Capítulo X

Seleção Genômica Ampla

Marcio F. R. Resende Jr.
Alexandre Alonso Alves
Carlos Felipe Barrera Sánchez
Marcos Deon Vilela de Resende
Cosme Damião Cruz

1. Introdução

O melhoramento genético tem, há vários anos, proporcionado com muito sucesso o aumento de produtividade e a melhoria de várias características de interesse na agricultura e na pecuária. Embora muitos métodos tenham surgido ao longo dos anos no intuito de otimizar o processo de seleção de indivíduos mais produtivos e, ou, que exibem características de interesse, a estratégia básica utilizada até hoje é 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, existe a expectativa de que informações genotípicas (obtidas por meio dos marcadores moleculares), uma vez correlacionadas com características fenotípicas de interesse, possam ser amplamente utilizadas na identificação e seleção de indivíduos com maiores valores genéticos. Adicionalmente, espera-se que a seleção com base em informações genotípicas possa ser realizada precocemente, o que no caso do melhoramento animal, ou de espécies vegetais perenes, tende a elevar os ganhos (RESENDE, 2008).

Nesse sentido, o primeiro método proposto para o uso de marcadores no melhoramento ficou conhecido como seleção assistida por marcadores (MAS) (LANDE; THOMPSON, 1990; PATERSON et al., 1991). Essa metodologia se baseia na análise de progênies segregantes (uma ou algumas famílias) para identificação e mapeamento de regiões controladoras de características quantitativas (QTL – *Quantitative trait Loci*). A geração de progênies oriundas de poucos indivíduos gera alto desequilíbrio de ligação (DL) dentro de famílias ou cruzamentos, o que permite a identificação de marcadores que cossegregam com a característica fenotípica, mesmo quando reduzido número de marcadores é utilizado. O conceito de DL refere-se à associação não aleatória entre dois genes ou entre um QTL e um marcador. Nesse caso, quando as frequências alélicas e genotípicas de dois locos são constantes de uma geração para a outra e as frequências genotípicas são determinadas pelas frequências alélicas, diz-se que esses locos encontram-se em equilíbrio de Hardy-Weinberg e de ligação (EL). Em razão da ligação gênica, dois genes/marcadores ligados apresentam associação que não se dá ao acaso, e diz-se, então, que esses genes estão em DL. Dado ao grande interesse pela técnica da MAS, um grande número de QTLs foi detectado e mapeado nas mais variadas culturas (COOPER et al., 2009; FRARY et al., 2000; YANO et al., 2000). No entanto, grande parte desses QTLs detectados e mapeados em cada espécie não foi aplicado de forma prática nos seus programas de melhoramento (BERNARDO, 2008). Uma das causas desse insucesso é a necessidade do estabelecimento de associações entre os marcadores e os QTLs de cada família avaliada. Isso acontece, pois os níveis populacionais de desequilíbrio de ligação em uma população de melhoramento são muito inferiores quando comparados com o desequilíbrio analisado na progênie segregante. Além disso, uma segunda razão que limitou o uso prático da SAM foi o fato de apenas pequeno número de QTLs de grande efeito ter sido detectado e mapeado, os quais, devido à natureza poligênica e à alta influência ambiental dos caracteres quantitativos, não explicam suficientemente toda a variação genética (DEKKERS, 2004).

Em razão das limitações inerentes à técnica de MAS (revisados em maiores detalhes a seguir), novas metodologias foram propostas para utilização de marcadores moleculares na identificação de *locos*

