MÁRCIO FERNANDO RIBEIRO DE RESENDE JÚNIOR

SELEÇÂO GENÔMICA AMPLA NO MELHORAMENTO VEGETAL

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

VIÇOSA
MINAS GERAIS – BRASIL
2010

MÁRCIO FERNANDO RIBEIRO DE RESENDE JÚNIOR

SELEÇÃO GENÔMICA AMPLA NO MELHORAMENTO VEGETAL

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

APROVADA: 16 de abril de 2010.

Prof. Marcos Deon Vilela de Resende
(Co-orientador)

Prof. Fabyano Fonseca e Silva
(Co-orientador)

Profª. Eveline Teixeira Caixeta

Prof. José Marcelo Soriano Viana

Prof. Cosme Damião Cruz
(Orientador)

Se eu algum dia pude ver mais longe, foi por estar de pé sobre ombros de gigantes

(Sir Isaac Newton)

Ao meu pai Márcio, à minha mãe Ana, minhas irmãs Ju, Fê e Carol e à minha noiva Kelly, pelo apoio, amor, sacrifícios e abdicações,
Dedico.

AGRADECIMENTOS

Ao meu pai Márcio Fernando Ribeiro de Resende, minha mãe Ana Maria Lara de Resende e minhas irmãs Juliana, Fernanda e Ana Carolina pelo exemplo, pelo apoio, amizade e compreensão em todos os momentos.

À Universidade Federal de Viçosa, ao Departamento de Engenharia Florestal e ao Departamento de Biologia Geral pela oportunidade de realização do curso de graduação e de mestrado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro.

Aos orientadores Dr. Cosme Damião Cruz e Dr. Marcos Deon Vilela de Resende pelo exemplo de pesquisadores, pela amizade, orientação, pelos sábios ensinamentos responsáveis pelo meu crescimento e por estarem sempre disponíveis e atenciosos para prontamente me atenderem mesmo com várias outras obrigações e compromissos.

Ao Dr. Dario Grattapaglia, pelos ensinamentos, conselhos, amizade e pela confiança ao permitir que eu integrasse o grupo de pesquisa de GWS.

Aos estudantes de doutorado da UnB, César Daniel Petroli e Carolina Paola Sansaloni, que embora eu não os conheça, fizeram um trabalho árduo e brilhante na geração dos marcadores DArTs, material de suas dissertações e também utilizados nesta tese.

Ao professor Dr. Fabyano Fonseca da Silva pela coorientação, atenção e pelas valiosas contribuições ao trabalho.

Às empresas CENIBRA Papel e Celulose e FIBRIA Papel e Celulose por disponibilizarem o material de análise.

Aos amigos de laboratório que me apoiaram nesta caminhada: Leo, Edmar, Tatiana, Rafael, Rafael, Danielle, Caio, Felipe, Otávio, Tetsu, Eliel, Marcelo, Moysés, Lívia e Talles.

Aos amigos do GENMELHOR que estiveram sempre presentes nas organizões de eventos e cursos

Aos amigos da área de genética florestal: Alexandre, Ricardo e Leandro pelas horas compartilhadas em Viçosa, estudando ou descansando.

Aos amigos da Engenharia Florestal e da república que tornaram os anos em Viçosa mais divertidos: Leo, Fred, Vítor, Serjão, Guilherme, Ana, Bruna, Mila, Nanda, Douglas, Ricardinho, Gustavo, Gustavinho, Leo e Lu, Reno, Marcelão, Bernardo, demais amigos da ENF-2004 e outros que cruzaram meu caminho e que foram sempre companheiros nesta cidade.

À Edna e à Dona Rita, secretárias do programa de Pós-Graduação, sempre dispostas a ajudar

A todos os professores que já tive em toda a minha vida acadêmica, por terem contribuído com a minha formação e crescimento.

Ao meu País, Brasil, que financiou toda minha formação acadêmica.

E por último, mas não menos importante, à Kelly por tornar a minha vida mais feliz, pela companhia, compreensão e sobretudo, AMOR!...

Meu muito obrigado!!!

BIOGRAFIA

MÁRCIO FERNANDO RIBEIRO DE RESENDE JÚNIOR, filho de Márcio Fernando Ribeiro de Resende e Ana Maria Lara de Resende, nasceu em Belo Horizonte, MG, no dia dois de outubro de 1985.

Em 2004 ingressou no curso de Engenharia Florestal pela Universidade Federal de Viçosa graduando-se em janeiro de 2008.

Em março de 2008, iniciou o curso de Mestrado em Genética e Melhoramento na Universidade de Viçosa (UFV), submetendo-se à defesa de tese em abril de 2010.

Em março de 2010, obteve aprovação para cursar o doutorado na Universidade da Flórida – USA, sob a supervisão do professor Dr. Matias Kirst com início previsto para maio de 2010.

ÍNDICE

RESUMO

RESENDE JÚNIOR, Márcio Fernando Ribeiro, M.Sc., Universidade Federal de Viçosa, Abril de 2010. **Seleção genômica ampla no melhoramento vegetal.** Orientador: Cosme Damião Cruz. Coorientadores: Marcos Deon Vilela de Resende e Fabyano Fonseca e Silva.

A seleção genômica ampla (GWS) foi idealizada no ano de 2001 como forma de predizer o fenótipo futuro de uma população baseado em informações de marcadores moleculares, cujos efeitos genéticos aditivos, que estes explicam, já foram previamente estimados. Esta tecnologia já é pesquisada e integrada aos programas de melhoramento animal. Embora em plantas nenhum trabalho com dados reais tenha sido descrito, a GWS tem grandes perspectivas de utilização também no melhoramento genético vegetal, o que pode permitir melhores acurácias e seleção precoce. O objetivo deste trabalho foi, em um primeiro momento, fornecer subsídios para melhor entender a seleção genômica ampla e fazer uma comparação de sua utilização com marcadores dominantes e codominantes. Em uma segunda etapa, a aplicação dessa tecnologia foi então proposta em *Eucalyptus* e seu impacto foi avaliado no melhoramento florestal. Foram simuladas uma característica de controle oligogênico e outra controlada por muitos genes em diferentes situações de desequilíbrio de ligação com os marcadores. Em cada característica, o número de locos que controlava o caracter foi estabelecido entre 100, 200 e 400 e as herdabilidades entre 20%, 30% e 40%. Foi avaliada a correlação dos valores fenotípicos observados com os valores fenotípicos preditos via informação de marcadores e a acuáracia de seleção. A partir das estimativas de acurácia, calculou-se também o ganho de seleção por unidade de tempo comparado com a seleção fenotípica. Os resultados das simulações demonstraram altos valores de acurácias que proporcionaram ganhos de até 500% caso o tempo do ciclo de geração seja reduzido. Observou-se que se o número de marcadores dominantes disponíveis foi superior ao número de marcadores codominantes, essa maior densidade proporciona acurácias maiores. A segunda etapa do trabalho foi realizada em duas populações de Eucalipto utilizando marcadores dominantes DArTs e avaliando as características Altura total, Diâmetro a Altura do Peito (DAP) e penetração do Pilody. As acurácias

máximas obtidas foram de 0,67 para Altura e 0,69 para DAP em uma população, e de 0,53, 0,62, e 0,53 para Altura, DAP e Pilodyn, respectivamente, na segunda população. Estes valores proporcionaram ganhos que variaram entre 430% e 723% caso o ciclo de geração seja reduzido em 7 anos, situação possível no melhoramento de Eucalipto. Este trabalho demonstrou resultados animadores e o uso GWS se mostrou factível em plantas nas simulações e no conjunto de dados reais.

ABSTRACT

RESENDE JÚNIOR, Márcio Fernando Ribeiro, M.Sc., Universidade Federal de Viçosa, April 2010. **Genome wide selection in plant breeding.** Advisor: Cosme Damião Cruz. Co-Advisors: Marcos Deon Vilela de Resende and Fabyano Fonseca e Silva.

The Genome Wide selection was proposed in 2001 to predict the phenotype values based in molecular markers information. In a previus step, the effect of each of each marker in controlling the genetic variance is estimated. This technology has already been used in animals, however, no report of its use was made for plants. This work aimed to study the impact of GWS, first in a simulated dataset, then in two *Eucalyptus* populations. Besides that, on the simulated data, it was compared the efficiency of using dominant markers versus the use of codominant ones. The simulation generated one population controlled by many genes (polygenic) and one population with olligogenic control. There were different situations of linkage disequilibrium among the marker and the QTL, different number of markers controlling the trait and heritabilities of 20, 30 and 40%. It was evaluated the prediction ability and the accuracy of the GWS. The results of accuracy were high, which turn in to a selection gain of up to 500% in the selection dataset and the use dominant markers at higher densities is more efficient than the use of dominant markers with lower densities. The *Euvalyptus* populations were genotyped for total height, diameter at breast height (DBH) and Pilodyn. The values of accuracy were 0,67 for height and 0,69 for DBH in the first population and ,53, 0,62, and 0,53 for total height, DBH and Pilodyn respectively. Those result turn in to a selection gain that varied from 430% to 723% with a reduction of 7 years in the breeding cycle. This showed significant results and gave evidences that the use of GWS in plants is possible to improve the way plant breeding is actually performed.

INTRODUÇÃO GERAL

O melhoramento genético, associado às pesquisas de práticas agrícolas e silviculturais como o manejo do solo, controle de pragas e doenças dentre outros, proporcionou avanços consideráveis para a agricultura e para o setor florestal. Embora estas tecnologias continuam, ainda hoje, a proporcionar ganhos para as culturas, o advento da biotecnologia gerou expectativas de aumentar ou maximizar estas melhorias dos materiais para as diferentes características de interesse.

A partir de pesquisas de biologia molecular e genômica, foram identificados marcadores genéticos com potencial de aplicação na localização de regiões genômicas que controlam características de interesse (QTLs). Estes marcadores podem ser utilizados para elucidar a arquitetura genética de caracteres complexos em plantas via mapeamento genético e detecção de QTLs. Sua aplicação foi vislumbrada para auxiliar os procedimentos de seleção no melhoramento convencional, o que foi chamado de melhoramento genômico.

No entanto, até os dias atuais, a operacionalização do uso de marcadores moleculares se deu principalmente em medidas de controle de qualidade para detecção de cruzamentos contaminantes e proteção varietal, uma vez que essas aplicações requerem um baixo número de marcadores disponíveis, ao contrário do melhoramento genômico que requer a genotipagem de um grande número de marcadores para que este seja efeciente.

Para que esta integração se torne viável e eficiente, vários foram os esforços de pesquisa que culminaram com o desenvolvimento de tecnologias de genotipagem em larga escala a um custo muitas vezes inferior ao inicialmente proposto e um método de utilização destes marcadores conhecido como seleção genômica ampla. Com isso, o panorama atual demonstra perpectivas da integração dos marcadores moleculares nos programas de melhoramento genético vegetal, fato que já vem sendo feito no setor animal.

O objetivo deste trabalho foi, em um primeiro momento, fornecer subsídios para melhor entender a seleção genômica ampla, técnica que permite a seleção de indivíduos baseado apenas na informação dos seus marcadores. Em uma segunda etapa, a aplicação dessa tecnologia foi então proposta em *Eucalyptus*,. e o impacto de seu uso no melhoramento florestal foi avaliado.

CAPÍTULO

SELEÇÂO GENÔMICA AMPLA VIA DADOS SIMULADOS

VIÇOSA

MINAS GERAIS – BRASIL

2010

1. INTRODUÇÃO

O melhoramento genético tem, a vários anos, proporcionado, com muito sucesso, o aumento de produtividade e a melhoria de várias outras características de interesse na agricultura e na pecuária. Embora muitos métodos surgiram ao longo dos anos, a estratégia básica utilizada foi a de predizer o valor genético do indivíduo, baseado em informações fenotípicas e em alguns casos de genealogia. No entanto, com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, criou-se a expectativa de que as informações genotípicas dos marcadores moleculares, uma vez correlacionados com características fenotípicas de interesse, pudessem ser amplamente utilizadas na obtenção e seleção de indivíduos com maior valor genético. Esta técnica ficou conhecida como seleção assistida por marcadores moleculares (*MAS - Marker Assisted Selection*).

Com a perspectiva de um aumento nos ganhos de seleção e redução nos ciclos de melhoramento via seleção assitida por marcadores, muitas pesquisas foram feitas e QTLs foram detectados e mapeados nas mais variadas culturas (Frary, *et. al.* 2000; Yano *et. al.* 2000; Takahashi *et. al.* 2001; El-Din El-Assal *et. al.* 2001; Liu *et. al,* 2002). No entanto, grande parte destes QTLs detectados e mapeados em cada espécie, não foram aplicados de forma prática nos seus programas de melhoramento (Bernardo *et al.* 2008)

As principais causas deste insucesso foram a necessidade do estabelecimento de associações entre os marcadores e os QTLs para cada família avaliada e o fato de serem feitas apenas a detecção de um pequeno número de QTLs de grande efeito, os quais, devido à natureza poligênica e à alta influência ambiental dos caracteres quantitativos, não explicam suficientemente toda a variação genética (Dekkers, 2004). Além disso, pode se destacar, também, que a seleção baseada em marcadores moleculares só é superior em relação á seleção fenotípica quando esta é aplicada em uma família com tamanho superior a 500 indivíduos. (Resende, 2007)

A partir do início do século XXI, os avanços de tecnologias de genotipagem em larga escala, a descoberta de novos marcadores como os SNPs e a automação do processo de genotipagem de marcadores (Jenkins and Gibson, 2002) proporcionaram a redução do preço por *data point* e permitiram que um grande número de marcadores fosse usado para várias culturas.

Uma vez gerado um grande número de marcadores espalhados por todo o genoma de um indivíduo, alguns destes marcadores estarão muito perto do QTL e em desequilíbrio de ligação (LD) com este (Hastbacka *et. al.,* 1994). O conceito de desequilíbrio de ligação refere-se à associação não aleatória entre dois genes ou entre um QTL e um loco marcador. Quando as freqüências alélicas e genotípicas de um ou mais locos autossômicos são constantes de uma geração para a outra e as freqüências genotípicas são determinadas pelas freqüências alélicas, diz se que este loco se encontra em equilíbrio de ligação. Com a ligação gênica, dois genes ligados apresentam uma associação que não se dá ao acaso e estão em desequilíbrio de ligação. Com os eventos de recombinação, a cada nova geração, os locos tendem a uma situação de equilíbrio, e o tamanho de um dado segmento cromossômico que contém dois locos quaisquer e que não sofreu recombinação diminui, o que consequentemente reduz o LD. Por essa razão, o uso eficiente de marcadores moleculares para auxiliar o melhoramento genético requer grande número, para que, mesmo em uma população que já passou por várias gerações e sucessivas recombinações históricas, exista um marcador tão próximo do QTL que apresente uma associação com este que não tem razão aleatória.

Uma vez que um marcador se encontra em LD com o QTL, alguns alelos destes marcadores, estarão correlacionados com efeitos positivos dos QTLs em todas as famílias e podem ser utilizados sem que seja necessário o estabelecimento da fase de ligação em cada família. (Meuwissen *et.al.,* 2001). Além disso, características quantitativas são controladas por muitos genes e para que se tenha grande parte da variação genética explicada pelos marcadores moleculares, é preciso obter um marcador em desequilíbrio de ligação com cada loco controlador da característica quantitativa.

A seleção genômica ampla (*Genome Wide Selection* - GWS), proposta inicialmente no ano de 2001 por Meuwissen *et. al.* (2001), consiste na análise de

um grande número de marcadores amplamente distribuídos no genoma. Após a obtenção destes marcadores, seus efeitos são estimados baseados em dados fenotípicos de uma população conhecida como população de estimação. Uma vez que seus efeitos são modelados e estimados, estes são testados em uma população de validação e ,então, seleciona-se os marcadores que explicam parte da variância genética do caráter em estudo para que sua informação seja efetivamente incorporada à etapa de seleção do programa de melhoramento.

