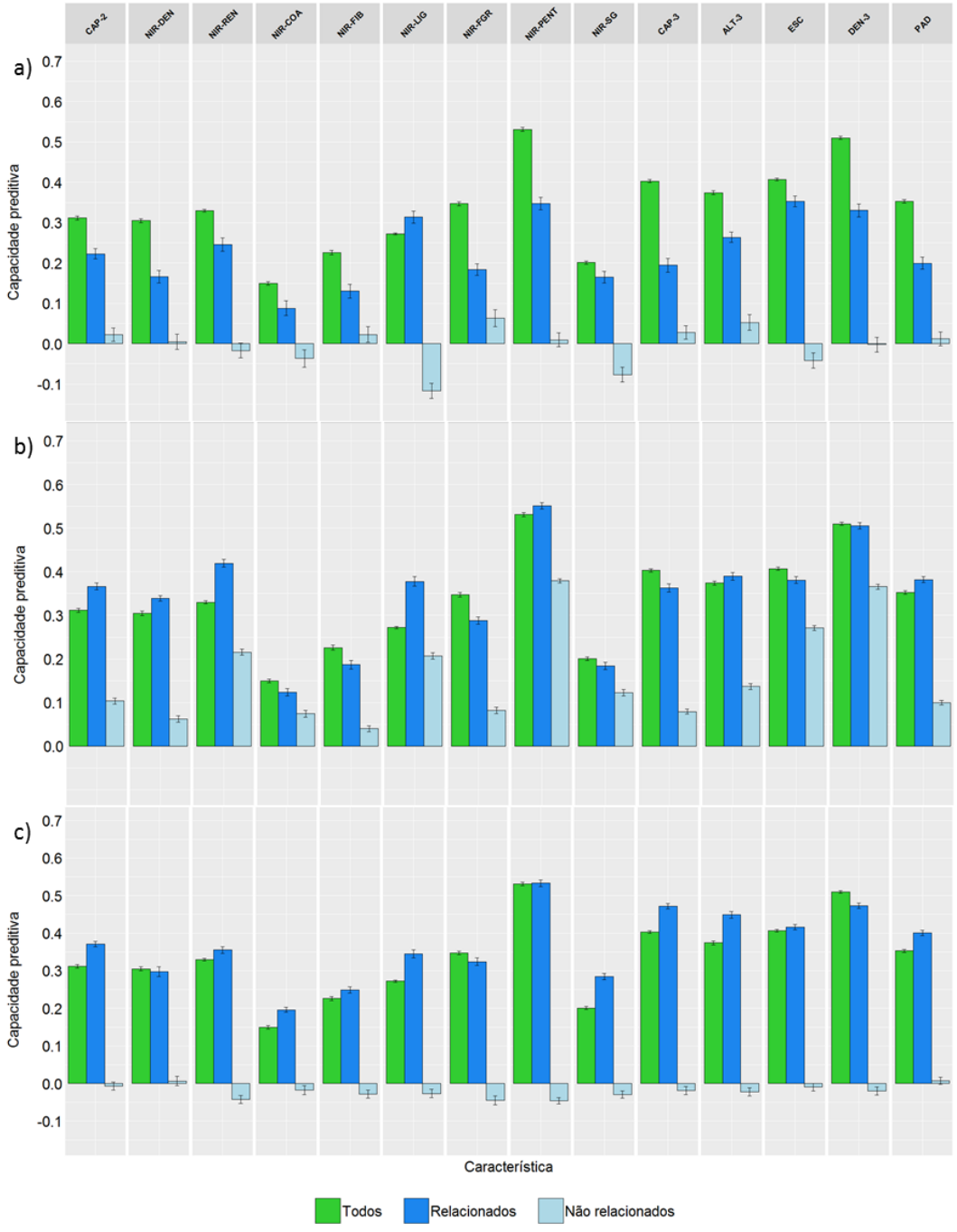
Já para o último método de estratificação por meio do agrupamento bayesiano, foram verificadas estimativas negativas, para as 13 características avaliadas, exceto para a variável padronizada (PAD), quando o grupo de estimação não foi relacionado ao grupo de validação (Figura 7c). Para nove características, as predições foram superiores ao método de amostragem aleatório, incluindo a variável padronizada. Somente em três situações a amostragem aleatória foi superior ao método de amostragem, via estratificação, por meio da estratificação pela agrupamento bayesiano (Indivíduos-Structure). Na média geral de todos os caracteres avaliados, a estratificação de Indivíduos-Structure levou a um aumento na capacidade preditiva de 9% sobre a análise considerando todos indivíduos. Para a variável padronizada, o aumento foi de 14%.