que controlam características de interesse ao melhoramento, por exemplo, a genética de associação (GWAS) e a seleção genômica ampla (GWS). A proposição dessas novas metodologias foi somente possível graças ao extraordinário avanço das tecnologias de genotipagem. A partir do início do século XXI, métodos que permitiam a descoberta e genotipagem em larga escala de Single Nucleotide Polymorphism (SNPs) em plataformas de microarranjos multiplicaram-se, de modo que hoje a maioria das espécies de interesse econômico dispõe de número relativamente elevado (na ordem de algumas centenas/milhares) de marcadores passíveis de uso em programas de melhoramento (JENKINS; GIBSON, 2002). Com o advento de tecnologias de genotipagem por sequenciamento (Genotyping by Sequencing, GBS) (BAIRD et al., 2008; ELSHIRE et al., 2011; RESENDE et al., 2012D), espera-se que haja incremento ainda maior no número de marcadores disponíveis, concomitantemente com a redução do preço por *data point*, o que tende a viabilizar o uso rotineiro de marcadores em programas de melhoramento.

A teoria da genética de associação e da seleção genômica baseia-se no fato de que, com grande número de marcadores espalhados pelo genoma, aumenta-se a probabilidade de que QTLs de interesse estejam em forte DL com os marcadores (HASTBACKA et al., 1992). No entanto, essas metodologias diferem quanto ao modo como os efeitos dos QTLs/marcadores na expressão fenotípica do caráter são ajustados em um modelo biométrico e o tipo de população utilizado. No caso da GWAS, o ajuste é feito, de maneira geral, marcador a marcador em grandes populações naturais, não estruturadas, em que os indivíduos suspostamente se relacionam entre si, por meio de um ancestral comum em dado tempo. O principal intuito dessa técnica é identificar genes candidatos para o controle genético de determinada característica. Já no caso da GWS o ajuste é realizado para todos os QTLs/marcadores simultaneamente, na própria população de melhoramento, e.g. no conjunto de famílias segregantes (híbridas ou não) em teste. O intuito da GWS é obter um modelo que prediz o valor genético do indivíduo, mas que não necessariamente determina genes específicos envolvidos no controle do caráter. Cabe destacar que o tipo de população utilizada tem impacto relevante sobre os padrões de DL e, consequentemente, sobre o número de marcadores necessários

para identificar genes que controlam características de interesse ao melhoramento e selecionar indivíduos superiores. Em razão dessas características e de outras detalhadas a seguir, a GWS tem chamado mais atenção de melhoristas recentemente pela possibilidade real de sua operacionalização em programas de melhoramento (HAYES et al., 2009). Grande número de trabalhos tem sido publicados abordando aspectos teóricos da GWS, assim como aspectos práticos. Nesse contexto, destacam-se, principalmente, os trabalhos de simulação que indicam que o método é altamente acurado (GRATTAPAGLIA; RESENDE, 2011; MEUWISSEN et al., 2001), e mais recentemente os trabalhos de prova de conceito, que indicam que a GWS pode ser efetivamente implementada como diferencial em programas de melhoramento de espécies vegetais perenes e animais (HAYES et al., 2009; RESENDE et al., 2012B; VANRADEN et al., 2009).

Posto isso, o objetivo deste capítulo é delinear as principais características da genética de associação (fazendo um paralelo a MAS) e, posteriormente, apresentar a metodologia de seleção genômica ampla. Abordam-se também os conceitos e a evolução desta última metodologia desde 2001, quando foi proposta (MEUWISSEN et al., 2001), até os dias de hoje. Posteriormente, são apresentados os principais métodos estatísticos utilizados na estimação dos efeitos dos marcadores e algumas de suas aplicações no melhoramento de plantas e animais. Ao final da leitura do capítulo, o leitor será capaz de diferenciar as metodologias de uso de marcadores moleculares para identificação de QTLs e seleção de indivíduos superiores, bem como suas características e peculiaridades. Espera-se também que o leitor possa ter clara noção das potencialidades da GWS para os programas de melhoramento genético de plantas e animais.

2. Genética de Associação

O desenvolvimento de métodos de análise de SNPs em larga escala, aliado à grande quantidade de sequências gênicas (e em alguns casos genomas completos) disponíveis publicamente, tem possibilitado o mapeamento de características complexas por meio da estratégia de associação. O princípio do genética de associação é semelhante àquele utilizado nos métodos de mapeamento de QTLs

em famílias segregantes, em que se identificam locos cuja frequência alélica está correlacionada com a variação fenotípica em uma população. Nos dois métodos, diferenças significativas entre os valores fenotípicos observados nos indivíduos que herdaram alelos distintos sugerem que o loco está em desequilíbrio de ligação com o loco que efetivamente controla a característica fenotípica.