Na população de validação, utiliza-se um conjunto de dados menor do que aquele da população de estimação e contempla indivíduos genotipados e fenotipados para a característica de interesse. Esta amostra independente é utilizada para testar e verificar as acurácias das equações de predição de valores genéticos genômicos. Para computar essa acurácia, os valores genéticos genômicos são preditos (usando os efeitos estimados na população de estimação) e submetidos a análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos marcadores, os erros nas estimativas dos valores genéticos genômicos e dos valores fenotípicos são independentes e toda covariância entre esses valores é de natureza genética (Resende, 2007, 2008).

O ponto chave da análise destes marcadores é a estimação de seus efeitos, uma vez que o número de parâmetros que precisam ser estimados é muito superior ao número de observações fenotípicas disponíveis. Vários métodos de predição de valores genéticos genômicos foram propostos, dos quais se pode destacar o de quadrados mínimos, BLUP/GWS, BayesA e BayesB (Meuwissen et al., 2001), aprendizado de máquina (AM de Long et al., 2007), regressão RKHS (*Reproducing Kernel Hilbert Spaces*) (Gianola et al., 2008), LASSO Bayesiano (de los Campos, 2009), Bayes C (Gredler *et al.*, 2009), Bayes B Acelerado (Meuwissen, 2009), Regressão via Quadrados Mínimos Parciais (PLSR) (Solberg et al., 2009) e Regressão via Componentes Principais (PCR) (Solberg et al., 2009).

A GWS tem sido amplamente pesquisada e desenvolvida para aplicação na pecuária sendo que vários estudos analisaram as perspectivas da Seleção Genômica avaliando parâmetros e procedimentos analitcos que influenciam na

predição dos valores genômicos e qual seriam o impacto de seu uso no melhoramento animal (Calus et al. 2008; Dekkers 2007; Long et al. 2007; Muir 2007; Schaeffer 2006; Solberg et al. 2008). Em plantas, esta técnica tem também possibilidades de grandes impactos e revisões e pesquisas recentes, com dados simulados, demonstraram excelentes resultados na utilização desta tecnologia em diferentes culturas (Bernardo & Yu 2007; Resende *et. al.* 2008; Wong & Bernardo 2008; Heffner et al. 2009; Zhong et al. 2009). No entanto, até o presente momento, nenhum trabalho com o uso da seleção genômica foi descrito com aplicação em dados reais de culturas vegetais.

No melhoramento animal, os marcadores moleculares utilizados para as pesquisas e aplicações da GWS são os polimorfismos de base única, SNPs, uma vez que um grande número destes marcadores já está disponível e são comercializados em painéis de vários milhares de marcas (Matukumalli *et. al.*, 2009).

Recentemente, um novo tipo de marcador conhecido como DArTs (*Diversity Array Technology*) foi desenvolvido com perspectivas para aplicação em análises de seleção genômica ampla. Este marcador apresenta como característica a genotipagem em larga escala baseada em análises de micro arranjo, o baixo custo de desenvolvimento e de genotipagem e sua natureza dominante (Wenzl *et al.,* 2004). Para cada amostra de DNA estudada, são desenvolvidas representações genômicas através da digestão do DNA genômica com enzimas de restrição. Os framentos amplificados por PCR são clonados e arranjados em um suporte sólido (microarranjo) resultando em um arranjo de descoberta*.* Representações genômicas preparadas a partir de genomas individuais alvo de estudo são hibridizados a esse arranjo de descoberta para identificação de polimorfismos. Os clones polimórficos (marcadores DArT) mostram intensidade de sinais de hibridização variáveis para diferentes indivíduos conforme os diferentes genótipos (Jaccoud *et al.,* 2001).

Este marcador tem se mostrado como uma alternativa disponível para a aplicação de seleção genômica em culturas vegetais que não têm painéis de SNPs desenvolvidos, como é o caso do Eucalipto. Os DArTs podem reduzir as implicações de uma das limitações para a aplicação prática da seleção genômica

que é o custo desde o desenvolvimento de um grande número de marcadores até a genotipagem de um grande número de indivíduos. No entanto, nenhuma pesquisa foi encontrada utilizando este tipo de marcador, principalmente devido ao fato de as pesquisas de GWS na área vegetal ainda serem limitadas e de que as pesquisas animais se concentram em marcadores SNPs. Assim, o objetivo deste trabalho foi fazer uma primeira avaliação, via dados simulados, sobre o uso da seleção genômica ampla e sua perspectiva no melhoramento de plantas, avaliando a acurácia da GWS utilizando marcadores dominantes e codominantes.

2. MATERIAL E MÉTODOS

Nos estudos genéticos, caracteres quantitativos são, em geral, regulados por vários genes com pequena magnitude de efeitos, ao passo que os caracteres qualitativos são controlados por alguns poucos genes de maior efeito.

As estratégias de melhoramento genético aplicada a características quantitativas como produção de grãos e volume de madeira podem ser diferentes das estratégias utilizadas no melhoramento de características oligogênicas como a resistência a doenças. Dessa forma, para estudar o impacto da seleção genômica ampla em características de controle poligênico e oligogênico, foram realizadas duas simulações, uma com magnitude dos efeitos genéticos semelhantes e outra com alguns poucos genes de maior efeito reponsáveis por grande parte da variação genética da característica

2.1 SIMULAÇÃO DOS DADOS GENOTÍPICOS E FENOTÍPICOS

Para proceder as análises de seleção genômica, foram simulados dados genotípicos e fenotípicos considerando ausência de dominância, diferentes herdabilidades, desequilíbrios de ligação, número de locos controlando a característica e tipos de marcadores.

A simulação dos dados foi feita de dois modos diferentes. Em ambas as situações, o caráter quantitativo foi simulado em três situações, considerando dois alelos por loco, efeitos aditivos e o controle por 100, 200 e 400 locos. No primeiro caso, o efeito do alelo favorável em cada loco foi simulado como sendo

$$a_L = \frac{(L-1)}{(L+1)}$$

em que L refere-se ao L-ésimo QTL. (Lande e Thompson ,1990 e Bernardo e Yu, 2007) (Figura 1). O efeito do alelo menos favorável foi tomado como $-a_L$. Foram simulados 1000 indivíduos com fenótipos gerados segundo o modelo $f = g+e$, em que g são os efeitos genéticos totais dados pelo somatório dos efeitos genéticos em cada loco e e são os efeitos ambientais, gerados segundo uma distribuição

normal com média 10 e variância compatível com as três herdabilidades testadas: 20%, 30% e 40% de acordo com a fórmula:

$$\sigma_M^2 = \frac{\sigma_g^2(1-h^2)}{h^2},$$

em que σ_M^2 é a variância ambiental, σ_g^2 a variância genética e h² a herdabilidade testada

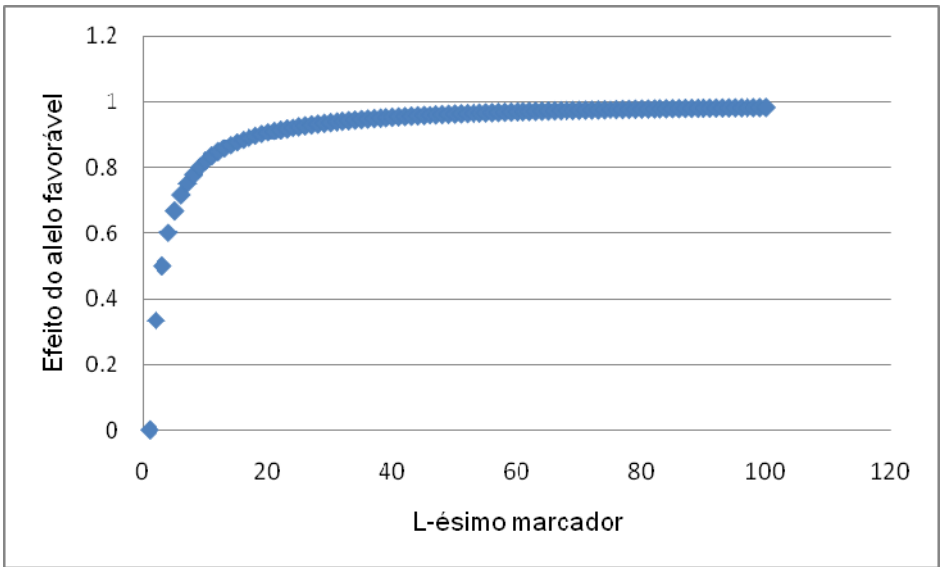

Figura 1 – Distribuição dos efeitos dos alelos favoráveis de cada marcadore simulados de acordo com Lande e Thompson (1990).

No segundo caso, a simulação foi feita considerando novamente as três situações de 100, 200 e 400 locos. No entanto, os efeitos foram simulados de acordo com a expressão

$$a_L = \frac{1}{L(L+1)}$$

em que L refere-se ao L-ésimo QTL. A distribuição dos efeitos a_L, neste caso, se assemelha mais a uma situação biológica de característica oligogênica e qualitativa, uma vez que, embora a característica seja controlada por vários genes, existem alguns poucos QTLs de maior efeito e a grande maioria tem efeitos próximos de zero como pode ser observado na figura 2, O número de indivíduos, o modelo para simulação dos fenótipos e as herdabilidades testadas foram as mesmas da primeira simulação.

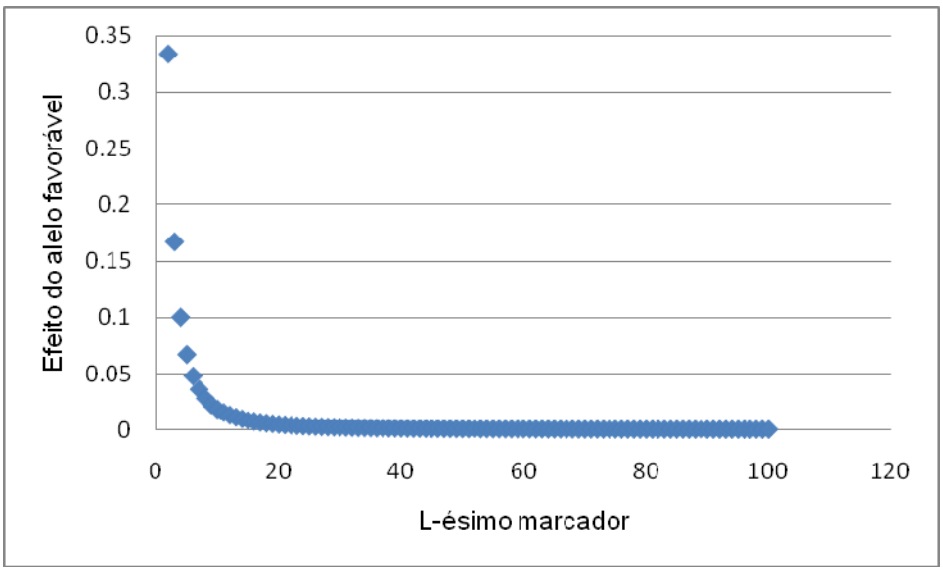

Figura 2 – Distribuição dos efeitos de “a” de acordo com a segunda simulação.

Em cada indivíduo, genótipos do tipo A_1A_1, A_1A_2, ou A_2A_2 foram sorteados aleatoriamente em cada loco e codificados como 0, 1 e 2, respectivamente. O número 1 dos genótipos do tipo A_1A_1 e A_1A_2 denotam o alelo desfavorável e o número 2 denota o alelo favorável.

Para avaliar o efeito da densidade de marcadores utilizados e do nível de desequilíbrio de ligação (LD) obtido entre o QTL e pelo menos um marcador foram simuladas quatro situações de LD igual a 0,5; 0,7; 0,9 e 1,0, Nesta simulação, o objetivo não foi representar o LD médio da população e sim uma densidade hipotética de marcadores capaz de explicar 50, 70, 90 e 100% dos QTLs controladores da característica. Dessa forma, em cada situação, o valor de LD simbolizou a porcentagem de locos marcadores que estavam em desequilíbrio de ligação com um QTL e a proprorção da variância genética que estava sendo explicada pelos marcadores. Nas situações em que o loco marcador estava em desequilíbrio de ligação, foi considerado que este marcador é efetivamente um QTN (Quantitative Trait Nucleotide) estando diretamente ligado ao QTL e tendo efeito igual ao do QTL. Na situação oposta, em que foi observado equilíbrio de ligação do marcador com o QTL, o genótipo do marcador teve seu efeito atríbuido como zero. Sendo assim, considerando o genótipo A_1A_1 do QTL, teve seu efeito associado ao genótipo M_1M_1 do marcador quando estes se encontravam em LD.

Neste caso, o efeito do marcador M_1M_1 foi o mesmo do efeito do QTL. Na situação contrária, o genótipo A_1A_1 teve seu efeito aleatoriamente associado a qualquer um dos genótipos M_1M_2, ou M_2M_2 representando uma associação aleatória, conceito de equilíbrio de ligação.

Embora em uma situação prática para uma aplicação eficiente da Seleção Genômica Ampla o número de marcadores necessários seja grande e superior ao número de QTLs que controlam a característica, nesta simulação, o número de marcadores simulados foi fixado como o número de locos utilzados em cada cenário para explicar a característica quantitativa. No entanto, a avaliação da densidade de marcadores utilizada pode ser extrapolada pelo nível de LD considerado, ou seja, uma situação de LD igual a 0,5 será obtida quando um pequeno número de marcadores for utilizado de modo a apenas 50% deles se encontrar em LD com o QTL. Quando muitos marcadores forem utilizados, será possível obter uma marca em LD com cada QTL (LD = 1).

A simulação dos marcadores dominantes foi realizada da mesma maneira que o método descrito acima. No entanto, ao final da simulação, todos os genótipos de código 2 (genótipo A_2A_2) foram substituídos pelo código 1 de maneira que o conjunto de dados dispôs apenas de 2 classes, 0 e 1, e a última com as informações confundidas do genótipo heterozigoto (A_1A_2) e do genótipo homozigoto para o alelo 2.

2.2 METODOLOGIA DE ANÁLISE

A partir dos dados fenotípicos gerados, foram estimados os efeitos de cada um dos locos marcadores que somados, compõem o valor genético genômico predito de cada indivíduo. Em situações práticas, o número de marcadores genotipados é maior que o número de fenótipos avaliados. A estimação dos efeitos do elevado número de parâmetros (locos) a partir de um número limitado de dados fenotípicos conduz ao problema da estimação por quadrados mínimos com número insuficiente de graus de liberdade para ajustar todos esses efeitos simultaneamente. Assim, os efeitos foram preditos por meio do procedimento BLUP/GWS que permite ajustar todos os efeitos alélicos simultaneamente.

O seguinte modelo linear misto geral foi usado para estimar os efeitos dos marcadores, conforme Resende (2008):

$y = Xb + Zh + e,$

em que *y* é o vetor de observações fenotípicas, *b* é o vetor de efeitos fixos, *h* é o vetor dos efeitos aleatórios dos marcadores e *e* refere-se ao vetor de resíduos aleatórios. *X* e *Z* são as matrizes de incidência para *b* e *h*. A estrutura de médias e variâncias no modelo em questão é definida como:

$$h \sim N(0,\ G) \qquad E(y) = Xb$$

$$e \sim N(0,\ R = I\sigma_e^2) \qquad Var(y) = V = ZGZ' + R$$

$$G = \sum_i^n I\sigma_{gi}^2 = I\sigma_g^2$$

As equações de modelo misto genômicas para a predição de *h* via o método BLUP/GWS equivalem a:

$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{h} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$ em que σ_g^2 refere-se à variância genética total do caráter e σ_e^2 é a variância residual. O valor genético genômico global do indivíduo *j* é dado por $VGG = \hat{y}_j = \sum_i Z_i \hat{h}_i$,

em que Z_i equivale a 0, 1 ou 2 para os genótipos A_1A_1, A_1A_2 e A_2A_2, respectivamente, para marcadores bialélicos e codominantes. Para marcadores dominantes, A_1A_2 e A_2A_2 ficam confundidos. Neste caso, foram testados quatro valores para um peso médio constante k que foi inserido na matriz Z para tentar explicar melhor a estrutura dos dados genotípicos. Dessa forma, a matriz Z quando marcadores dominantes foram utilizados foi composta pelos valores 0 e k. Os 4 valores de k testados foram: (1) k assumiu um valor de 1 o que seria equivalente a considerar que todos os indivíduos que receberam código 1 possuiam genótipo heterozigoto, (2) k assumiu o valor de 1,33 que seria o peso médio a se inserir na matriz Z em uma situação de equilíbrio de Hardy e Weinberg, dado por 0,33 x 2 + 0,66 x 1, (3) k assumiu o valor de 1,5, em que foi considerado uma situação intermediária, e (4) k assumindo o valor de 2, equivalente à dizer

que todos os indivíduos da população com o código na matriz Z diferente de zero eram na realidade homozigotos do tipo 22.