Quando foram comparados os métodos de estratificação, Indivíduos-PCA, Famílias-PCA e Indivíduos-Structure, foi possível verificar que o método Indivíduos-PCA foi superior somente para a característica NIR-LIG. Nos demais, esse método foi o que teve comportamento inferior ou igual aos demais métodos avaliados.

Os métodos de estratificação Famílias-PCA e Indivíduos-Structure apresentaram diferenças, na magnitude das capacidades preditivas, para dez características avaliadas, no entanto não apresentaram diferença significativa para a variável padronizada.

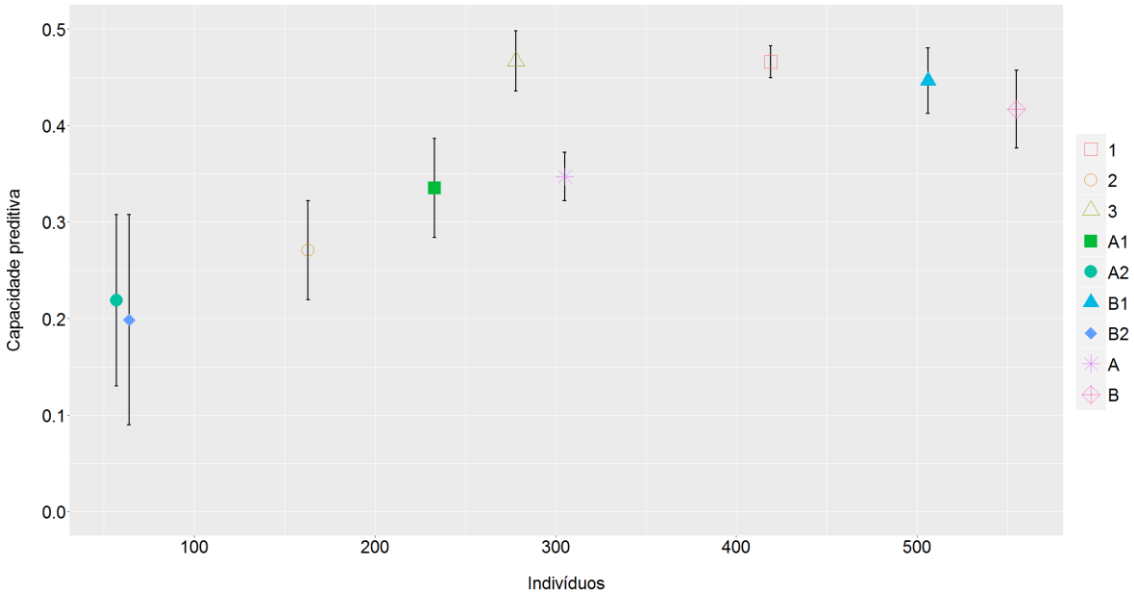
Com o intuito de verificar o comportamento das capacidades preditivas entre as subpopulações relacionadas, foram dispostas, na Figura 8, as médias das capacidades preditivas de cada uma das subpopulações utilizadas, para predição de indivíduos desta mesma subpopulação, para a variável padronizada. Os resultados obtidos mostraram que maiores populações obtiveram maiores capacidades preditivas. Foi possível visualizar que maiores médias foram obtidas com populações de tamanho superior a 278 indivíduos (subpopulação 3). Somente para a subpopulação A, obtida pelo método Famílias-PCA, houve uma redução da capacidade preditiva. Os resultados indicaram que as subpopulações obtidas pela estratificação Indivíduos-Structure geraram maiores capacidades preditivas, principalmente, ao desconsiderar subpopulações com menos de 200 indivíduos que apresentaram capacidades preditivas muito baixas.

Figura 7 - Capacidades preditivas do modelo de seleção genômica rr-BLUP, incluindo todos indivíduos, as subpopulações compostas por indivíduos relacionados e entre subpopulações de indivíduos não relacionados para as 14 variáveis respostas avaliadas. Todos: alocação aleatória dos indivíduos na população de treinamento e validação. Relacionados: alocação aleatória dos indivíduos de uma mesma subpopulação no conjunto de treinamento e validação. Não relacionados: subpopulações diferentes foram utilizadas para criar a população de treinamento e validação. a) Estratificação de todos os indivíduos por PCA. b) Estratificação de famílias por PCA e c) Estratificação de todos os indivíduos por STRUCTURE.