O mapeamento de QTLs (ou mapeamento de ligação, ML) baseia-se na análise de uma população com *pedigree* conhecido. Como a população analisada normalmente passa por apenas um ou dois cruzamentos, o número de eventos de recombinação entre dois locos localizados em um mesmo grupo de ligação é limitado. Dessa forma, os blocos em DL são extensos (> 10 cM), e assim, quando se identifica um marcador molecular associado à variação fenotípica, o loco causador pode estar na realidade a uma distância genética de vários centimorgans (cM) desse marcador. O nível de resolução pode ser melhorado analisando número maior de indivíduos na população segregante e saturando o mapa genético com número maior de marcadores, como se faz no mapeamento fino. Ainda assim, a resolução dificilmente será menor do que 1 cM, o que ainda representa enorme porção física do genoma. Para fins de comparação de 1 cM equivale a 15 Mpb em espécies como *Pinus taeda* (Kirst, 2007) ou a 300 Kpb em espécies de *Eucalyptus* (Grattapaglia; Bradshaw, 1994).
O mapeamento de associação baseia-se, entretanto, na análise de uma população de indivíduos não relacionados. A razão dessa condição está relacionada com o objetivo do mapeamento de associação, em que se querem detectar e mapear os genes controladores de características fenotípicas. Assim, para que se atinja uma resolução em nível de um gene na determinação de associação entre um marcador e o fenótipo, é necessário que o tamanho do bloco de ligação seja extremamente reduzido. Além disso, o uso de uma população não estruturada permite que os locos amostrados potencialmente capturem toda a variabilidade genética da população em estudo e não apenas a variabilidade dos dois genótipos parentais. No entanto, o número de marcadores necessários para identificar genes associados ao fenótipo é inversamente proporcional à extensão do desequilíbrio de ligação. Assim, como essa ausência de estrutura das populações leva a um reduzido nível de desequilíbrio de ligação, as análises de GWAS idealmente requerem

elevadíssima densidade de marcadores. Por exemplo, em *Eucalyptus grandis*, se a extensão média do desequilíbrio de ligação for de 1.000 pares de base e o genoma tiver 630 Mpb, será necessária a genotipagem de 630.000 marcadores SNPs (GRATTAPAGLIA, 2007). Se o desequilíbrio de ligação estender apenas por 500 pares de bases, o número de marcadores dobra para 1,26 milhão de SNPs. Em outras palavras, com as tecnologias atuais de genotipagem de SNPs, essa abordagem não é possível economicamente.

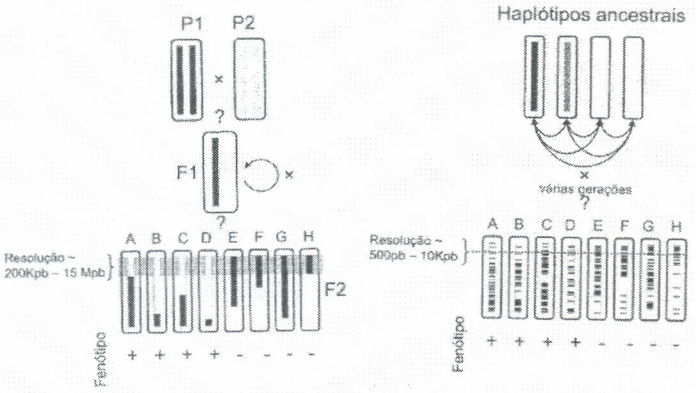


Figura 1.**A** - Mapeamento de QTLs (mapeamento de ligação) em uma população segregante (F_2) com *pedigree* definido. Como apenas um evento de recombinação efetivo ocorreu no genoma dos dois parentais, a resolução representada pela área hachurada é baixa. **B** Mapeamento de associação em uma população não estruturada. Após várias gerações de cruzamento, chega-se à população atual (representada aqui pelos indivíduos A-H). Como o número de eventos de recombinação entre os haplótipos ancestrais e a população atual é alto, a resolução também é muito superior.