As equações de predição apresentadas acima assumiram *a priori* que todos os locos explicam iguais quantidades da variação genética. Assim, a variação genética explicada por cada loco é dada por σ_g^2 / n, em que σ_g^2 é a variação genética total e *n* é o número de marcadores utilizados em cada um dos cenários com os três tipos de seleção de marcadores testados. Essa estratégia foi adotada por Meuwissen et al. (2001), Muir (2007), Bernardo (2007) e Kolbehdari et al. (2007).

Na predição dos efeitos aleatórios via BLUP/GWS, não há necessidade de uso da matriz de parentesco (Schaeffer, 2006). A matriz de parentesco baseada em *pedigree* usada no BLUP tradicional foi substituída pela própria matriz Z´Z que é uma matriz de parentesco estimada pelos marcadores.

A seleção genômica ampla requer o uso de uma população de estimação para estimar os efeitos dos marcadores e uma população de validação, para analisar a eficiência da estimação destes efeitos na recuperação do valor genômico em uma população independente. Após a validação e estimação de um conjunto de marcadores, estes serão utilizados em uma população que dispõe apenas de genotipagem (Figura 3).

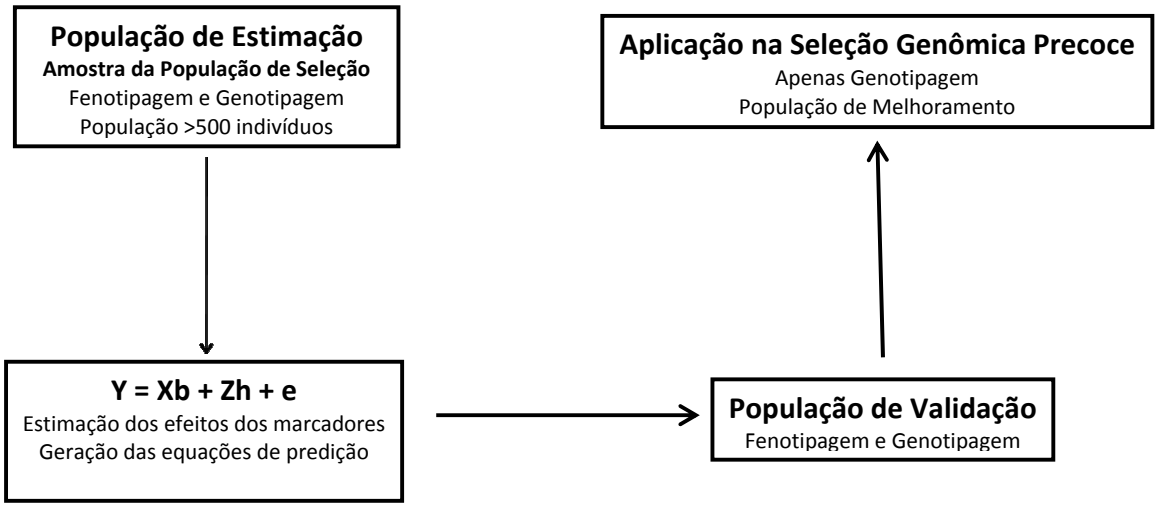

Figura 3 – Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético.

Cada população de 1000 indivíduos foi separada em 10 grupos de 100 indivíduos. A análise foi repetida 10 vezes de acordo com um procedimento *Jacknife* de maneira que em cada repetição, um grupo era removido do conjunto de dados para compor a população de validação e os outros 900 indivíduos eram utilizados na estimação dos valores genômicos preditos. Uma vez estimados todos os efeitos, estes eram aplicados na população de validação para predizer o valor genômico e somados à média geral estimada para compor o fenótipo realizado na população de validação. Uma vez que em cada repetição, um grupo era utilizado na validação, ao final da análise obteve-se o fenótipo realizado para todos os 1000 indivíduos, no entanto, sem que isso comprometesse a independência necessária entre a estimação e a validação.

2.3 NÚMERO DE MARCADORES UTILIZADOS NA GWS

Quando um grande número de marcadores são genotipados, vários destes não apresentam nenhuma associação com nenhum QTL. Em um primeiro estudo, é necessário a genotipagem de um máximo de marcadores possível, uma vez que a informação prévia da associação ou não destes com o QTL não é conhecida. No entanto, após uma primeira análise de associação e de estimação dos efeitos, o pesquisador pode selecionar apenas marcadores que tiveram sua associação detectada por algum teste estatístico específico.

2.3.1 UTILIZAÇÃO DE TODOS OS MARCADORES

A escolha do número de marcadores selecionados para estimação de seus efeitos foi feita de diferentes formas. Em um primeiro cenário, todos os marcadores genotipados foram utilizados na análise, independente de seus efeitos estarem associados ou não ao QTL. Neste caso, todos os efeitos dos marcadores são estimados, o que acarreta a estimação de um número maior de parâmetros.

2.3.2 TESTE DE ASSOCIAÇÃO VIA MARCA SIMPLES E TESTE F.

Em uma segunda abordagem, os marcadores foram préviamente testados quanto a associação da característica com a marca por um teste estatístico. Para isso, foi avaliada a associação via regressão em marcas simples. O seguinte modelo de regressão em marcas simples foi empregado para testar a associação entre marcador e QTL, conforme Resende (2008).

Y = 1u + Xm +e

Em que y é o vetor de observações fenotípicas, 1 é um vetor com valores 1, u é o escalar referente à média geral, m é o efeito fixo do marcador e e refere-se ao vetor de resíduos aleatórios. X é a matriz de incidência para m. As equações de quadrados mínimos para a estimação dos efeitos da média geral e do marcador equivalem a:

$$\begin{bmatrix} 1'1 & 1'X \\ X'1 & X'X \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} 1'y \\ X'y \end{bmatrix}$$

A hipótese da nulidade, ou seja, de que o marcador não apresenta qualquer efeito sobre o caráter, pode ser avaliado pelo teste F cuja estatística é calculada de acordo com a fórmula:

$$F = \frac{QMRegressão}{\hat{\sigma}_e^2} = \frac{\hat{m}X'y + \hat{u}1'y - (1/n)(1'y)^2}{(y'y - \hat{m}X'y - \hat{u}1'y)/(n-2)}$$

A hipótese alternativa é a de que o marcador afeta o caráter, ou seja, de que o marcador e QTL encontram-se em desequilíbrio de ligação e existe uma associação entre eles. A comparação com o valor tabelado de F foi feita a um nível de significância de 40%. Os marcadores que demonstraram associação com o QTL e foram considerados significativos pelo teste F, foram selecionados para as análises de estimação de seus efeitos.

2.3.3 CRITÉRIO DA TAXA DE FALSA DESCOBERTA (FDR).

Uma das questões a serem consideradas quando se realizam testes conjuntos para várias marcas moleculares é o estabelecimento do nível crítico de significância. Neste caso, o nível nominal de significância adotado para cada teste não corresponde àquele realizado em todo o experimento.

O terceiro cenário considerado foi o de avaliar o modo de seleção dos marcadores a terem seus efeitos estimados baseado no teste F e no critério da taxa de descobertas falsas (FDR). Este critério propõe controlar a razão de falsas descobertas, ou seja, a proporção de hipóteses nulas verdadeiras entre as hipóteses nulas rejeitadas. (Benjamini e Rochberg, 1995; Cruz, 2000).

2.4 AVALIAÇÃO DOS DADOS

‘ Ao estimar os efeitos de cada marcador, o fenótipo dos indivíduos são preditos, multiplicando a matriz de incidência Z, que corresponde aos genótipos dos marcadores para a população, pelos efeitos estimados de cada marcador e somando à média geral estimada. Como na validação dos resultados, o valor fenotípico é conhecido, a Seleção Genômica e cada situação testada foram avaliadas ao calcular a correlação do valor genético predito com o fenótipo observado nos 1000 indivíduos. Esta correlação é conhecida como capacidade preditiva ($r_{y\hat{y}}$) da seleção genômica em estimar os fenótipos e ela é dada teoricamente pela acurácia de seleção ($r_{g\hat{g}}$) multiplicada pela raiz quadrada da herdabilidade individual (h) ou, em outras palavras, $r_{y\hat{y}} = r_{g\hat{g}}h$ (Resende, 2008). Assim, foi isolado dessa equação o valor de $r_{g\hat{g}}$ ao dividir a correlação pela raiz quadrada da herdabilidade e foi também avaliada a acurácia de seleção, removendo a influência da herdabilidade na capacidade de predição.

A seleção genômica foi ainda comparada com a seleção fenotípica quanto ao ganho de seleção por unidade de tempo. Os únicos parâmetros variáveis no calculo do ganho de seleção ao comparar a seleção genômica com a fenotípica são a acurácia de seleção e, possívelmente, o tempo de seleção. Assim, as

acurácias obtidas pela GWS foram comparadas com o valor máximo de acurácia possível de se obter via seleção fenotípica (0,68 para a seleção de indivíduos em testes de progênie e herdabilidade igual a 0,2) (Resende, 2002). A relação foi avaliada considerando a expectatica de redução do tempo de geração em ½ e ¾ ao utilizar a seleção precoce via dados genotípicos.

2.5 FERRAMENTAS DE ANÁLISES

Todas as análises, tanto de simulação quanto de estimação foram desenvolvidas e serão futuramente implementadas como um pacote do *software* R, que é uma linguagem e um ambiente distribuído gratuitamente e utilizado para computação estatística e elaboração de gráficos.

Além disso, foi implementado no R (R Development Core Team, 2009) três técnicas de análise Bayesiana: BayesA, BayesB (Meuwissen et al., 2001) e Bayes B Acelerado (Meuwissen, 2009). No entanto, estes métodos, da forma como foram implementados necessitam de uma grande demanda computacional não sendo possível a análise dos diferentes cenários e sendo necessárias melhorias na programação para que estas se tornem viáveis do ponto de vista prático.

3. RESULTADOS E DISCUSSÃO

3.1 CARACTERÍSTICA SIMULADA COM LOCOS DE MAIOR EFEITO (OLIGOGÊNICA)

As características simuladas foram analisadas quanto ao número de marcadores significativos pelo teste de associação entre o marcador e a característica fenotípica, quanto à capacidade preditiva da seleção genômica (medida via correlação) e quanto à acurácia de seleção.

3.1.1 NÚMERO DE MARCADORES SIGNIFICATIVOS EM LD IGUAL A 1

Ao avaliar o impacto da Seleção Genômica no melhoramento de uma característica controlado por locos de maior efeito, como pode ser o caso da resistência de plantas à doenças, avaliou-se: O número de marcadores em que foi possível detectar associação com o QTL, ou seja que eram significativos pelo teste estatístico F, as acurácias obtidas via seleção baseada nos marcadores e os ganhos de seleção por unidade de tempo comparados com a seleção fenotípica.

Pôde se observar, quando o desequilíbrio de ligação considerado foi igual a um, que o número de marcadores significativos avaliados pela estatística F a um nível de significância de 40% variou entre 40 a 48% do total de marcadores codominantes utilizados. A mesma amplitude de variação pode ser observada quando foram utilizados marcadores do tipo dominante. (Tabela 1)

Quando o critério da taxa de falsa descoberta (FDR) foi utilizado o número de marcadores significativos, variou, para ambos os tipos de marcadores, de 6 a 9 % quando foram avaliados 100 locos marcadores, de 2 a 3% quando foram avaliados 200 locos, e de 1 a 3% quando 400 locos marcadores foram utilzados na genotipagem. (Tabela 1)

Tabela 1 – Número médio de marcadores significativos em cada um dos tipos de marcadores (Dominante e Codominante) e utilizando apenas o teste F e o critério de FDR. Desequilíbrio de ligação foi igual a 1

Número de Locos	Herdabilidade (h^2)	Marcador Codominante Tipo: SNP		Marcador dominante Tipo: DARTs	
		Teste F	FDR	Teste F	FDR
100	0,2	46,1	9,8	44,9	6,7
	0,3	42,2	6,4	41,2	8
	0,4	47,9	6,8	47,3	9,6
200	0,2	91,6	6,2	93,6	5,4
	0,3	82,6	7,5	86,5	4,7
	0,4	89,6	6,5	87,1	6,9
400	0,2	163,1	11,6	160,6	15,9
	0,3	168,1	5,9	161,8	7,1
	0,4	161,5	11,3	165	11,6

É possível observar na tabela 1 que o número de marcadores significativos em cada cenário de 100, 200 e 400 locos simulados foi aproximadamente constante em todas as herdabilidades e proporcional ao número de locos. De maneira análoga, observou-se que o teste FDR detectou em todos os cenários, um número de marcadores significativos constantes e próximos de dez em todos os cenários. A razão destes valores constantes é o fato da simulação ter uma distribuição que se assemelha a uma característica oligogênica. Assim, a soma de 50% dos locos de menor efeito representam apenas 2% da variação genotípica para 100 locos, 1% para 200 locos, e 0,5% para os 400 marcadores simulados. O poder dos testes de associação para detectar um QTL pelo uso de marcadores é a probabilidade de que o experimento rejeitará corretamente a hipótese de nulidade quando um QTL realmente existe na população. De acordo com Hayes (2009), dois dos fatores, dentre outros que influenciam este poder, são a proporção da variância fenotípica total explicada pelo QTL e o número de observações fenotípicas. Embora, mesmo na situação de desequilíbrio de ligação igual a um,

em que os marcadores moleculares são simulados diretamente ligados aos QTLs, a associação não é detectada devido à magnitude dos efeitos.

A mesma explicação pode ser extrapolada para a significância via o critério de FDR. Dessa forma, os 10 primeiros locos representaram 85, 84 e 83% da variação genética nas situações de 100, 200 e 400 locos, respectivamente. Como a razão de falsa descoberta considera um nível de significância mais rigoroso e pela proporção da variância total explicada pelos primeiros locos ser aproximadamente a mesma nos três casos, um número menor, e aproximadamente constante de marcas foram detectadas como significativas.

No entanto, embora os marcadores SNPs, por serem codominantes, são mais informativos, o teste de significância via regressão em marca simples foi capaz de detectar a associação em aproximadamente o mesmo número de marcas do tipo DArTs. Este foi considerado como o primeiro indício da aplicabilidade dos marcadores dominantes nos estudos de associação.

3.1.2 NÚMERO DE MARCADORES SIGNIFICATIVOS EM LD DIFERENTE DE 1

Ao analisar os cenários cuja população simulada apresentava um desequilíbrio de ligação com os marcadores inferior a um, observou-se uma redução no número de marcadores significativos. Este número variou de 36 a 45% do total de marcadores utilizados do tipo codominante. Ao avaliar os marcadores dominantes, este número variou de 33 a 43%, com excessão de um único caso, de desequilíbrio de ligação igual a 0,7, herdabilidade igual a 0,2 e 100 locos simulados que o número de marcadores significativos foi 48% do valor total de marcadores (Tabela 2).