Fonte: Do autor (2017).

Figura 8 - Capacidades preditivas da variável padronizada, em função do tamanho da população entre indivíduos relacionados, para os três métodos de estratificação. Os pontos representam a capacidade preditiva média das validações cruzadas entre indivíduos relacionados dentro da mesma subpopulação, em que os grupos 1, 2 e 3 são referentes ao método de estratificação de Indivíduos-Structure; A1, A2, B1 e B2 referentes à estratificação Indivíduos-PCA; A e B referem-se às subpopulações do método Família-PCA.



Fonte: Do autor (2017).

5 DISCUSSÃO

O tamanho da população de estimação tem importante papel na seleção genômica, a qual deve conter número suficiente de indivíduos que representem a população de melhoramento (LORENZ; SMITH; JANNINK, 2012). Foi possível verificar que, quando se aumentou o número de indivíduos na população de treinamento, as capacidades preditivas também aumentaram para todas as características avaliadas. Esta situação foi verificada em diversos estudos para culturas agrícolas como trigo, arroz, ervilha, aveia e soja (ASORO et al., 2011; ISIDRO et al., 2015; JARQUÍN et al., 2014; SPINDEL et al., 2015; TAYEH et al., 2015).

Neste trabalho, verificou-se que diferentes tamanhos de populações, quando utilizadas para treinamento, levaram à obtenção de predições divergentes e, principalmente, quando foram criados estratos com poucos indivíduos, as predições foram muito baixas e o método de estratificação da população se mostrou pouco eficiente. Esta situação corrobora o resultado apresentado por Isidro et al. (2015).

Este aumento não é dependente da metodologia utilizada para otimização do grupo de treinamento, assim como relatado por Akdemir, Sanchez e Jannink (2015), que não verificaram diferenças significativas, nas acurácias da GS entre diferentes métodos de seleção dos indivíduos, para compor a população de treinamento.

A contribuição do efeito da estrutura de população pode ser visualizada pelo importante papel do LD nas predições genômicas. O efeito da estrutura de população pode ser verificado, uma vez que o LD apresentou comportamento distinto, quando o r^2_S foi corrigido para estrutura populacional. Pode-se verificar um decaimento do LD consistente com as taxas de recombinação variáveis relatadas, previamente, para populações de *Eucalyptus* (SILVA-JUNIOR; GRATTAPAGLIA, 2015).

Desse modo, os resultados aqui obtidos coincidem com estudos prévios os quais indicam que o relacionamento ou não entre indivíduos e entre as populações de treinamento e validação podem afetar a implementação da seleção genômica, sub ou superestimando as predições (MÜLLER et al., 2017). A otimização das populações de treinamento para predição genômica, por meio da estratégia de estratificação das famílias, minimizou o relacionamento entre os indivíduos da população de treinamento ou estimação e maximizou o relacionamento entre o conjunto de indivíduos entre a população de validação e de treinamento.

Faz-se importante ressaltar que é de grande valia ter informações da população, na qual se deseja aplicar a seleção genômica, para, assim, otimizar a população de treinamento e

obter melhores predições. Com isso, a estratificação, por meio da estrutura de população, pode minimizar o relacionamento entre indivíduos da população de treinamento e aumentar o relacionamento entre os indivíduos da população de validação e de treinamento e elevar a acurácia dos modelos preditivos. Isso significa que os indivíduos da população de treinamento não necessariamente devem estar intimamente relacionados, mas devem representar a população como um todo (ISIDRO et al., 2015).

O grande benefício de utilizar a estrutura de população, para predição de indivíduos relacionados, é que considera-se a frequência dos alelos que compõem os diferentes genótipos das diferentes famílias, desde que tenha o número mínimo de 300 indivíduos, para que seja possível amostrar a variabilidade genética da população. A maior eficiência da seleção genômica, quando se faz a estruturação dos indivíduos com maior relacionamento, já foi relatada por Valente et al. (2016), em estudo de simulação para emprego seleção genômica de progênies com diferentes estruturas populacionais.