Fonte: Adaptado de KIRST, 2007.

A Seleção Genômica, como será detalhada nos próximos tópicos, requer número menor de marcadores e tem outros objetivos. No entanto, para introduzir o leitor aos modelos e métodos utilizados, é

interessante exemplificar uma das formas mais utilizadas para testar a associação entre vários marcadores e uma característica: o teste F de Snedecor da análise de regressão em marca simples, utilizando todos os marcadores, um de cada vez. Essa associação pode ser testada pelo modelo:

$$y = 1u + Xg_i + e$$

em que y é um vetor coluna de fenótipos, 1 é um vetor coluna que contém N (número de indivíduos) vezes o número 1, X é um vetor coluna de incidência que aloca o genótipo de um loco marcador a cada indivíduo, g_i é um escalar contendo o efeito de um dos alelos do loco marcador e é um vetor de erros aleatórios com distribuição $e_{ij} \sim N(0, \sigma_e^2)$, em que σ_e^2 é a variância residual. Um fator importante desse modelo é que o efeito de cada marcador é tratado como fixo. A pressuposição desse modelo é de que o marcador irá afetar o fenótipo, caso esteja em desequilíbrio de ligação com o QTL. Assim, a hipótese de nulidade é de que o marcador não tem nenhum efeito associado ao fenótipo. Entretanto, a hipótese alternativa é de que o marcador está em desequilíbrio de ligação com o QTL e apresenta, assim, associação com o fenótipo. A significância é avaliada a partir de um teste F, normalmente com alguma correção para múltiplos testes (FDR e Bonferroni).

Exemplo: Como exemplo da análise de marca simples, considere um conjunto de 10 plantas genotipadas com um marcador codominante com dois alelos *A* e *a*. Os códigos de genotipagem usados são 0 para *aa*, 1 para *Aa* e 2 para AA.

Indivíduo	Fenótipo (Vetor y)	Genótipo (Vetor X)
1	83,00	2
2	14,55	1
3	7,35	0
4	52,99	2
5	52,29	1
6	12,82	0
7	24,91	2
8	26,52	0
9	49,50	1
10	47,01	2

Segundo esse modelo proposto, a média e efeito do marcador podem ser estimados da seguinte maneira:

$$\begin{bmatrix}\hat{\mu}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}1_n{}' 1_n & 1_n{}' X\\ X' 1_n & X' X\end{bmatrix}^{-1}\begin{bmatrix}1_n{}' y\\ X' y\end{bmatrix}$$

Resolvendo cada termo da equação, tem-se:

$X' X = 19$;

$1_n{}' X = 11$;

$1_n{}' 1 = 10$;

$1_n' y = 370{,}94$; e

$X' y = 532{,}16$.

Assim, resolvendo a equação com os valores anteriores, a média e o efeito do marcador são:

$$\begin{bmatrix}\hat{\mu}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}17{,}31\\ 17{,}99\end{bmatrix}$$

O teste F pode ser calculado de acordo com a fórmula a seguir (RESENDE, 2008):

$$F = \frac{QMRegressão}{\sigma_e^2} = \frac{\hat{g}X' y + \hat{\mu}1_n' y - {}^1/_n (1' y)^2}{(y' y - \hat{g}X' y - \hat{\mu}1_n' y)/(n-2)}$$

Assim:

$$F = \frac{17{,}99 * 532{,}16 + 17{,}31 * 370{,}94 - 370{,}94 * 37{,}094}{\frac{(18845{,}27 - 17{,}99\, x\, 532{,}16 - 17{,}31\, x\, 370{,}94)}{(10-2)}} = 6{,}2717$$

A estatística F deve ser comparada com um valor de F tabelado com 1 e (n-2) graus de liberdade para o numerador e denominador, respectivamente. Nesse caso, ao checar a tabela de valores de F, obtém-se o valor de 5.32 com 5% de significância. Assim, rejeita-se a

hipótese H0 de que o marcador é independente, e, dada a informação disponível, declara-se que o marcador está em associação com a característica. Essa estratégia pode ser usada para identificação de genes candidatos ou de regiões do genoma que indicam associação com o fenótipo.