Quando o critério FDR foi utilizado, novamente o número de marcadores significativos foi inferior à situação de desesquilíbrio de ligação igual a 1, Estes valores variaram de 0,02 a 14 % (Tabela 2), sendo possível observar, em várias repetições em diferentes casos, um número de marcadores não significativos iguais ao número total de marcadores simulados, especialmente para o LD igual a 0,5.

Tabela 2 – Número de marcadores significativos considerando o desequilíbrio de ligação (LD) igual a 0,5, 0,7 e 0,9 para os marcadores tipo SNPs e DArTs em diferentes herdabilidades.

			SNPs		DARTs	
N° de Locos	LD	h²	Teste F	FDR	Teste F	FDR
100		0,2	41,6	0,5	33,6	0,3
100		0,4	36,7	1,0	39,3	0,1
200		0,2	80,9	1,8	73,2	1,8
200	0,5	0,4	73,1	1,8	80,3	2,1
400		0,2	170,8	2,3	173,9	2,8
400		0,4	164,4	0,1	158,8	2,6
100		0,2	45	5,4	48	14,3
100		0,4	41,1	6,3	39,9	5,5
200		0,2	81,9	4,8	81,6	2,8
200	0,7	0,4	76,4	4,1	79,1	3,7
400		0,2	165	2,5	150,9	2,7
400		0,4	166,6	3,2	162,4	1,9
100		0,2	41,5	4,5	58,1	4,8
100		0,4	40,7	7,5	56,1	5,9
200		0,2	88,8	6,1	80,1	7
200	0,9	0,4	82,7	8,8	84,2	6,2
400		0,2	161,2	5,8	156,6	6,2
400		0,4	160,1	6,7	158,1	5,8

Quando os casos de desequilíbrio de ligação diferentes de um foram analisados, o número de marcadores significativos reduziu mais ainda. Pritchard e Przeworski (2001) declararam ainda que a medida de desequilíbrio de ligação é outro fator que afeta o poder de detecção de um QTL via um marcador.

3.1.3 ESTIMATIVAS DE ACURÁCIA E CORRELAÇÃO

Após ser feita a genotipagem na população de estimação e validação, o número de marcadores que tiveram seus efeitos estimados para posterior validação foi escolhido de acordo com os tratamentos testados. Ao todo foram analisadas 11 formas de selecionar os marcadores para a etapa de estimação, das quais algumas não foram apresentadas neste trabalho. No conjunto de dados que apresentava desequilíbrio de ligação completo, isto é, igual a um, as estimativas de correlação entre o valor fenotípico real e o valor fenotípico recuperado a partir das informações genotípicas e fenotípicas da população de estimação variaram entre 0,19 e 0,64, As acurácias da seleção genômica ampla, calculadas após dividir as estimativas de correlação pela raiz quadrada da herdabilidade variaram de 0,44 a 0,99.

Nesta simulação de característica oligogênica, em todos os casos testados, para os diferentes marcadores e valores de constantes K utilizados, as acurácias de seleção se mostraram menores quando foi feita a análise com todos os marcadores genotipados, do que a análise baseada em uma seleção prévia dos marcadores via teste F (Tabela 3). A razão desta inferioridade foi porque muitos dos QTLs explicam uma fração muito pequena da variação genética. Neste caso, mesmo quando são obtidos marcadores ligados a estes QTLs, o erro na estimação dos efeitos é maior que o ganho que estes efeitos poderiam proporcionar à acurácia realizada.

Tabela 3 – Acurácias de seleção utilizando todos os marcadores e utilizando apenas aqueles significativos pelo teste F com nível de significância igual a 40%. D_K significa o marcador dominante do tipo DArT, e a constante K testada. N indica o número total de locos genotipados.

N	H^2	Seleção baseada em todos os Marcadores					Seleção baseada nos marcadores significativos pelo teste F				
		SNPs	D_1	D_1,33	D_1,5	D_2	SNPs	D_1	D_1,33	D_1,5	D_2
100	0,2	0,90	0,75	0,76	0,76	0,76	0,91	0,75	0,75	0,74	0,72
100	0,3	0,83	0,70	0,70	0,70	0,68	0,88	0,71	0,71	0,71	0,69
100	0,4	0,92	0,82	0,82	0,81	0,80	0,95	0,86	0,86	0,85	0,84
200	0,2	0,70	0,56	0,55	0,55	0,52	0,74	0,56	0,55	0,55	0,53
200	0,3	0,75	0,63	0,64	0,64	0,63	0,80	0,65	0,65	0,65	0,63
200	0,4	0,85	0,69	0,69	0,68	0,65	0,91	0,72	0,71	0,69	0,62
400	0,2	0,62	0,52	0,53	0,54	0,54	0,66	0,55	0,55	0,55	0,54
400	0,3	0,64	0,49	0,49	0,49	0,45	0,74	0,60	0,59	0,59	0,55
400	0,4	0,70	0,60	0,60	0,60	0,56	0,77	0,65	0,64	0,63	0,59

É possível observar na Tabela 3 que, de maneira geral, a medida que o número de locos aumenta, a acurácia reduz. Como os primeiros locos simulados, são os que explicam a maior parte da variância genética, a medida que o número de parâmetros a serem estimados aumentam, e os efeitos relacionados a estes locos têm magnitude muito pequena, a acurácia reduz pela mesma razão que as acurácias estimadas quando todos os marcadores foram utilizados foi inferior.

No total, foram analisadas 45 situações onde a seleção foi baseada somente no teste F, e 45 situações em que a seleção foi baseada no critério de FRD. De um total de 45 comparações, a acurácia da seleção dos marcadores após ser considerado o critério de FDR foi superior em 44 vezes, num total de 98%.

O resultado comparando a multiplicação da matriz de incidência Z, pela contante K para tentar explicar melhor a estrutura da população, demonstrou que de um total de 115 comparações, o valor de K igual a um se mostrou melhor em valor absoluto em 76% das vezes e o valor de K igual a 1,33 se mostrou superior em 16% das vezes. No entanto, estas diferenças entre K igual a 1, 1,33, e 1,5

foram praticamente nulas. A diferença entre a acurácia obtida com o uso de marcadores codominantes, tipo SNPs, e o uso de marcadores DArTs, variou entre zero e 0,29, com média igual a 0,13.

O valor de constante K que gerou os melhores resultados analisados foi quando a matriz Z foi multiplicada por 1. Este resultado foi diferente do esperado, uma vez que a o sorteio dos genótipos no procedimento de simulação, foi feito de forma aleatória, com $P(A_1A_1) = P(A_1A_2) = P(A_2A_2) = 1/3$. Dessa forma, a freqüência de genótipos confundidos na classe de código 1 é $P(A_1A_2) = P(A_2A_2) = ½$ e o valor de K esperado seria então ½ x 1 + ½ x 2 = 1,5. Embora as diferenças entre os valores de acurácias para cada constante K tenha sido muito pequena, a alternativa de multiplicar a matriz Z por um peso, mesmo que este explique a estrutura da população, não parece ser aconselhada nestas populações simuladas. Entretanto, os resultados para K igual a 1, 1,33 e 1,5 foram praticamente idênticos.

3.1.4 OUTRAS MEDIDAS DE DESEQUILÍBRIO DE LIGAÇÃO

A análise dos dados simulados com desequilíbrio de ligação diferente de 1, foi feitas apenas para os métodos utilizando o Teste F e a proteção de FDR, uma vez que estes métodos se mostraram superiores. A constante utilizada para K foi apenas o valor 1 e foram consideradas as herdabilidades de cada extremo, 0,2 e 0,4.

Em algumas situações quando o desequilíbrio de ligação foi igual a 0,2 e a taxa de falsa descoberta foi considerada, a correlação e acurácia não puderam ser estimadas. A razão disto foi o fato de que, em algumas repetições do *Jacknife*, todos os marcadores foram não significativos, o que impossibilitou a recuperação dos valores fenotípicos para este grupo, uma vez que não havia marcadores pra terem seus efeitos estimados. Como a correlação e a acurácia foi estimada, baseada em todos os indivíduos, a ausência de fenótipo para alguns grupos de indivíduos não permitiu a estimativa (Tabela 5).

Assim como na situação de desequilíbrio de ligação igual a 1, as estimativas de correlação e de acurácias foram superiores quando o critério de

FDR foi adotado ao comparar com a situação em que apenas a estatística F foi utilizada (Tabela 5). As estimativas de correlação variaram de -0,09 a 0,48 e as estimativas de acurácia variaram de -0,13 a 0,77.

Tabela 5 – Estimativas de acurácia da seleção genômica em diferentes desequilíbrios de ligação (LD) e considerando dois critérios de significância: Teste F e FDR.

			SNPs		DArTs	
N° de Locos	LD	h^2	Teste F	FDR	Teste F	FDR
100		0,2	-0,09	-0,10	-0,07	-
100		0,4	-0,14	0,19	-0,08	-
200		0,2	0,02	0,17	0,06	0,10
200	0,5	0,4	0,05	0,13	0,11	0,06
400		0,2	0,07	-	0,19	-
400		0,4	0,07	-	0,05	-0,04
100		0,2	0,19	0,35	0,33	0,31
100		0,4	0,48	0,50	0,34	0,40
200		0,2	0,23	0,39	0,07	0,21
200	0,7	0,4	0,29	0,47	0,2	0,42
400		0,2	0,20	0,55	0,06	0,40
400		0,4	0,14	0,38	0,08	0,35
100		0,2	0,51	0,64	0,48	0,62
100		0,4	0,65	0,71	0,57	0,61
200		0,2	0,45	0,65	0,28	0,53
200	0,9	0,4	0,63	0,72	0,5	0,60
400		0,2	0,46	0,77	0,35	0,56
400		0,4	0,59	0,76	0,46	0,63

3.1.5 GANHOS DE SELEÇÃO.

Os ganhos de seleção foram calculados em relação a acurácia de 0,68 que é a acurácia máxima obtida via seleção fenotípica (Resende, 2002; Resende,

2007a). Em espécies de ciclo longo, como o Eucalipto, uma das vantagens da seleção genômica pode ser a redução do ciclo de seleção ao fazer a seleção precoce. Assim, o ganho por unidade de tempo foi avaliado considerando a redução de 0, 25, 50, 75 e 88,5% no tempo do ciclo de melhoramento o que potencializou os ganhos (Figura 4).

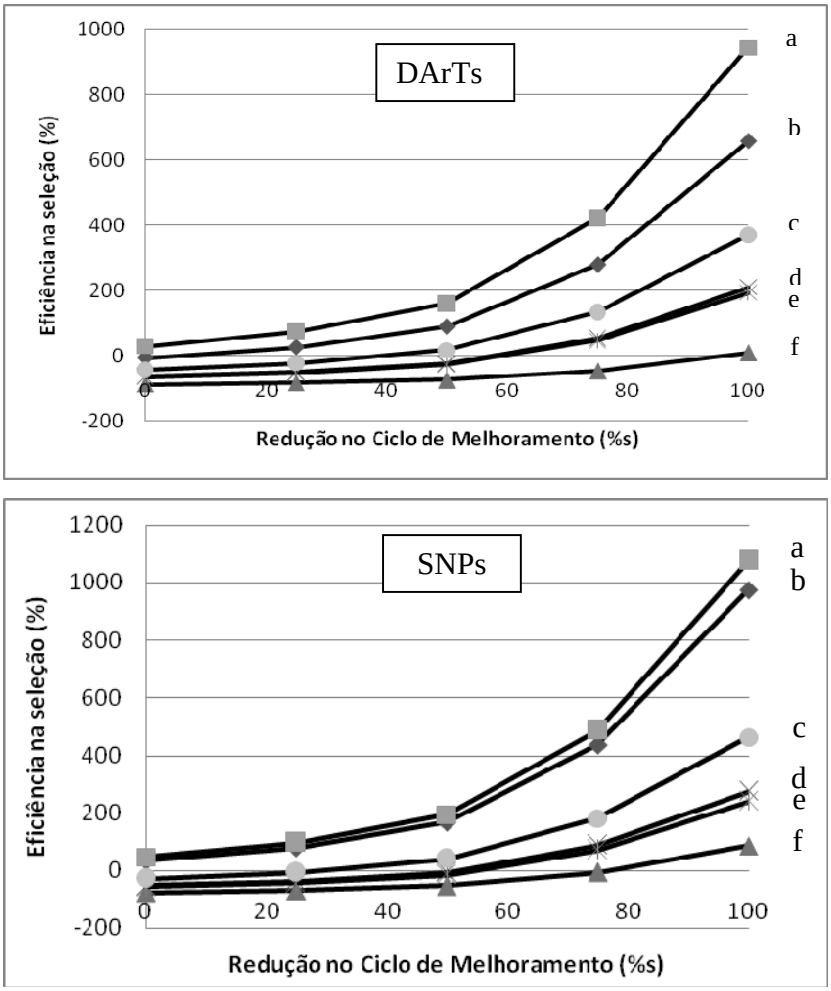

Figura 4 – Gráfico com eficiência da GWS em relação a seleção fenotípica. (a) – Valor máximo de acurácia obtido ao analisar casos com os 3 números de locos e 3 herdabilidades e ainda LD igual a 1 (b) – Valor mínimo de acurácia quando LD foi igual a 1; (c) e (d) – Valor máximo e mínimo de acurácia quando o LD foi igual a 0,9; (e) e (f) – Valor máximo e mínimo para LD igual a 0,7.

Percebe se aqui, o impacto que a seleção genômica pode ter principalmente em espécies de ciclo longo como as perenes com ganhos que chegam a 200% caso o ciclo seja reduzido para a metade do tempo usual. Nesta situação, após o mapeamento de poucos QTLs de maior efeito, metodologias que

abordam a descoberta de genes candidatos poderiam ser aplicadas para identificação de marcadores diretamente ligados a estes genes e aplicação da MAS. No entanto, o grupo de QTLs de pequenos efeitos possivelmente nao seria detectado, o que reduziria os ganhos.

3.2 CARACTERÍSTICA SIMULADA COM EFEITOS INFINITESIMAIS (POLIGÊNICA)

A simulação realizada de acordo com Lande e Thompson (1990) obteve um número de marcadores significativos pelo teste F que variou de 33 a 75% no caso dos marcadores dominantes tipo DArT, e de 34 a 85% nos casos dos marcadores codominantes tipo SNP. Após a aplicação da proteção de razão de falsa descoberta, o número de marcadores signicativos diminuiu, e variou de 0 a 68%, para os DArTs e 0 a 81% para os SNPs (Tabela 7).

Quando foi avaliada a variação em porcentagem separada por número de locos marcadores, estes valores ficaram entre 35 e 85% para 100 locos, 33 e 70% para 200 locos e 38 e 57% para 400 locos. Padrão semelhante foi observado ao avaliar o teste FDR, que obteve variações de 0 a 80%, de 0 a 51% e de 0 a 20% para 100, 200 e 400 locos respectivamente.

Na tabela 7 pode-se observar que o número de marcas significativas, tanto pelo teste F quanto pelo critério FDR, aumentou ao comparar com a característica oligogênica, uma vez que os locos que explicavam pequenas proporções da variação fenotípica total eram poucos.

Tabela 7 – Número de marcadores significativos avaliados em diferentes situações de herdabilidade, desequilíbrio de ligação igual a 0,5, 0,7, 0,9 e 1, e 100, 200 e 400 locos controlando o caracter quantitativo.