Foi indicado por Massman et al. (2013) que, quando se estimam os GEBVs em populações que não apresentam estrutura similar às populações de estimação dos efeitos dos marcadores, estas informações obtidas não são acuradas e reduzem o poder da seleção genômica, consequente à seleção de indivíduos não superiores.

Os indivíduos que terão seus GEBVs preditos devem estar relacionados à população, na qual os efeitos dos marcadores serão estimados, caso contrário, qualquer relacionamento entre os indivíduos da população de treinamento e validação não repercutirá em maiores capacidades preditivas do modelo genômico (VALENTE et al., 2016); portanto, quando se estruturam os indivíduos em populações com maior relacionamento genético ou parentesco, maior será a eficiência da seleção genômica.

Já foi demonstrado, em alguns estudos no melhoramento animal que, adicionando à população de treinamento indivíduos derivados de outras raças ou populações de melhoramento, pode haver um amento da acurácia preditiva da seleção genômica (ERBE et al., 2012; HAYES et al., 2009). Para a cultura do milho, Technow et al. (2014) demonstraram que, quando se reúnem indivíduos de diferentes origens ou germoplasma, pode-se favorecer o aumento das capacidades preditivas. Desse modo, é importante ressaltar que, neste estudo, diversas famílias provenientes de diferentes cruzamentos interespecíficos foram agrupadas, mostrando, assim, que, quando se utilizou o modelo com todos indivíduos da população, os valores da capacidade preditiva foram muito próximos aos valores daqueles em que se agruparam com base na estrutura de população.

Não se pode esquecer, também, que, quando se agrupam os indivíduos com base na estrutura de população, também, consideram-se os indivíduos de diversas famílias, com origens diferentes. Foi também demonstrado para gado leiteiro que agrupamento de indivíduos de raças diferentes propiciou maiores acurácias de predição dentro das raças (ERBE et al., 2012). Estes resultados puderam ser verificados por meio dos efeitos dos marcadores avaliados nas diferentes populações. Foi visto em milho que os efeitos de marcadores estimados dentro de cada população apresentaram pouca ou nenhuma vantagem sobre os modelos em que se consideraram efeitos dos marcadores entre diferentes populações (TECHNOW; BÜRGER; MELCHINGER, 2013). Salienta-se que a consistência da fase de ligação foi alta o suficiente, nos painéis com elevado número de marcas, indicando, assim, a capacidade dos modelos em capturarem pequenas diferenças nos efeitos dos marcadores entre populações diferentes.

Quando ajustados os efeitos dos marcadores dentro das subpopulações, foram obtidas predições diferentes comparadas às predições para todos os indivíduos conjuntamente. Assim, a predição das performances dos híbridos interespecíficos pode ser guiada, uma vez que se têm indivíduos aparentados dentro da mesma população de treinamento. Esses resultados corroboram as afirmações feitas para híbridos de milho e também nos estudos de simulação (TECHNOW et al., 2014).

Nos agrupamentos utilizados no presente estudo, quando se utilizou o agrupamento por família, foi possível verificar que, quando havia indivíduos aparentados na população de treinamento, foi possível obter capacidades preditivas elevadas, quando comparados às populações de treinamento que continham indivíduos não relacionados aos da população de validação.

Não somente verificado pelo parentesco dos indivíduos, mas também pelo coeficiente de relacionamento obtido, por meio da identidade por estado, que leva em consideração somente as informações dos marcadores. Quando se estratificou a população com informações das famílias, foram obtidas predições superiores de quando se levou em consideração somente a informação dos SNPs (Indivíduos-PCA). O agrupamento, considerando a estrutura de população obtida, por meio da estimativa bayesiana, teve comportamento intermediário.

A importância de se ter indivíduos com maior grau de parentesco nas populações de treinamento também foi relatado por Legarra et al. (2008). Deste modo, é possível selecionar indivíduos superiores, quando há envolvidos, na população de treinamento, parentes mais próximos daqueles preditos que compõem a população de interesse.

Nas situações em que se tem pequeno tamanho efetivo da população de estimação, maior é a importância do relacionamento entre os indivíduos e da estrutura populacional para o modelo de predição genômica (WIENTJES; VEERKAMP; CALUS, 2013). No presente estudo, o tamanho efetivo populacional é Ne=27, sendo considerado uma população com reduzido tamanho efetivo, daí a grande importância de se ter um bom entendimento de ambas as populações, tanto a de estimação quanto a população em que se deseja selecionar os indivíduos superiores por meio dos GEBVs.