3. Seleção Genômica Ampla

Um dos grandes atrativos da genética molecular em benefício do melhoramento genético de plantas e, ou, animais é a possibilidade de utilização direta das informações de DNA na seleção. Essa característica permite alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos (RESENDE et al., 2010). Contudo, conforme exposto anteriormente, as metodologias de MAS e GWAS apresentam limitações que restringem sua aplicação direta em programas de melhoramento. Visando efetivamente implementar o uso de marcadores moleculares em programas de melhoramento, um novo método de seleção denominado seleção genômica (GS), ou seleção genômica ampla (GWS), foi proposto por Meuwissen et al. (2001).

A GWS, diferentemente da MAS, pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético. Além disso, a GWS apresenta alta acurácia seletiva para a seleção baseada exclusivamente em marcadores e não exige prévio conhecimento das posições (mapa) dos QTLs, não estando sujeita aos erros tipo II associados à seleção de marcadores ligados a QTLs (RESENDE, 2008). Esse método permaneceu discreto por vários anos, devido ao fato de os marcadores moleculares disponíveis na época serem caros e restritos. Recentemente, entretanto, observou-se um incrível avanço nas tecnologias de genotipagem, de modo que hoje centenas a milhares de marcadores podem ser efetivamente genotipados virtualmente em qualquer população a baixo custo e curto espaço de tempo. Em razão disso, o método tornou-se muito atrativo, e geneticistas e melhoristas renomados têm demonstrado e confirmado a superioridade e exequibilidade prática (BERNARDO; YU 2007; GODDARD, 2009; GODDARD, HAYES, 2007; HEFFNER et al., 2009; RESENDE, 2008; SCHAEFFER, 2006; SMARAGDOV, 2009). Esses

trabalhos evidenciaram, definitivamente, que a seleção genômica terá grande utilidade no melhoramento genético.

Definição

A GWS pode ser, de modo simplificado, definida como um método de predição de fenótipo ou valor genético, visando à seleção de indivíduos superiores para uma ou mais características, com base exclusivamente em informações genotípicas. Nesse caso, é necessária a determinação *a priori* dos efeitos genéticos dos *n* marcadores utilizados na seleção sobre as características de interesse. Essa etapa é realizada em uma amostra da população de melhoramento, comumente chamada de população de estimação, de treinamento ou de descoberta. Essa população, além de genotipada para os marcadores moleculares, é fenotipada para as características de interesse. Os efeitos de todos os marcadores sobre as características de interesse são, então, estimados simultaneamente (de modo contrastante a MAS), e modelos para predizer o valor genético genômico dos indivíduos em gerações futuras são elaborados (detalhado a seguir). Como a densidade de marcadores utilizada na GWS é relativamente elevada, esses cobrem o genoma de maneira densa e idealmente uniforme. Assim, espera-se que cada um dos genes que controlam dado caráter quantitativo esteja em DL com pelo menos um dos marcadores (RESENDE, 2008). Desse modo, pelo fato de todos os QTLs, sejam eles de grande ou pequenos efeitos, estarem em DL com marcadores moleculares, ao utilizar todos os marcadores no modelo preditivo, quase a totalidade da variação genética do caráter quantitativo será capturada. Isso eleva a acurácia da estimação do valor genético-genômico dos indivíduos para valores muito acima daqueles observados para a MAS (BERNARDO; YU, 2007; GODDARD, 2009; GODDARD; HAYES, 2007; RESENDE, 2008). Além disso, como os efeitos dos marcadores são estimados em uma amostra de indivíduos (população de treinamento) pertencentes às várias famílias da população de melhoramento, os padrões de DL amostrados são representativos da população de melhoramento, em que a GWS visa ser aplicada. Isso torna a GWS operacionalmente muito mais interessante que MAS, pois pode ser aplicada em toda a população, não se restringindo a uma família específica. Cabe, no entanto, ressaltar que os efeitos dos marcadores não serão necessariamente