Significativos pelo teste FDR									
N	h²	DARTs				SNPs			
		0,5	0,7	0,9	1	0,5	0,7	0,9	1
100	0,2	2,1	20,9	14,8	46,1	3	19	22,9	46,6
100	0,3	0	7,1	45,9	50,9	0,1	6,8	50,2	67,3
100	0,4	2,5	14	54,8	68,4	1,3	25,4	66,6	80,5
200	0,2	5,8	9,2	14	30,2	5,9	11,9	25,7	45,2
200	0,3	0	4,6	15,9	49,4	0	12,1	43	77
200	0,4	3	10,3	59,6	78,4	1,3	16,4	70,1	102,8
400	0,2	6,7	0,4	4,5	12,6	6,6	0,1	13,8	38,5
400	0,3	0	1,4	46,2	30,4	0	0,1	58	62,4
400	0,4	0,5	6,6	53,1	75,9	0,1	12	69,7	81,8
Significativos Via o Teste F (α = 0,4)									
N	h²	DARTs				SNPs			
		0,5	0,7	0,9	1	0,5	0,7	0,9	1
100	0,2	44,7	57	51,2	64,3	44,3	55,7	56,4	68
100	0,3	37,6	40,1	64	70,2	35,8	45,8	69,4	75,6
100	0,4	45,7	54	69,3	75,1	41,8	57,6	76	84,6
200	0,2	80,8	92,6	103,4	102,1	82,5	92	111,3	112,3
200	0,3	66,1	91,7	104,6	109,5	69,7	98	116,4	127,5
200	0,4	90,2	88,9	114,8	129,7	94,7	93,4	117,3	140,8
400	0,2	176,5	156,8	175,6	189,4	168,8	153,6	182,3	202,9
400	0,3	173,1	177,3	203,7	204,5	164,7	183,2	208,2	217,8
400	0,4	154,2	180,4	199,4	211,4	153	179,2	220,9	231,3

3.2.1 ESTIMATIVAS DE ACURÁCIAS E CORRELAÇÃO

As estimativas de correlação e acurácia de seleção na simulação de dados que se assemelhou, do ponto de vista biológico, a uma característica quantitativa, variaram de -0,14 a 0,59 para correlação e 0 a 0,93 para a acurácia.

Nesta simulação, em que nenhum QTL de efeito maior e diferenciado dos outros foi simulado, as estiimativas de correlação e acurácia calculadas após aplicar critério do FDR foram inferiores às estimativas calculadas após a seleção via teste F. Este último por sua vez, obteve acurácias inferiores às estimativas obtidas quando todos os marcadores foram utilizados. Esta situação pode ser vizualizada na Tabela 8, que ilustra a acurácia obtida via marcadores SNPs nas situações de desequilíbrio de ligação igual a 0, 9 e 1,0. Os demais desequilíbrios de ligação obtiveram resultados de acurácia baixo, variando de 0 a 0,24 para LD igual a 0,5 e de 0 a 0,44 para LD igual a 0,7 (Resultados não apresentados).

Tabela 8 – Estimativas de Acurácia obtidas via seleção baseada em todos os marcadores, seleção dos marcadores significativos via teste F e via FDR. Desequilíbrios de ligação igual a 0,9 e 1, Marcador SNP

		Todas as Marcas		Teste F		FDR	
N	h²	0,9	1	0,9	1	0,9	1
100	0,2	0,51	0,82	0,48	0,81	0,32	0,76
100	0,3	0,63	0,89	0,60	0,85	0,55	0,83
100	0,4	0,63	0,93	0,55	0,91	0,52	0,89
200	0,2	0,56	0,65	0,41	0,47	0,27	0,42
200	0,3	0,47	0,75	0,39	0,67	0,21	0,54
200	0,4	0,60	0,80	0,58	0,64	0,50	0,53
400	0,2	0,50	0,68	0,48	0,57	0,21	0,31
400	0,3	0,52	0,70	0,56	0,62	0,34	0,29
400	0,4	0,56	0,69	0,51	0,60	0,35	0,45

Nesta situação, pôde se observar também, que as acurácias foram muito afetadas pela h² individual, uma vez que muitos locos contribuíram na variação fenotípica e a herdabilidade foi então gerada a partir da herdabilidade individual de muitos marcadores, o que alterou as estimativas.

Os valores altos de acurácia obtidos indicam a possibilidade de aplicação da seleção genômica com ambos os tipos de marcador. Os resultados obtidos concordam com os valores obtidos via abordagem determinística feita por Resende (2008). O melhor método de seleção foi quando a seleção genômica foi realizada considerando todos os marcadores. Como o objetivo da seleção genômica é maximizar o ganho de seleção e consequentemente a acurácia, nesta situação quantitativa, em que muitos locos controlam o carácter com magnitudes de efeitos semelhantes, o fato de utilizar marcas que se encontram em LD, mesmo que a associação não tenha sido detectada, maximiza estes ganhos. Quando a simulação foi feita com LD igual a um, existiu um marcador que estava em desequilíbrio de ligação com cada QTL. No entanto, é importante ressaltar que, em situações reais, para que seja genotipado um marcador dentro do bloco genômico que se encontra em LD com cada QTL é necessária a saturação do genoma com um grande número de marcas para que se tenha a densidade suficiente e obter estes marcadores em LD. Uma vez que não se tem conhecimento do posicionamento dos QTLs no genoma, esta genotipagem tem que ser ampla, o que acarreta também na genotipagem de marcas que não estão em associação nenhuma com nenhum QTL. Nesta situação pode ser necessário descartar alguns marcadores, os quais não foi observada a associação para reduzir o número de parâmetros que serão estimados.

Outra observação interessante foi a respeito da comparação entre marcadores codominantes e dominantes. Obviamente que quando analisados em situações idênticas, os marcadores SNPs proporcionam maior acurácia, por serem mais informativos. No entanto, considerando a disponibilidade de um número de SNPs com densidade suficiente para obter 90% dos QTLs em associação com pelo menos um marcador (LD = 0,9), e um número maior de DArTs neste mesmo genoma, com densidade de marcadores suficientes para detectar associação com todos os QTLs, as acurácias obtidas via marcadores DArTs foram superiores.

A metodologia BLUP se adequou bem as análises de seleção genômica e demonstrou potencial em sua aplicação. Outras metodologias foram relatadas como superiores (Meuwissen, 2001, Meuwissen, 2009) fato também observado e relatado no próximo capítulo com a metodologia BayesA na análise de dados reais

em Eucalipto. O modelo pode ainda conter o efeito poligênico, quando alguns locos não estiverem sendo explicados por marcadores (Hayes, 2009)

Os ganhos de seleção foram calculados para os desequilíbrios de ligação de 0,9 e 1. Embora em alguns casos a acurácia obtida via Seleção Genômica foi menor que o máximo possível para a seleção fenotípica, quando o ganho é calculado em função do tempo de seleção, é possível observar uma superioridade. Para o LD igual a 0,9, o ganho com a redução do ciclo em 50% variou de -32 a 62% no caso de DArTs e de 46 a 84% no caso de SNPs. Quando o LD foi igual a 1,0, a variação foi de 90 a 134% para SNPs e de 27 a 131% para DArTs (Figura 5). Pode se observar nesse caso que a GWS pode impactar o melhoramento genético com grandes possibilidades de ganho, sendo no entanto necessário o uso de um grande número de marcadores.

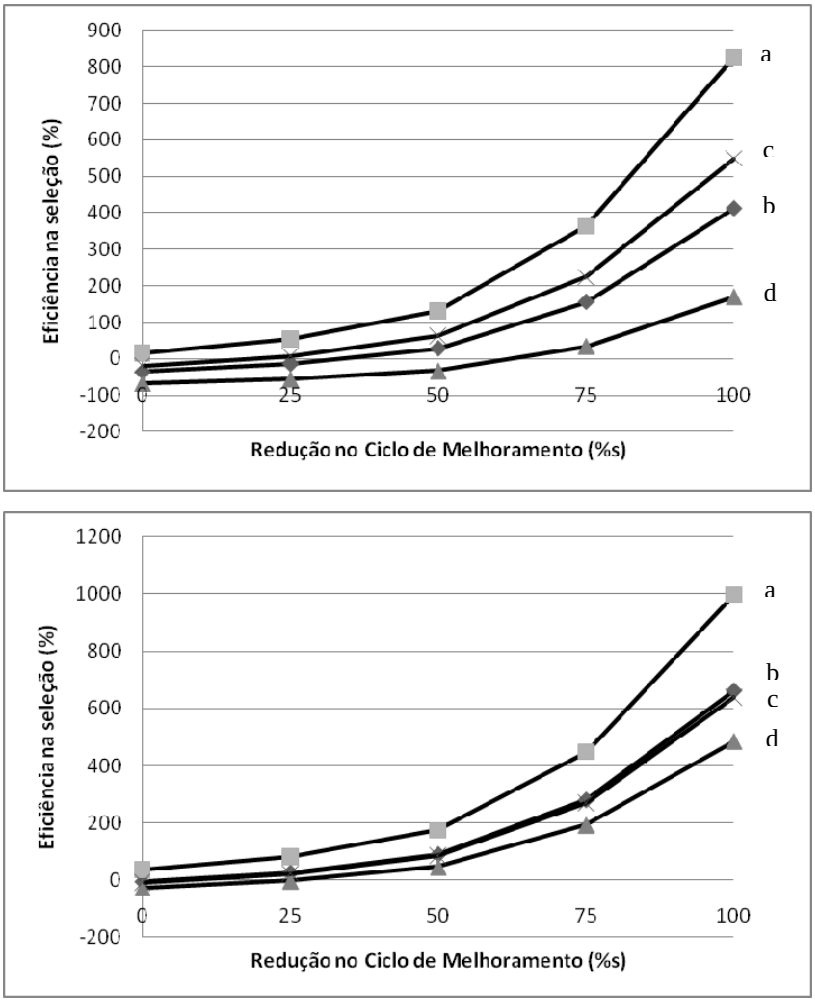

Figura 5 – Eficiência da GWS em relação a seleção fenotípica quando foi considerado: (a) e (b) Valores máximos e mínimos de acurácia para os cenários testados com LD igual a 1; (c) e (d) Valores máximos e mínimos para os cenários com LD igual a 0,9

4. CONCLUSÃO

A partir deste estudo, pôde se concluir que a seleção genômica ampla tem perspectivas de serem aplicadas em quaisquer organismos. A real concretização desta tecnologia pode alterar completamente a maneira como o melhoramento genético é feito. No entanto, ainda são necessários a disponibilidade de um grande número de marcadores cobrindo de maneira ampla, todo o genoma. Nos próximos anos, algumas espécies terão seus genomas seqüenciados, situação esta que permite a descoberta de um número quase infinito de marcadores SNPs. Entretanto, espécies que não têm esta disponibilidade, podem perfeitamente, desenvolver marcadores dominantes do tipo DArTs para utilização na GWS.

Novos estudos precisam ser realizados, avaliando o custo de aplicação, considerando taxas reais de desequilíbrio de ligação em culturas vegetais e testando novas metodologias. Este estudo demonstrou, no entanto que a metodologia BLUP atende bem às análises e pode ser utilizada como forma de estimação dos efeitos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Bernardo R (2008). Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Science, 48:1649-1664.

Bernardo R, Yu JM (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082-1090.

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

de los Campos, G, Naya, H, Gianola, D, Crossa, J, Legarra, A, Manfredi, E, Weigel, K, Cotes, JM (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree Genetics 182: 375-385 .

Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. Journal of Animal Science 82:313-328.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. Journal of Animal Science 85:2104-2114.

El-Din El-Assal S, Alonso-Blanco C, Peeters AJ, Raz V, Koornneef M (2001): A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2, Nat Genet, 29:435-440.

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD (2000): fw2,2: a quantitative trait locus key to the evolution of tomato fruit size. Science, 289:85-88.

Gianola D, Van Kaam JBCHM (2008) Reproducing Kernal Hilbert Spaces Regression methods for genomic assisted prediction of quantitative trais. Genetics 178:4:2289-2303.

Gredler,B, Nirea,KG, Solberg TR, Egger-Danner, C, Meuwissen, T, Sölkner, J (2009). A comparison of methods for genomic selection in Austrian dual purpose Simmental cattle. Proc. Assoc. Advmt. Anim. Breed. Genet. 18:568-571.

Hastbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A.(1994). The diastrophic dysplasia

gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell*; 78: 1073–87.

Hayes B (2009): Whole Genome Association and Genomic Selection – Course Notes. 120p.

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic Selection for Crop Improvement. Crop Sci 49:1-12.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25.

Jenkins, S., and N. Gibson. 2002, High-throughput SNP genotyping. Comp. Funct. Genom. 3:57–66.

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124.

Liu J, Van Eck J, Cong B, Tanksley SD (2002): A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci USA , 99:13302-13306.

Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S. (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics, 124:377-389.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, et al. 2009 Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4(4): e5350.

Meuwissen T, Hayes B, Goddard M (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829.

Meuwissen T, Solberg TR, Shepherd R, Wooliams JA (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution. 41:2.

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

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Resende MDV (2008) Notas de aula de Melhoramento Florestal.

Resende MDV, Lopes PS, Silva RL, Pires IE (2008) Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. Pesquisa Florestal Brasileira 56:63-77.

Resende, MDV (2002). Genética biométrica e estatística no melhoramento de plantas perenes. Brasília: Embrapa Informação Tecnológica, 2002, 975p.

Resende, MDV (2007). Matemática e estatística na análise de experimentos e no melhoramento genético. Colombo: Embrapa Florestas. 561p.

Resende MDV (2008) Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais. Colombo:EMBRAPA Florestas. 330p.

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123:218-223.

Schuster I, Cruz CD (2004) Estatística Genômica aplicada a populações derivadas de cruzamentos controlados. Editora UFV, Viçosa.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Reducing dimensionality for prediction of genome-wide breeding values. *Genet Sel Evol.* 2009;**41**:29.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using different marker types and densities. *J Anim Sci.* 2008;**86**:2447–2454.

Takahashi Y, Shomura A, Sasaki T, Yano M (2001): Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the a subunit of protein kinase CK2, Proc Natl Acad Sci USA, 98:7922-7927.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915-9920.

Wong, CK., Bernardo, R (2008). Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116:815–824.

Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto K, Umehara Y, Nagamura Y, Sasaki T (2000): Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell, 12:2473-2484.

Zhong, S, Dekkers, JCM., Fernando, RL., Jannink, JL (2009). Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study Genetics. 182: 355-364.

CAPÍTULO II

SELEÇÂO GENÔMICA AMPLA EM *Eucalyptus*

VIÇOSA

MINAS GERAIS – BRASIL

2010

1. INTRODUÇÃO

Atualmente as espécies do gênero *Eucalyptus* são as mais utilizadas em plantações florestais no país, sendo que o Brasil ocupa lugar de destaque como um dos países de maior área plantada com eucalipto do mundo. Estas plantações de rápido crescimento são o suporte para uma indústria multimilionária baseada na produção de papel, celulose, carvão vegetal e produtos sólidos (Grattapaglia *et al.*, 2009).

Os programas de melhoramento de eucalipto são em geral baseados na obtenção de híbridos interespecíficos férteis que capitalizam os efeitos da heterose para os caracteres de crescimento e a propagação clonal que explora, além dos efeitos aditivos, os efeitos devido a dominância (Resende e Assis, 2008). A estratégia de seleção utilizada é a seleção recorrente recíproca interespecífica, e o processo de recomendação que inicia na hibridação, passa por testes de progênie e testes clonais para então indicação de um material comercial. Esse processo demanda em média, de 12 a 16 anos.

Assim, em espécies florestais, as perspectivas do uso de marcadores moleculares para auxilio no melhoramento genético (Seleção assitida por marcadores – MAS) são ainda maiores. Nestes casos, a seleção precoce permitiria a seleção nos primeiros estágios das mudas, não sendo necessário aguardar até que as árvores expressassem o fenótipo desejado o que poderia então, representar um substituto para a seleção fenotípica (Dekkers, 2004). No caso de *Eucalyptus*, os testes clonais amplos derivados do intercruzamento de algumas dezenas de parentais elites de diferentes espécies se apresentam como uma condição favorável a MAS (Grattapaglia 2007; Grattapaglia and Kirst 2008), pois geram grande quantidade de LD. Com o uso da seleção precoce assitida por marcadores é possível aumentar o ganho genético por unidade de tempo. Esta oportunidade é rara em espécies anuais, mas é evidente em espécies perenes com longo ciclo de vida. No caso de *Eucalyptus,* este ganho é ainda maior pelo fato da seleção precoce permitir a instalação imediata de testes clonais a partir de

mudas jovens antes mesmo de estas serem submetidas a teste de progênie convencional.