Observa-se que, para a predição dos valores genômicos, considerando modelo aditivo, características com maiores herdabilidades (h^2), serão aqueles que apresentarão maiores capacidades preditivas, quando comparados àqueles com baixa h^2 (LIN; HAYES; DAETWYLER, 2014). Neste estudo, mesmo os caracteres que apresentaram altas herdabilidades, quando se obtiveram as predições entre indivíduos não relacionados, as capacidades preditivas foram muito baixas comparadas às obtidas entre indivíduos relacionados. Esses resultados também foram reportados por Müller et al. (2017), para duas populações de melhoramento de duas espécies de *Eucalyptus*, em que se obtiveram predições negativas, quando as populações de treinamento eram não relacionadas às populações de validação. Neste trabalho, os autores avaliaram duas populações obtidas do cruzamento intraespecífico de *E. pellita* e *E. benthamii*, porém, mesmo utilizando populações que envolviam uma única espécie, foi possível verificar o efeito da estrutura de população nas predições genômicas.

Na situação em que se tem populações oriundas de cruzamentos multiparentais e, principalmente, no caso de *Eucalyptus* em que se tem cruzamentos interespecíficos, é de grande importância que sejam realizados estudos da estrutura populacional antes de se construir os modelos de predição. Essa situação se apresenta um pouco diferente para populações biparentais (MARULANDA; MELCHINGER; WÜRSCHUM, 2015).

Ao avaliar a composição da população de treinamento, foi possível verificar que a capacidade preditiva aumentou com o aumento do tamanho da população de treinamento, assim como esperado. O maior aumento foi verificado, quando se passaram de 50 indivíduos para 300 indivíduos que compunham a população de treinamento. Porém o aumento foi pequeno da capacidade preditiva, quando se compararam os tamanhos de 400 e 800 indivíduos utilizados na predição dos modelos. Portanto aumentos consideráveis aconteceram até 300 indivíduos, na população de treinamento, indicando, assim, que, para se obter acurácias maiores na predição de indivíduos em populações de eucaliptos, é indicado ter em torno de 300 árvores para compor a população de treinamento. Números inferiores a esse

podem não representar fielmente a variabilidade da população e induzir a predições pouco acuradas dos indivíduos de interesse.

Foi mostrado em diversos trabalhos que, quando se reduziu pela metade o número de indivíduos na população de treinamento, as predições não tiveram decréscimo significativo. Isto também foi verificado por Albrecht et al. (2011) em milho, em que eles relataram que a contribuição para este fato foi pelas populações de treinamento ainda serem compostas por indivíduos com estreito relacionamento aos que compuseram as populações de validação. Isso é verificado, pois, no esquema de validação cruzada, se amostram indivíduos da população como um todo, e aqueles que compõem a população de validação têm normalmente irmãos completos na população de treinamento.

Mesmo com todas essas discrepâncias nas acurácias preditivas, em alguns casos, não foi possível verificar diferença significativa nas predições dos modelos nos diferentes cenários avaliados. Isso corrobora o comentário de Technow et al. (2014), ao enfatizarem uma importância desproporcional para a predição genômica de indivíduos aparentados. Foi relatado no caso de milho que, mesmo considerando subpopulações com diferentes origens e relacionamentos, não foram verificadas diferenças em todos os casos estudados, uma vez que as linhagens são muito próximas umas das outras em um programa de melhoramento. No presente trabalho, essa mesma situação foi verificada para diferentes características avaliadas. Isso se deve à condição de que, embora sendo uma população que contenha híbridos interespecíficos, as espécies que os originaram são intimamente relacionadas, refletindo na presença de um alto grau de parentesco entre os indivíduos da população em estudo.

As estimativas da capacidade preditiva dos modelos apresentaram grande variação, quando são considerados indivíduos relacionados e não relacionados na população de treinamento. Já a menor variação entre os casos, nos quais foram utilizados todos indivíduos da população e indivíduos relacionados para predição dos modelos, pode ser explicada pela garantia de se ter indivíduos relacionados nas populações de treinamento. No entanto, mesmo considerando indivíduos relacionados, diferentes tamanhos da população de treinamento contribuíram para as diferenças detectadas neste estudo, levando-se a ter diferenças significativas nas capacidades preditivas dos modelos.

O efeito do tamanho da população de treinamento é realçado no agrupamento dos indivíduos, por meio da estrutura de população, pois, quando se obtiveram estratos com menor número de indivíduos os quais foram utilizados como população de treinamento, a capacidade de predição do modelo de seleção genômica foi baixa.

De acordo com os resultados obtidos, pode-se indicar que as populações de estimação devem conter um número grande de indivíduos que estejam relacionados aos indivíduos e às populações de validação, no entanto, dependendo dos custos, obter os dados fenotípicos de todos indivíduos genotipados em uma população, em vez de avaliar fenotipicamente somente um grupo de indivíduos em diversas repetições ou locais, pode ser mais eficiente (ENDELMAN et al., 2014; LORENZ, 2013), o que mostra um balanço nos custos de fenotipagem, sendo restrito basicamente a indivíduos a serem avaliados fenotipicamente e genotipados em um único local e/ou repetição. Assim, delineando melhor a população de treinamento, pode-se alocar melhor os recursos, a fim de obter maiores estimativas de acurácia providas por maiores populações de estimação.

É possível verificar neste estudo que as informações providas pelos SNPs se apresentam vantajosas e serviram como informação a priori, para otimização das populações de estimação em populações de cruzamentos múltiplos, especialmente, para cruzamentos interespecíficos, que foi o caso da população em estudo. Isso sugere que, mesmo na ausência da informação de pedigree e na falta de informações fenotípicas prévias dos indivíduos, é possível se delinear de forma satisfatória as populações de estimação com as informações dos SNPs dos indivíduos que serão preditos. Isso se mostra importante para aplicação de seleção, em estágios mais precoces e, mantendo indivíduos relacionados na população de estimação, poderá maximizar a capacidade de predição da seleção genômica.

6 CONCLUSÕES

Com esse estudo, pode-se afirmar que um maior número de indivíduos na população de treinamento proporciona maior capacidade de predição do modelo genômico, sendo assim recomendam-se populações com no mínimo 300 indivíduos para aplicação da seleção genômica em *Eucalyptus* spp.

A estrutura de população apresenta um papel importante, para se otimizar as populações de treinamento, proporcionando maior eficiência dos modelos de predição, quando se tem indivíduos relacionados nas populações de estimação e validação.

A estratificação da população pelo método bayesiano apresenta maior eficiência no agrupamento de indivíduos relacionados, mostrando-se, assim, um bom critério para otimização das predições genômicas.

REFERÊNCIAS

AKDEMIR, D.; SANCHEZ, J. I.; JANNINK, J. L. Optimization of genomic selection training populations with a genetic algorithm. **Genetics, Selection, Evolution**, Paris, v. 47, n. 1, 2015. Disponível em: <https://gsejournal.biomedcentral.com/articles/10.1186/s12711-015-0116-6>. Acesso em: 10 mar. 2017.

ALBRECHT, T. et al. Genome-based prediction of testcross values in maize. **Theoretical and Applied Genetics**, Berlin, v. 123, n. 2, p. 339-350, 2011.

ASORO, F. G. et al. Accuracy and training population design for genomic selection on quantitative traits in elite North American Oats. **The Plant Genome Journal**, Madison, v. 4, n. 2, p. 132-144, July 2011.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, Austin, v. 186, n. 2, p. 713-24, Oct. 2010.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLoS ONE**, San Francisco, v. 3, n. 10, p. 1-8, 2008.

DONG, S. et al. Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. **Genome Research**, Cold Spring Harbor, v. 11, n. 8, p. 1418-1424, Aug. 2001.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, Irvine, v. 19, p. 11-15, 1987.

EARL, D. A.; VONHOLDT, B. M. Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. **Conservation Genetics Resources**, Berlin, v. 4, n. 2, p. 359-361, 2012.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R Package rrBLUP. **The Plant Genome Journal**, Madison, v. 4, n. 3, p. 250-255, 2011.

ENDELMAN, J. B. et al. Optimal design of preliminary yield trials with genome-wide markers. **Crop Science**, Madison, v. 54, p. 48-59, Jan./Feb. 2014.

ERBE, M. et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. **Journal of Dairy Science**, Champaign, v. 95, n. 7, p. 4114-4129, 2012.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software structure: a simulation study. **Molecular Ecology**, Oxford, v. 14, n. 8, p. 2611-2620, 2005.