Vários trabalhos descreveram o sucesso na identificação de QTLs para componentes de produtividade (crescimento volumétrico, forma), qualidade da madeira (densidade básica, teor de lignina, rendimento em celulose), resistência a estresses abióticos (tolerância ao frio, à seca) e resistência a patógenos, principalmente fungos (O' Malley et al. 1996; Sewell e Neale 2000; Neale e Savolainen 2004; Grattapaglia e Kirst 2008). No entanto, apesar de dezenas ou mesmo centenas de QTLs terem sido mapeados, a informação gerada não tem sido imediatamente útil para a seleção assistida no melhoramento. A razão deste insucesso pode ser extrapolada de estudos de associação entre vários marcadores e diferentes características (Eckert *et al*. 2009; Gonzalez Martinez *et al*. 2007; Gonzalez Martinez *et al*. 2008). A magnitude dos efeitos da associação individual destes marcadores, raramente excedeu 5 a 10% da variância genética. Estes resultados são similares aos obtidos em estudos de associação em humanos (Visscher e Montgomery, 2009), animais domésticos (Goddard e Hayes 2009) e plantas (Buckler *et al*. 2009) e demonstram a natureza complexa das características quantitativas e questionam o uso de algumas poucas associações discretas na melhoria de caracteres quantitativos em espécies florestais. Para que a seleção assistida por marcadores possa ser eficiente, é necessária a captura simultânea de uma grande proporção da variação genotípica de uma característica.

Para efetivamente alcançar esse benefício um novo método de seleção denominado seleção genômica (GS) ou seleção genômica ampla (GWS) foi proposto (Meuwissen *et al*. 2001). A GWS pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético de espécies alógamas, apresenta alta acurácia seletiva para a seleção baseada exclusivamente em marcadores (após terem seus efeitos genéticos estimados a partir de dados fenotípicos em uma amostra da população de seleção) e não exige prévio conhecimento das posições (mapa) dos QTLs (Resende et al. 2008).

A maior limitação na implementação da seleção genômica é o grande número de marcadores necessários e os altos custos de genotipagem (Goddard

and Hayes 2007). No entanto, uma abordagem que permite a genotipagem paralela de eucalipto a baixo custo e dado os altos níveis de polimorfismo no genoma do eucalipto (Grattapaglia and Kirst 2008) é o marcador DArT (Diversity Array Technology) (Jaccoud *et al.* 2001; Wenzl *et al*. 2006). Uma vez disponível um elevado número de marcadores, a probabilidade de se encontrar um QTL em LD com pelo menos um marcador é muito alta. Este aspecto é extremamente importante uma vez que somente os marcadores em LD com os QTLs serão úteis na determinação dos fenótipos e na explicação da variação genética.

É importante salientar que o desequilíbrio de ligação entre dois locos é afetado pela freqüência alélica, pelas taxas de recombinações e pelo tamanho efetivo populacional (Flint-Garcia *et. al*, 2003). As taxas de recombinação entre um QTL e um marcador podem ser controladas pela densidade de marcadores, uma vez que com um grande número de marcadores, espera-se encontrar um marcador mais próximo do QTL e, consequentemete, com menor taxa de recombinação. No entanto, o tamanho efetivo populacional é uma característica da população de melhoramento que têm algumas limitações quanto a redução muito acentuada do tamanho efetivo com eventual perda de variabilidade genética.

Assim, este trabalho relata a implementação da seleção genômica em duas populações elites de *Eucalyptus* com diferentes tamanhos efetivos utilizando um grande número de marcadores DArTs. Trabalhos semelhantes já foram realizados no melhoramento animal (Van Raden *et al*., 2009; Hayes *et al*. 2009), no entanto este é o primeiro estudo de seleção genômica com dados reais de plantas.

2. MATERIAL E MÉTODOS

Neste primeiro experimento de prova de conceito da GWS em *Eucalyptus*, duas populações foram utilizadas na estimação de efeitos genéticos e validação dos marcadores DArTs. Uma população, oriunda da CENIBRA S.A., foi composta por 783 indivíduos amostrados aleatoriamente de 51 famílias derivadas de 11 parentais (N_e igual a 11). Todos os indivíduos foram fenotipados para altura total e diâmetro a altura do peito (DAP) aos três anos e genotipados com 2343 marcadores DArTs de alta qualidade. A segunda população foi oriunda da empresa FIBRIA S.A., e foi composta por 920 indivíduos amostrados em uma população com um tamanho efetivo de 120. O fenótipo de todos os indivíduos foi coletado para as características altura total, DAP, e densidade básica avaliada via Pilodyn. A genotipagem envolveu 3564 marcadores DArTs de alta qualidade.

Os marcadores DArTs foram gerados na Austrália como parte da dissertação de doutorado dos estudantes da Universidade de Brasília, César Petroli e Carolina Sansaloni, sob orientação do pesquisador da EMBRAPA – CENARGEN, Dr. Dario Grattapaglia.

Os valores fenotípicos foram derregressados (*deregressed*) corrigindo os dados da CENIBRA para efeitos de blocos e efeitos de parcela e os dados da FIBRIA para efeitos de blocos, parcela e falhas no experimento. Análises de associação individual baseada em regressão em marca simples foram conduzidas considerando os efeitos de marcadores como fixos.

2.1 METODOLOGIA DE ANÁLISE

Foram obtidas as acurácias da seleção genômica baseada em dois métodos de estimação dos efeitos de marcadores propostos por Meuwissen (2001): BLUP/GWS e BayesA. Em ambos os casos, a validação foi realizada através da reamostragem de um grupo de indivíduos via procedimento Jacknife. A metodoloia generalizada do Jacknife baseia-se na divisão do conjunto de N dados amostrais em g grupos de tamanho igual a k, de forma que N = gk. A população da CENIBRA foi dividida em 9 grupos de 87 indivíduos de maneira que a mesma

análise foi realizada nove vezes. Em cada repetição, um grupo foi removido da população e utilizado para a formação da população de validação e 696 indivíduos (8 grupos x 87 indivíduos) foram utilizados na população de estimação dos efeitos dos marcadores. O mesmo procedimento foi adotado na população da FIBRIA, neste caso, foram separados 10 grupos com 92 indivíduos em cada e 9 grupos foram utilizados na estimação dos efeitos dos marcadores.

No modelo *Jacknife*, a validação é feita em uma população independente, pertencente à mesma população e às mesmas famílias. No caso da população da FIBRIA, outra forma de validação dos efeitos estimados dos marcadores foi feita. A validação foi realizada em uma população independente pertencente à mesma população, porém pertencente à diferentes famílias. Assim foram removidos da população de 920 indivíduos, todos aqueles que pertenciam a famílias oriundas de quatro genitores aleatoriamente selecionados. Dessa forma, um grupo de 102 indivíduos formou a população de validação pertencente a diferentes famílias da população utilizada na estimação. Esta validação é considerada mais precisa pois retrata o padrão de desequilíbrio de ligação que é permanente na população, sem ser afetado pela genealogia.

2.2 ESTIMAÇÃO DOS EFEITOS DE MARCADORES VIA BLUP/GWS

Os marcadores que tiveram seus efeitos estimados a partir dos dados fenotípicos foram selecionados após uma análise de associação individual via regressão em marca simples e a associação declarada pela estatística F a um nível de significância de 5%, conforme descrito no capítulo anterior. O valor da constante k utilizada para compor a matriz Z foi igual a 1,0. Análises anteriormente realizadas demonstraram que as acurácias foram inferiores quando todos os marcadores foram utilizados e quando diferentes valores da constante k foram selecionados. De um total de 2343 marcadores utilizados na população da CENIBRA, 555 marcadores tiveram sua associação declarada como significativa para altura total e 944 para DAP. Quando o teste FDR foi aplicado a 5%, o número de marcadores significativos caiu para 210 e 757 para altura e DAP, respectivamente. Já na população da FIBRIA, o número de marcadores

significativos pelo teste F foi de 1081,1414 e 816, respectivamente, para altura, DAP e Pilodyn e após as análises de FDR, esse número caiu para 624, 1041 e 308 marcadores. Estes resultados sugerem que a característica DAP é possivelmente controlada por um número maior de locos nestas populações. Em todas as situações, os resultados de acurácia da seleção genômica com os marcadores selecionados após o teste FDR foram inferiores à situação onde apenas o teste F foi aplicado e os resultados então não são aqui apresentados. Estes resultados concordam com os obtidos no capítulo anterior ao calcular as acurácias da GWS na simulação de características quantitativas. A possível explicação para este fato é que o nível de significância global utilizado é rigoroso a ponto de não detectar algumas associações que auxiliam no incremento da acurácia.

Os efeitos dos marcadores significativos foram estimados pelo procedimento BLUP proposto por Meuwissen *et. al.* (2001).

O modelo linear misto usado conforme Resende(2008) foi :

y = Xb + Zm + e,

onde y é um vetor de observações fenotípicas, b é um vetor de efeitos fixos, m é o vetor de efeitos dos marcadores assumidos como aleatórios e e se refere ao vetor de erros aleatórios. X e Z são as matrizes de incidência para b e m.

A equação de modelos mistos para a predição dos valores genômicos é dada por:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$ onde σ_g^2 se refere a variância genetica da característica e σ_e^2 a variância residual. O valor genômico do indivíduo é dado por:

$$GBV = \hat{y}_j = \sum_i Z_i \hat{m}_i$$,

O valor de σ_g^2 foi estimado via metodologia REML através do software SELEGEN (Resende 2007b) e também a partir de diferentes valores de herdabilidade relatados na literatura como padrão para cada característica em estudo

2.3 ESTIMAÇÃO DOS EFEITOS DE MARCADORES VIA BayesA.

O método BayesA equivale ao método BLUP com variâncias heterogêneas, pois as variâncias dos segmentos cromossômicos diferem para cada segmento e são estimadas sob esse modelo, considerando a informação combinada dos dados e da distribuição *a priori* para estas variâncias. Essa distribuição é tomada como uma qui-quadrado invertida e escalonada. Para obtenção dessa informação combinada ou da distribuição das variâncias, adota-se o procedimento de amostragem de Gibbs.

O seguinte modelo foi utilizado

y = Xb + Zg + e, onde:

y : vetor de dados fenotípicos.

b : vetor de efeitos fixos.

g : vetor de valores aleatórios dos marcadores

e : vetor de erros.

X, Z : matrizes de incidência que associam b e h aos dados fenotípicos (y).

De maneira resumida, os passos para estimação de cada efeito podem ser apresentados da seguinte forma, conforme Resende (2008).

1, Foram fornecidos os valores iniciais dos parâmetros de locação e dispersão do modelo. A média geral $\bar{y}$ como o único efeito fixo, foi inicialmente calculada como a média aritmética das observações. O vetor dos efeitos de marcadores foi inicializado com um número positivo e de pequena magnitude igual a 0,01.

2, Os valores da variância σ^2_{gi} para o i-ésimo segmento cromossômico, foi amostrado da distribuição condicional completa $P(\sigma^2_{gi}|g_i) = \chi^{-2}(\upsilon + q, S^2 + g_i{}'g_i)$ com $\upsilon = 4.012$ e $S^2 = 0.002$ conforme Meuwissen *et al* (2001).

3, Dados gi e $\bar{y}$, foi calculado os valores de e via $e = (y - Xb - Zg)$, em que Z é a matriz de incidência para os efeitos de marcadores. Então, atualize a variância residual por meio da amostragem de $\chi^{-2}(n-2, e_i{}'e_i)$.

4, A média geral foi amostrada de uma distribuição normal com média (1/n) $(y - Zg)$ e variância σ^2_e / n após a variância residual ter sido atualizada

5, Por fim, os efeitos dos marcadores g_{ij} foram amostrados de uma distribuição normal com média $\frac{Z_{ij}^{'} y - Z_{ij}^{'} Z g_{ij=0} - Z_{ij}^{'} 1_n u}{Z_{ij}^{'} Z_{ij} + \sigma_e^2 / \sigma_{gi}^2}$ e variância $\sigma_e^2 / (Z_{ij}^{'} Z_{ij} + \sigma_e^2 / \sigma_{gi}^2)$, dado a amostragem mais recente da média, σ_e^2 e σ_{gi}^2, em que Zij é o vetor coluna de Z com efeitos g_{ij} e, g_{ij}=0 equivale a g com efeito gij igualado a zero.

6, Os passos de (2) a (5) foram repetidos 10 mil vezes. Após todas as iterações, foram descartados os 2000 primeiros ciclos e selecionados um valor de cada parâmetro, $\overline{y}$, σ_{gi}^2, σ_e^2, e e gi, a cada 100 ciclos de modo que ao final da análise obteve-se 80 repetições. Uma análise visual do gráfico identificou convergência nas estimações. A partir daí, o valor dos efeitos estimados para cada marcador foram calculados pela média aritmética das 80 iterações.

Este algotirmo foi implementado no software R, no entanto, a análise de cada repetição do procedimento Jacknife demorou aproximadamente 12 dias em um servidor quad core, 2,33Ghz com 12 Gb de memória RAM. Dessa forma, esta metodologia foi utilizada apenas para a característica altura total na população da CENIBRA.

2.4 AVALIAÇÃO DOS DADOS

‘ A avaliação dos dados foi realizada baseada na correlação do valor genético predito com o fenótipo observado. Foi isolado também a acurácia de seleção, removendo a influência da herdabilidade na capacidade de predição.

A seleção genômica foi ainda comparada com a seleção fenotípica quanto ao ganho de seleção por unidade de tempo. Os únicos parâmetros variáveis no cálculo do ganho de seleção ao comparar a seleção genômica com a fenotípica são a acurácia de seleção e possívelmente o tempo de seleção. Assim, as acurácias obtidas pela GWS foram comparadas com o valor máximo de acurácia possível de se obter via seleção fenotípica e a relação foi avaliada considerando a expectatica de redução do tempo de geração em ½ e ¾ ao utilizar a seleção precoce via dados genotípicos.

3. RESULTADOS E DISCUSSÃO

3.1 POPULAÇÃO CENIBRA

3.1.1 ACURÁCIAS CALCULADAS NA POPULAÇÃO DE ESTIMAÇÃO

As acurácias obtidas na predição dos valores genômicos aditivos na população de estimação foram obtidas via a variância do erro de predição (PEV) e inversão da matriz dos coeficientes da equação de modelos mistos e são demonstradas na Tabela 1. Estes valores representam o valor máximo que pode se obter de acurácia nesta população da CENIBRA.

Tabela 1, Acurácias realizadas na população de estimação obtidas via PEV e inversão da matriz dos coeficientes da equação de modelos mistos sob diferentes herdabilidades

H^2	Altura total	DAP
	2343 Marcadores	2343 Marcadores
0,20	0,70	0,66
0,30	0,74	0,67
0,40	0,77	0,70

Os resultados na Tabela 1 indicam pequena superioridade nas acurácias obtidas para Altura, em função do menor número de marcadores com efeitos significativos o que proporciona menor número de parâmetro a serem estimados. As análises foram feitas considerando as herdabilidades individuais de árvores e variou no intervalo das estimativas comumente relatadas na literatura para estas características em *Eucalyptus* em condições experimentais similares.

Para comparar as acurácias realizadas foram gerados os valores de acurácias esperadas utilizando a abordagem determinística proposta por Resende *et. al* (2008) para estimar a acurácia esperada da seleção genômica. Os parâmetros utilizados foram uma população com tamanho efetivo Ne igual a 10 e a distância entre marcadores igual a 0,005M (2 marcadores a cada cM) (Tabela 2). Esta distância entre marcadores foi escolhida, pois equivale a aproximadamente

2400 marcadores analisados em eucalipto, uma vez que o tamanho do genoma do eucalipto é igual a 1200cM (D. Gratappaglia, comunicação pessoal).