GANAL, M. W. et al. 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, 2011. Disponível em: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028334>. Acesso em: 10 mar. 2017.

GANAL, M. W. et al. Large SNP arrays for genotyping in crop plants. **Journal of Biosciences**, Bangalore, v. 37, n. 5, p. 821-828, 2012.

GILMOUR, A. R. et al. **ASReml user guide release 3.0**. Hemel Hempstead: VSN International, 2009. Disponível em: <http://www.vsni.co.uk>. Acesso em: 3 nov. 2014.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, Dordrecht, v. 136, n. 2, p. 245-257, 2009.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, London, v. 10, n. 6, p. 381-391, 2009.

GRAPES, L. et al. Comparing linkage disequilibrium-based methods for fine mapping quantitative Trait Loci. **Genetics**, Austin, v. 166, n. 3, p. 1561-1570, 2004.

GRAPES, L. et al. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. **Genetics**, Austin, v. 172, n. 3, p. 1955-1965, 2006.

GRATTAPAGLIA, D. et al. Genomic selection for growth traits in Eucalyptus: accuracy within and across breeding populations. **BMC Proceedings**, London, v. 5, p. O16, 2011. Supplement 7.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, Heidelberg, v. 7, n. 2, p. 241-255, Oct. 2011.

GUO, Z. et al. The impact of population structure on genomic prediction in stratified populations. **Theoretical and Applied Genetics**, Berlin, v. 127, n. 3, p. 1-14, Mar. 2013.

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

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.

HARRIS, B. L.; JOHNSON, D. L. The impact of high density SNP chips on genomic evaluation in dairy cattle. **Interbull Bulletin**, Uppsala, v. 42, n. 42, p. 40-43, 2010.

HAYES, B. J. et al. Invited review: genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, Champaign, v. 92, n. 2, p. 433-443, 2009.

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

HESLOT, N. et al. Genomic selection in plant breeding: a comparison of models. **Crop Science**, Madison, v. 52, n. 1, p. 146-160, 2012.

ISIDRO, J. et al. Training set optimization under population structure in genomic selection. **Theoretical and Applied Genetics**, Berlin, v. 128, p. 145-158, 2015.

JAKOBSSON, M.; ROSENBERG, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics**, Oxford, v. 23, n. 14, p. 1801-1806, 2007.

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

JARQUÍN, D. et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. **BMC Genomics**, London, v. 15, n. 1, 2014. Disponível em: <https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-740>. Acesso em: 10 mar. 2017.

JIANG, G. Molecular markers and marker-assisted breeding in plants. In: ANDERSEN, S. B. (Ed.). **Plant breeding from laboratories to fields**. London: Intech, 2013. p. 45-83.

JOMBART, T. et al. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genetics**, London, v. 11, n. 1, 2010. Disponível em: <http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-11-94>. Acesso em: 10 mar. 2017.

KILIAN, A. et al. The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In: INTERNATIONAL CONGRESS "IN THE WAKE OF THE DOUBLE HELIX: FROM THE GREEN REVOLUTION TO THE GENE REVOLUTION", 2003, Bologna. **Proceedings…** Bologna, 2003. p. 443-461.

LAWSON, D. J. et al. Inference of population structure using dense haplotype data. **PLoS Genetics**, San Francisco, v. 8, n. 1, p. 11-17, 2012.

LEE, S. H. et al. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS Genetics**, San Francisco, v. 4, n. 10, Oct. 2008. Disponível em: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000231>. Acesso em: 10 mar. 2017.

LEGARRA, A. et al. Performance of genomic selection in mice. **Genetics**, Austin, v. 180, n. 1, p. 611-618, 2008.

LEHERMEIER, C. et al. Usefulness of multiparental populations of maize (*Zea mays* L.) for Genome-Based Prediction. **Genetics**, Austin, v. 198, n. 1, p. 3-16, Sept. 2014.

LIMA, B. M. de. **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 f. Thesis (Doctor in Genetics and Improvement of Plants)- Universidade de São Paulo, São Paulo, 2014.

LIN, Z. A.; HAYES, B. J. A.; DAETWYLER, H. D. A. Genomic selection in crops, trees and forages: a review. **Crop & Pasture Science**, Collingwood, v. 65, n. 11, p. 1177-1191, 2014.