Tabela 2, Acurácias esperadas através de simulações determinísticas para uma população com Ne= 10, distância entre marcadores de 0,005 Morgans, herdabilidade de 20%, 30% e 40%, um número variável de locos (N) controlando a característica um nível de LD igual a 0,83 (proporção da variação genética explicada pelos marcadores (r^2_{mq})).

Herdabilidade (h2)	N	Acurácia	r^2_{mq}	Distância entre marcas(M)
0,2	50	0,77	0,83	0,005
0,2	100	0,67	0,83	0,005
0,2	150	0,61	0,83	0,005
0,2	200	0,56	0,83	0,005
0,2	300	0,49	0,83	0,005
0,3	50	0,81	0,83	0,005
0,3	100	0,73	0,83	0,005
0,3	150	0,68	0,83	0,005
0,3	200	0,63	0,83	0,005
0,3	300	0,56	0,83	0,005
0,4	50	0,83	0,83	0,005
0,4	100	0,77	0,83	0,005
0,4	150	0,72	0,83	0,005
0,4	200	0,68	0,83	0,005
0,4	300	0,61	0,83	0,005

Comparando a Tabela 1 com 2343 marcadores com a Tabela 2 para herdabilidade igual a 20%, é possível inferir que o número mais provável de locos que controlam as duas características é um número próximo de 100 para a altura (acurácia de 67% na Tabela 2, próximo de 70% na Tabela 1) e entre 100 e 150 locos para DAP (acurácia de 67% e 61% na tabela 2, próximo de 66% obtido na tabela 1). Para a herdabilidade de 30%, a mesma análise indica que

aproximadamente 100 locos controlam altura (acurácia de 73% na tabela 2 e de 74% na tabela 1) e 150 locos para DAP (acurácia de 68% na Tabela 2 e de 67% na Tabela 1). Mais uma vez, para herdabilidade igual a 40% pôde-se observar o mesmo padrão de 100 locos para altura total (acurácia de 77% em ambas as tabelas) e de 150 locos para DAP (acurácia de 72% na tabela 2 de 70% na tabela 1).

Estes resultados reforçam a hipótese de que um número maior de locos controlam a característica DAP quando comparado com a característica altura total nesta população.

3.1.2 VALIDAÇÃO CRUZADA: POPULAÇÃO DE ESTIMAÇÃO vs POPULAÇÃO DE VALIDAÇÃO

Para a validação cruzada foram utilizados 83 indivíduos para a validação e 696 indivíduos na população de estimação via uma estratégia de reamostragem *Jacknife*, o que forneceu independência entre os dados de estimação e validação. Neste esquema, todos os fenótipos dos 783 indivíduos foram usados na validação da população e submetidos as análises de correlação com os valores genômicos preditos usando os efeitos dos marcadores estimados

Na Tabela 3 são mostrados os resultados de correlação entre os valores genômicos preditos e os fenótipos dos indivíduos na população de validação associados às três diferentes medidas de herdabilidades. Nesta análise, calculou-se as medidas de correlação quando um número variável de marcadores significativos foram utilizados na estimação dos efeitos.

Tabela 3, Capacidade de predição do fenótipo pela seleção genômica. Correlação entre os valores genômicos preditos e os fenótipos dos indivíduos na população de validação associados a herdabilidades de 20, 30 e 40% e um número de marcadores variando de um até o total de marcas significativas.

2343 Marcadores						
Número de marcadores	Altura h^2= 20%	Altura h^2= 30%	Altura h^2= 40%	DAP h^2= 20%	DAP h^2= 30%	DAP h^2= 40%
1	-0,03	-0,01	-0,01	-0,05	-0,07	-0,07
5	0,09	0,07	0,08	0,06	0,01	-0,01
10	0,12	0,08	0,07	0,15	0,09	0,06
20	0,17	0,10	0,09	0,20	0,19	0,17
30	0,20	0,14	0,12	0,22	0,23	0,23
40	0,22	0,19	0,16	0,25	0,25	0,26
50	0,24	0,21	0,18	0,27	0,27	0,28
100	0,28	0,27	0,26	0,32	0,33	0,32
200	0,29	0,30	0,29	0,39	0,39	0,39
300	0,30	0,31	0,30	0,41	0,41	0,42
400	0,30	0,31	0,30	0,42	0,43	0,44
600	-	-	-	0,42	0,43	0,44
900	-	-	-	0,42	0,43	0,44

Estes resultados demonstram que as correlações estabilizaram quando foram utilizados os 200 marcadores de maiores efeitos para altura e os 300 marcadores para DAP (Figura 1). Estes resultados também mostram que o valor máximo das correlações variou entre 0,30, 0,31 e 0,30 para a altura e 0,42, 0,43, e 0,44 para DAP associados às herdabilidades de 20, 30 e 40%, respectivamente. Isto demonstra que as medidas de correlações são robustas para a incerteza sobre o valor paramétrico da herdabilidade.

Os valores de correlação foram maiores para DAP, o que sugere que esta característica, embora seja controlada por um número maior de locos tenha herdabilidade mais alta. Estes resultados indicam que 200 marcadores para altura

e 300 para DAP estão explicando, respectivamente, 100 locos para altura e 150 para DAP.

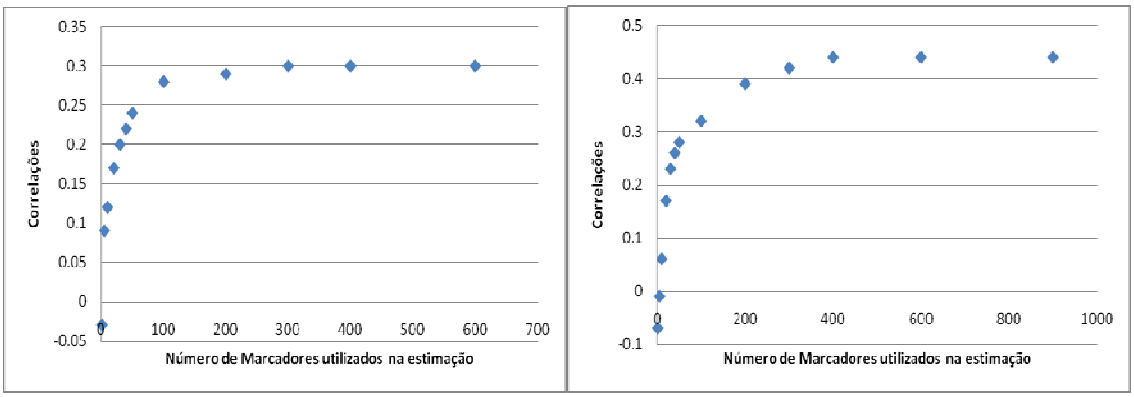

Figura 1 – Correlações entre os valores genômicos preditos e os valores fenotípicos (Altura e DAP, respectivamente) associados às herdabilidades de 20% para altura e 40% para DAP.

É importante salientar que os valores de correlação da Tabela 3 não são as acurácias da seleção genômica. Estas correlações são chamadas de capacidade preditiva da seleção genômica e sua acurácia é calculada ao remover a influência da herdabilidade dividindo a correlação pela raiz quadrada da herdabilidade associada.

Tabela 4, Acurácias da seleção genômica ampla calculadas na população de validação associadas a herdabilidades de 20, 30 e 40% e número de marcadores utilizados variando de um até o total de marcadores significativos.

2343 marcadores						
Número de marcadores	Altura h^2= 20%	Altura h^2= 30%	Altura h^2= 40%	DAP h^2= 20%	DAP h^2= 30%	DAP h^2= 40%
1	-0,06	-0,02	-0,01	-0,12	-0,13	-0,12
5	0,20	0,13	0,13	0,14	0,01	-0,02
10	0,26	0,15	0,12	0,33	0,16	0,10
20	0,39	0,17	0,14	0,44	0,34	0,28
30	0,45	0,26	0,20	0,49	0,42	0,36
40	0,50	0,35	0,25	0,57	0,46	0,41
50	0,54	0,39	0,29	0,60	0,49	0,44
100	0,63	0,50	0,41	0,72	0,60	0,51
200	0,65	0,54	0,46	0,86	0,71	0,62
300	0,67	0,56	0,47	0,91	0,75	0,67
400	0,67	0,57	0,48	0,94	0,78	0,70
600	-	-	-	0,94	0,78	0,69
900	-	-	-	0,93	0,78	0,69

Ao transformar a capacidade preditiva na acurácia de seleção pôde-se observar o impacto que a herdabilidade causa nos valores de acurácia. Para altura, a acurácia na validação foi de 67% na herdabilidade de 20%, 57% na herdabilidade de 30% e 48% na herdabilidade de 40%. Todos estes valores são consistentes com o valor máximo de acurácia obtido anteriormente via PEV na população de estimação os quais gerou 70, 74 e 77% nas três herdabilidades.

Observando a proximidade entre os valores de 67% na validação e de 70% na população de estimação para herdabilidade igual a 20% conclui-se que esta parece ser a herdabilidade mais plausível para a característica altura. Assim, 67% pode ser considerada como a acurácia mais provável para a seleção genômica no crescimento em altura. Este resultado é coerente com as expectativas teóricas apresentadas na Tabela 2 para uma herdabilidade de 20%, um número

aproximado de 100 locos controlando a característica, um Ne de 10 e uma distância entre marcadores de 0,005M.

Para DAP, a acurácia na população de validação variou entre 93, 78 e 69% para as herdabilidades de 20, 30 e 40%, respectivamente. Nem todos estes resultados são plausíveis uma vez que o valor máximo de acurácia possível obtido via PEV (Tabela 1) foi de 66, 67 e 70%, respectivamente para as três herdabilidades testadas. Apenas os valores de acurácia quando a herdabilidade foi de 40% teve resultados coerentes. Nos outros casos os valores são superestimados. Observando a proximidade dos valores de 69% na validação e de 70% na estimação pôde-se concluir que esta parece ser a herdabilidade desta característica nesta população. Assim, 69% pode ser considerado como o valor mais provável de acurácia da seleção genômica ampla para a característica DAP. Mais uma vez esses resultados são coerentes com as expectativas teóricas (Tabela 2) para uma característica com h² de 40%, controle por aproximadamente 150 a 200 locos, Ne igual a 10 e distância entre marcadores igual a 0,005M.

É importante apontar que estas seriam as herdabilidades ajustadas, uma vez que os dados fenotípicos foram inicialmente ajustados pela correção para os efeitos macro ambientais e os efeitos para as capacidades específicas de combinação.

As herdabilidades estimadas no experimento de campo inteiro de onde as árvores foram amostradas foram de 27 e 37%, respectivamente, para altura e DAP. Usando estes valores as acurácias na população de validação seriam 60% e 72%, respectivamente.

Quando a característica altura foi submetida à análise bayesiana pelo método BayesA, os efeitos dos marcadores estimados proporcionaram valores superiores de acurácia que variaram de 0,83 para herdabilidade de 20 % e 0,68 e 0,59 para as herdabilidades de 30 e 40 %. Estes valores são coerentes com os estudos de Meuwissen et al, (2001) que obtiveram acurácias pela metodologia Bayes A de 0,80, superiores também às acurácias de 0,73 obtidas vias BLUP.

3.2 POPULAÇÃO DA FIBRIA.

3.2.1 ACURÁCIA REALIZADA NA POPULAÇÃO DE ESTIMAÇÃO

As acurácias obtidas via Variância do Erro de Predição (PEV) foram superiores às obtidas para a população da CENIBRA (Tabela 5). Como as herdabilidades estimadas a partir dos dados do experimento de campo completo de onde foram amostrados os indivíduos foram superiores aos padrões observados na literatura, estas estimativas, de 0,6 para Altura total e Pilodyn, e 0,7 para DAP foram utilizadas em todos os cálculos. É importante salientar que estes valores de herdabilidade foram ajustados para o efeito de falhas, razão pela qual a herdabilidade estimada é mais alta que o padrão.

Tabela 5, - Acurácias realizadas na população de estimação obtidas via PEV e inversão da matriz dos coeficientes da equação de modelos mistos associadas às herdabilidades calculadas a partir dos dados obtidos no experimento de campo.

H^2	Altura	DAP	Pilodyn
	3564 marcadores	3564 marcadores	3564 marcadores
0,6	0,89	-	0,87
0,7	-	0,83	-

Nesta população foi observado mais uma vez que a característica DAP parece ser controlada por um número maior de locos por obter acurácias mais baixas. Era esperado que os valores de acurácia esperado fossem inferiores uma vez que quanto menor o desequilíbrio de ligação, menores as acurácias esperadas e o desequilíbrio de ligação é inversamente proporcional ao tamanho efetivo. Uma possível explicação para esse fato é a composição da população que foi formada por híbridos de várias espécies de *Eucalyptus*. Esta hibridação, principalmente interespecífica, gera desequilíbrio de ligação, embora este não seja permanente de uma geração para a outra. Dessa maneira, não foi possível fazer uma comparação das acurácias obtidas via PEV com as expectativas de acurácias obtidas na abordagem determinística de Resende *et al*. (2008). (Tabela 6). Isto acontece porque a abordagem determinística de Resende *et al*. (2008) foi feita

baseada na fórmula $E(r^2) = 1/(1+4N_ec)$ proposta por Sved (1971) para o cálculo da medida de desequilíbrio de ligação r^2: em que N_e é o tamanho efetivo populacional e *c* é a taxa de recombinação entre o marcador e o QTL. Esta fórmula é utilizada para calcular o LD entre dois locos em uma situação em que a população, após muitas gerações, que se encontra em equlíbrio. Como a estrutura da população da FIBRIA foi gerada por uma hibridação interespecífica, foi gerado também um desequilíbrio transitório, que não permanece na geração subseqüente.

Tabela 6 - Acurácias esperadas através de simulações determinísticas para uma população com distância entre marcadores de 0,0036 Morgans, herdabilidade de 60% e 70%, um número variável de locos (N) controlando a característica e um nível de LD igual a 0,36 e 0,9 (proporção da variação genética explicada pelos marcadores (r^2_{mq})).

Caractere	Herdabilidade (h^2)	N	Acurácia	r^2_{mq}	Distância entre marcas(M)
ALT	0,6	100	0,53	0,36	0,0036
	0,6	100	0,88	0,9	0,0036
DAP	0,7	150	0,52	0,36	0,0036
	0,7	200	0,50	0,36	0,0036
	0,7	200	0,83	0,9	0,0036

Pôde-se observar ao analisar a Tabela 6 que as acurácias esperadas considerando uma distância entre marcadores de 0,0036M, um número pré estabelecido de 100 e 150 locos baseado no conhecimento da outra população estudada, uma herdabilidade de 0,6 para Altura e Pilodyn e 0,7 para DAP e para uma medida de r^2_{mq} (proporção da variância genética explicada pelos marcadores) derivada de um tamanho efetivo de 120 foi de 0,53 e 0,52, Com os valores observados de acurácia via PEV e fixando o mesmo número de locos que controlou as características na população da CENIBRA, concluiu-se que o valor de r^2_{mq} em ambos os casos foi de 0,9, Entretanto, este r^2 é transitório e não permanece na próxima geração.

3,2,2 VALIDAÇÃO CRUZADA

A Tabela 7 mostra os resultados da capacidade preditiva da SG ao avaliar a correlação entre o fenótipo observado e o fenótipo predito a partir dos efeitos estimados dos marcadores.

Tabela 7 – Capacidade preditiva dos fenótipos via seleção genômica ampla calculadas na população de validação associadas a herdabilidades de 50, 60 e 70% e número de marcadores utilizados variando de um até o total de marcadores significativos.