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

LORENZ, A. J.; SMITH, K. P.; JANNINK, J. L. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. **Crop Science**, Madison, v. 52, n. 4, p. 1609-1621, 2012.

MANGIN, B. et al. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, Washington, v. 108, n. 3, p. 285-291, 2012.

MARRONI, F. et al. Nucleotide diversity and linkage disequilibrium in Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. **Tree Genetics and Genomes**, Heidelberg, v. 7, n. 5, p. 1011-1023, 2011.

MARULANDA, J. J.; MELCHINGER, A. E.; WÜRSCHUM, T. Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. **Plant Breeding**, Berlin, v. 134, n. 6, p. 623-630, 2015.

MASSMAN, J. M. et al. Genomewide predictions from maize single-cross data. **Theoretical and Applied Genetics**, Berlin, v. 126, n. 1, p. 13-22, 2013.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, n. 4, p. 1819-1829, Apr. 2001.

MÜLLER, B. S. F. et al. Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. **BMC Genomics**, London, v. 18, n. 1, p. 1-17, July 2017.

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

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

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Austin, v. 155, n. 2, p. 945-959, 2000.

PSZCZOLA, M. et al. Reliability of direct genomic values for animals with different relationships within and to the reference population. **Journal of Dairy Science**, Champaign, v. 95, n. 1, p. 389-400, 2012.

PURCELL, S. et al. 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.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R

Foundation for Statistical Computing, 2015. Disponível em: <http://www.r-project.org/>. Acesso em: 10 mar. 2015.

RAMOS, A. M. et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. **PloS ONE**, San Francisco, v. 4, n. 8, 2009. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2716536&tool=pmcentrez&rendertype=abstract>. Acesso em: 10 mar. 2017.

RAYMOND, C.; SCHIMLECK, L. Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in Eucalyptus globulus. **Canadian Journal of Forest Research**, Ottawa, v. 32, p. 170-176, 2002.

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

RESENDE, M. D. V. de et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, London, v. 194, p. 116-128, 2012.

RESENDE, M. D. V. de et al. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n. 56, p. 63-77, jan./jun. 2008.

RESENDE, M. F. R. et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytologist**, Cambridge, v. 193, n. 3, p. 617-624, Feb. 2012.

ROBERTSON, A.; HILL, W. G. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. **Genetics**, Austin, v. 107, n. 4, p. 703-718, 1984.

SANSALONI, C. P. et al. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. **Plant Methods**, London, v. 6, p. 16, 2010.

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

SILVA-JUNIOR, O. B.; FARIA, D. A.; GRATTAPAGLIA, D. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. **New Phytologist**, London, v. 206, n. 4, p. 1527-1540, June 2015.

SILVA-JUNIOR, O. B.; GRATTAPAGLIA, D. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. **New Phytologist**, London, v. 208, n. 3, p. 830-845, Nov. 2015.

SONG, Q. et al. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. **PLoS ONE**, San Francisco, v. 8, n. 1, p. 1-12, 2013.

SPINDEL, J. et al. Genomic Selection and Association Mapping in Rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS Genetics**, San Francisco, v. 11, n. 2, p. 1-25, 2015.

TAYEH, N. et al. Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. **Frontiers in Plant Science**, Minglin Lang, v. 6, n. 941, p. 1-11, 2015.

TECHNOW, F.; BÜRGER, A.; MELCHINGER, A. E. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. **G3 - Genes, Genomes, Genetics**, Bethesda, v. 3, n. 2, p. 197-203, 2013.

TECHNOW, F. et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. **Genetics**, Austin, v. 197, n. 4, p. 1343-1355, 2014.

UNTERSEER, S. et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. **BMC Genomics**, London, v. 15, n. 1, 2014. Disponível em: <http://www.biomedcentral.com/1471-2164/15/823>. Acesso em: 10 mar. 2017.

VALENTE, M. S. F. et al. Seleção genômica para melhoramento vegetal com diferentes estruturas populacionais. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 111, p. 1857-1867, 2016.

WIENTJES, Y. C. J.; VEERKAMP, R. F.; CALUS, M. P. L. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. **Genetics**, Austin, v. 193, n. 2, p. 621-631, 2013.

YU, K. et al. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. **Annals of Human Genetics**, London, v. 69, n. 5, p. 577-589, 2005.

ZHENG, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. **Bioinformatics**, Oxford, v. 28, n. 24, p. 3326-3328, 2012.