3564 Marcadores				
Número de Marcadores	Altura – $h^2 = 0,6$	DAP – $h^2 = 0,7$	Pilodyn – $h^2 = 0,6$	Pilodyn – $h^2 = 0,5$
1	0,01	0,05	-0,04	-0,03
10	0,07	0,11	0,07	0,06
20	0,12	0,20	0,18	0,13
30	0,18	0,27	0,24	0,15
40	0,20	0,30	0,28	0,17
50	0,22	0,32	0,29	0,19
100	0,30	0,42	0,35	0,30
200	0,35	0,47	0,34	0,34
300	0,37	0,50	0,34	0,35
400	0,38	0,51	0,34	0,35
600	0,39	0,52	0,34	0,35
900	0,40	0,51	0,34	0,35
2000	0,40	0,51	-	-
3000	-	-	-	-

Pôde-se observar na Tabela 7, que assim como na população da CENIBRA, a medida de correlação foi robusta e praticamente não alterou ao variar a herdabilidade da característica Pilodyn de 0,5 para 0,6 (0,34 para h^2 igual 0,50% e 0,35 para h^2 igual a 60%). A magnitude da correlação é maior quando o valor

paramétrico da herdabilidade é maior pois o fenótipo observado reflete de maneira mais precisa o valor genético. Assim, é possível concluir que a variável DAP realmente tem um valor paramétrico de herdabilidade superior às demais características. Por esta mesma razão, uma nova análise foi realizada para a característica Pilodyn com herdabilidade 0,5 uma vez que as correlações desta variável foram inferiores quando comparadas à variável altura total em um mesmo valor de herdabilidade (0,6).

A Tabela 8 mostra os resultados da acurácia de seleção para as três características, Altura, DAP e Pilodyn, obtidas após a estimação dos efeitos de um número variável de marcadores. Pôde-se observar que as acurácias na Tabela 8 foram inferiores às acurácias obtidas via PEV na população de estimação (0,89, 0,83 e 0,87 para Altura, DAP e Pilodyn) uma vez que esta última apresentou valores altos e diferentes do esperado devido a estrutura da população.

Os valores observados na Tabela 8 estabilizaram seu cresimento quando aproximadamente 300 marcadores foram utilizados nas características Altura e DAP e 200 marcadores utilizados na característica Pilodyn. Este resultado indica que um número menor de marcadores explica a característica Pilodyn ao ser comparado com a variável Altura e DAP

Tabela 8 - Acurácias da seleção genômica ampla calculadas na população de validação associadas a herdabilidades de 50, 60 e 70% e número de marcadores utilizados variando de um até o total de marcadores significativos.

3564 Marcadores				
Número de Marcadores	Altura– h^2 = 0,6	DAP– h^2= 0,7	Pilodyn – h^2 = 0,6	Pilodyn – h^2 = 0,5
1	0,00	0,06	-0,04	-0,05
10	0,10	0,13	0,08	0,10
20	0,16	0,24	0,17	0,25
30	0,22	0,32	0,20	0,34
40	0,25	0,36	0,22	0,39
50	0,29	0,39	0,24	0,42
100	0,39	0,50	0,39	0,49
200	0,44	0,56	0,44	0,48
300	0,47	0,60	0,46	0,47
400	0,49	0,61	0,45	0,48
600	0,51	0,62	0,45	0,48
900	0,51	0,61	0,45	0,48
2000	0,51	0,61	-	-
3000	-	-	-	-

3.2.3 VALIDAÇÃO EM POPULAÇÃO COM FAMÍLIAS DIFERENTES

Para fazer a validação em indivíduos oriundos de famílias diferentes das famílias utilizadas na população de estimação, foram removidos 102 indivíduos que juntos compuseram a população de validação. Os demais 818 indivíduos formaram a população de estimação. Os valores das correlações estimadas para as três características baseado nas herdabilidades de 0,6, 0,7, e 0,6 para Altura, DAP e Pilodyn, respectivamente, são mostrados na Tabela 9,

Tabela 9 – Capacidade preditiva da seleção genômica ampla calculadas na população de validação associadas a herdabilidades de 50, 60 e 70% e número de marcadores utilizados variando de um até o total de marcadores significativos.

3564 Marcadores				
Número de Marcadores	Altura– h² = 0,6	DAP– h²= 0,7	Pilodyn – h² = 0,6	Pilodyn – h² = 0,5
1	0,10	0,22	0,10	0,10
10	0,14	0,22	0,15	0,12
20	0,21	0,33	0,22	0,18
30	0,21	0,36	0,30	0,26
40	0,29	0,41	0,29	0,31
50	0,32	0,44	0,30	0,30
100	0,31	0,50	0,41	0,42
200	0,40	0,52	0,42	0,41
300	0,42	0,53	0,40	0,42
400	0,41	0,52	0,40	0,41
600	0,41	0,53	0,40	0,41
900	0,41	0,53	-	-
2000	-	-	-	-

As medidas de correlações observadas na Tabela 9 são praticamente iguais às medidas obtidas na validação cruzada dependente. para as variáveis altura total e DAP (Tabela 7). Os valores de correlação para a variável Pilodyn, embora constantes nas duas herdabilidades testadas, foram um pouco superiores aos obtidos na Tabela 7 para validação cruzada dependente. É possível que os valores obtidos via jacknife na validação cruzada (0,34 para a variável Pilodyn) sejam melhor estimados uma vez que foram utilizados 10 repetições, ao contrário desta validação independente, onde apenas uma repetição foi feita.

Os valores das acurácias de seleção estimada para as três características são apresentados na Tabela 10, Neste caso, as análises de seleção genômica

ampla não foram realizadas na herdabilidade 0,5 para a característica Pilodyn uma vez que altura e Pilodyn obtiveram os mesmos valores de correlação (0,41).

Tabela 10 - . Acurácias da seleção genômica ampla calculadas na população de validação associadas a herdabilidades de 60 e70% e número de marcadores utilizados variando de um até o total de marcadores significativos.

3564 Marcadores			
Número de Marcadores	Altura– h^2 = 0,6	DAP– h^2= 0,7	Pilodyn – h^2 = 0,6
Marcadores significativos →	1081	1414	816
1	0,12	0,26	0,14
10	0,19	0,33	0,15
20	0,27	0,40	0,23
30	0,27	0,41	0,33
40	0,37	0,42	0,40
50	0,42	0,44	0,39
100	0,40	0,56	0,54
200	0,52	0,60	0,53
300	0,54	0,62	0,54
400	0,53	0,62	0,53
600	0,53	0,63	0,53
900	0,53	0,62	0,53
2000	0,53	0,62	-
3000	-	-	-

Nesta validação, os valores de acurácias se estabilizaram quando foram utilizados 200 marcadores para as características de Altura e DAP e 100 marcadores para a característica Pilodyn. Estes resultados, diferente do esperado, foram próximos, porém inferiores dos valores estimados na validação cruzada dependente. Uma possível razão para este fato é o uso de repetições para

estimações dos valores na validação cruzada, fato que não pode ser realizado nesta validação. De qualquer maneira, os resultados revelam a eficiência do procedimento Jacknife em produzir uma validação independente.

Uma vez que estas acurácias foram obtidas nesta população de validação independente, concluí-se que estes valores poderiam ser também obtidos caso a seleção genômica fosse aplicada em todo o teste de progênie de onde estes indivíduos foram amostrados. Novos estudos devem agora ser conduzidos, para que uma nova validação seja feita nesta mesma população, porém na próxima geração de recombinação. Assim, isto permite que uma vez estimado os efeitos dos marcadores e devidamente validados, após a próxima recombinação e montagem do próximo teste de progênie, estas progênies podem ser selecionadas em estágios iniciais de mudas com a acurácia esperada obtida nesta população de validação de segunda geração.

Em próximos estudos considerando estas duas populações da FIBRIA e da CENIBRA, serão comparados quantos dos marcadores significativos para uma determinada característica em uma população foi também significativa para a outra. Será feito ainda a validação do modelo predito em uma empresa, na população da outra empresa. Espera-se que esta validação não proporcionará estimativas de acurácias altas, pois em diferentes populações, diferentes alelos podem estar presentes e os efeitos de marcadores estimados em uma população provavelmente não terão efeito em outra. No entanto, para algumas características, como resistência a doenças que são controlados por um número menor de genes, pode-se obter um modelo generalizado para aplicação em diferentes regiões do país.

3.3 GANHOS DE SELEÇÃO

As estimativas de ganho de seleção foram calculadas ao comparar a eficiência da seleção genômica ampla com o valor máximo obtido com a seleção fenotípica. Como uma das aplicações da GWS é a seleção precoce, o ganho foi avaliado por unidade de tempo, considerando assim a redução no ciclo de melhoramento de 25, 50 e 75% (Tabela 11).

Tabela 11 – Superioridade da seleção genômica ampla em relação a acurácia máxima possível de obter via seleção fenotípica (Acur F – 0,68 para h² igual a 20% e 0,80 para h² igual a 60%) e em função da redução do tempo de geração convencional em anos do melhoramento fenotípico (Temp F) para tempos menores (Temp G)

			Pop. CENIBRA			Pop. Fibria		
Acur F	Temp F	Temp G	ALT/ 0,67	DAP/ 0,70	**Acur F**	ALT/ 0,53	DAP/ 0,62	Pilodyn/ 0,53
0,68	8	4	97,06	105,88	**0,80**	32,50	55,00	32,50
0,68	8	3	162,75	174,51	**0,80**	76,67	106,67	76,67
0,68	8	2	294,12	311,76	**0,80**	165,00	210,00	165,00
0,68	8	1	688,24	723,53	**0,80**	430,00	520,00	430,00

É possível observar, que com a densidade de marcadores usada, os valores de acurácias obtidos via seleção genômica são inferiores aos valores máximos obtidos na seleção fenotípica. Neste experimento piloto, o grande impacto que esta tecnologia pode trazer é a possibilidade de seleção genômica precoce, que proporciona ganhos em torno de 100% na população da CENIBRA e de 40-50% na população da FIBRIA, caso o ciclo de geração de novos materiais seja reduzido pela metade.

Considerando novas pesquisas sobre alguns fatores cruciais para a aplicação da GWS como formas de induzir o florescimento precoce para recombinação dos genitores selecionados no teste de progênie, a precocidade da seleção pode reduzir o tempo de geração em um programa de melhoramento em até 7 anos. Esta redução proporciona ganhos por unidade de tempo de até 724%.

Percebe-se que aqui a GWS pode revolucionar a forma como é feita a seleção nos programas de melhoramento, principalmente de espécies perenes. É importante salientar ainda que, com o desenvolvimento de um número maior de marcadores, o desequilíbrio de ligação entre o loco e um marcador aumenta, o que proporciona medidas de acurácias que podem, per se, ultrapassarem os valores de acurácia via seleção fenotípica.

4. CONCLUSÃO

Com este estudo pioneiro em espécies vegetais, pôde-se concluir que a seleção genômica ampla tem grande potencial de aplicação e pode gerar um elevado ganho, principalmente em programas de melhoramento de espécies de longo ciclo onde a seleção precoce tem um maior potencial na geração de ganhos.

Além disso, percebe-se que, em aplicações práticas, é necessário a genotipagem do número máximo de marcadores disponíveis para aumentar a probabilidade de se encontrar vários marcadores em LD com os QTLs. No entanto, após a estimação dos efeitos dos marcadores, a acurácia máxima já é obtida com a estimação baseada no uso de um número menor de marcadores. Neste caso, em genotipagens futuras, é necessária a genotipagem apenas destes marcadores, que além de terem sido significativos para o teste de associação com o fenótipo, geraram estimativas de acurácias de magnitude aproximadamente igual às acurácias obtidas quando todos os marcadores foram utilizados.

Pôde se concluir também, que o procedimento *Jacknife*, cuja aplicação foi avaliada neste trabalho para obter um modo eficiente de validação, foi efetivo e pode ser utilizado para validar os efeitos estimados de marcadores.

Além disso, observou-se neste trabalho que os ganhos de seleção comparados apenas pela magnitude da acurácia e da diminuição do tempo de seleção foram expressivos. No entanto, novos estudos devem ser feitos, pois o valor agregado e competitivo de se liberar um material selecionado 8 anos antes de um concorrente pode atingir escalas muito altas aumentando ainda mais o impacto que o uso da GWS pode ter no melhoramento florestal.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Buckler et. al. (2009). The genetic architecture of maize flowering time. Science. 2009 Aug 7;325(5941):714-8.

Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. Journal of Animal Science 82:313-328.

Eckert A., Wegrzyn J.L., Pande B., Jermstad K.D., Lee J.M., Liechty J.D., Tearse B.R., Krutovsky K.V., Neale D.B. (2009) Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold-hardiness in Coastal Douglas-fir (Pseudotsuga menziesii var. menziesii). Genetics. 183:289-298.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler, IV. 2003, Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374,

Goddard, ME, Hayes, BJ (2007). Genomic Selection. J. Anim. Breed. Genet. 124 323–330.

Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10: 381–391.

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD,Neale DB: Association genetics in Pinus taeda L. I. Wood property traits. Genetics 2007, 175:399-409.

Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB: Association genetics in Pinus taeda L. II. Carbon isotope discrimination. Heredity 2008, 101:19-26.

Grattapaglia D (2007a) Mapas genéticos e seleção assistida por marcadores moleculares. In: Borem A (ed) Biotecnologia Florestal. Editora UFV, Viçosa, pp 201-230.

Grattapaglia D (2007b) Marker assisted selection in Eucalyptus. In: Guimarães E, Ruane J, Scherf B, Sonnino A, Dargie J (eds) Marker assisted selection - Current status and future perspectives in crops, livestock, forestry and fish, pp 251-283.

Grattapaglia D, Kirst M (2008) Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytologist 179:911-929.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. Curr Opin Plant Biol 12:148-156.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard, 2009 Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443,.

Meuwissen T, Hayes B, Goddard M (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829.

Meuwissen T, Solberg TR, Shepherd R, Wooliams JA (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution. 41:2.

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

Neale DB, Savolainen O: Association genetics of complex traits in conifers. Trends Plant Sci 2004, 9:325-330.

O´Malley *et. al.* (1996). Molecular markers, Forest genetics and tree breeding. In Gustafson, JP, Flavell, RB. Genomes of plants and animals. New York: Plenum Press. P. 87-102.

Resende MDV, Lopes PS, Silva RL, Pires IE (2008) Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. Pesquisa Florestal Brasileira 56:63-77.

Resende, MDV (2002). Genética biométrica e estatística no melhoramento de plantas perenes. Brasília: Embrapa Informação Tecnológica, 2002, 975p.

Resende, MDV (2007a). Matemática e estatística na análise de experimentos e no melhoramento genético. Colombo: Embrapa Florestas. 561p.

Resende, MDV (2007b). Selegen–Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos. Colombo: Embrapa Florestas, 361 p.

Resende MDV (2008) Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais. Colombo:EMBRAPA Florestas. 330p.

Resende MDV, Assis, TF (2008). Seleção Recorrente Recíproca entre populações sintéticas multi-espécies (SRR-PSME) de Eucalipto.Pesquisa Florestal Brasileira, Colombo n. 57:57-60.

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123:218-223.

Sewell M.M., Bassoni D.L., Megraw R.A., Wheeler N.C. , and Neale, D.B. (2000) Identification of QTLs influencing wood property traits in loblolly pine (Pinus taeda L.) I Physical wood properties. Theoretical Applied Genetics. 101:1273-1281.

VanRaden, P.M. et al. 2009 J. Dairy Sci. 92:16.

Visscher, PM, Montgomery, GW (2009). Genome-wide association studies and human disease: from trickle to flood. JAMA.;302(18):2028-9.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915-9920.

Wilcox PL, Echt CE, Burdon RD (2007) Gene-assisted selection: applications of association genetics for forest tree breeding Associated Mapping in Plants:211-247.

Wong, C.K., and R. Bernardo. 2008, Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. Theor. Appl. Genet. 116:815–824.

Xu Y, Crouch JH (2008) Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Sci 48:391-407.

Zhong, S, Dekkers, JCM., Fernando, RL., Jannink, JL (2009). Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study Genetics. 182: 355-364 .