os mesmos em diferentes populações e ambientes, de modo que os modelos preditivos são, em sua maioria, população e ambiente específicos (RESENDE et al., 2012B). A razão dessa especificidade dos modelos em cada ambiente aparenta estar relacionada com a interação genótipo x ambiente (RESENDE et al., 2012B), fator complicador também em programas tradicionais de melhoramento genético.

Tipos de população

Em termos operacionais, a fim de definir os efeitos genéticos dos marcadores no que concernem as características de interesse, duas ou três populações são necessárias. Primeiramente, uma população de descoberta ou treinamento é necessária (Figura 2). Essa população em geral consiste de um número moderado de indivíduos (800-1.000), analisados com grande número de marcadores moleculares (que pode ultrapassar 100 mil, dependendo do tamanho do genoma e da estrutura do DL) e também fenotipados para as características de interesse (GODDARD; HAYES, 2007). Essa população deve representar a população de melhoramento, uma vez que os modelos preditivos serão aplicados efetivamente em gerações futuras dessa população (RESENDE et al., 2010). Conforme já mencionado, o desequilíbrio de ligação populacional é inversamente proporcional ao tamanho efetivo populacional (Ne). Como o número de marcadores moleculares a serem genotipados depende da extensão média do DL na população de treinamento, esta deve ser constituída de modo a manter o Ne em faixas intermediárias, mas compatíveis com o tamanho efetivo usado no programa de melhoramento. Assim, a aplicação da GWS no melhoramento deve ser enfatizada em etapas mais avançadas do programa, quando o número de genitores intercruzados já for restrito. À medida que o preço de genotipagem reduzir será possível a obtenção de número muito elevado de marcadores, o que possibilitaria, assim, a incorporação da GWS no estabelecimento de populações-base. Uma vez levantados os dados fenotípicos e genotípicos (exemplos numéricos são apresentados em outras seções deste capítulo), as equações de predição dos valores genético-genômicos são obtidas conforme descrito a seguir. Simplificadamente, essas equações associam a cada intervalo (delimitado pelos marcadores moleculares), ou a cada marcador individual, o efeito na característica de interesse.

Após a obtenção dos modelos preditivos, esses devem ser avalidados em amostra independente. Para tanto, operacionalmente deve existir uma segunda população, chamada de população de validação (Figura 2). Essa pode ser menor que a população de descoberta, uma vez que os dados levantados serão usados apenas para fins de validação e não de estimação (RESENDE et al., 2010). Do mesmo modo que no caso da população de treinamento, a população de validação deve consistir de indivíduos analisados com os marcadores moleculares e também fenotipados. As equações de predição são testadas nessa população, e a acurácia dos modelos é estimada nessa amostra independente. Para computar essa acurácia, os valores genético-genômicos são preditos usando os efeitos estimados com base na população de treinamento e submetidos à análise de correlação com os valores genéticos obtidos via análise dos dados fenotípicos, conseguidos por metodologias-padrão (BLUP) (MEUWISSEN et al., 2001). Como a população de validação é independente e não é envolvida na predição dos efeitos dos marcadores, os erros associados aos valores preditos e observados são também independentes, e toda a correlação entre esses valores é de natureza genética e indica a capacidade preditiva da GWS (GODDARD; HAYES, 2007; RESENDE, 2008). Maiores considerações sobre a acurácia de predição dos modelos de GWS serão feitas no tópico 4.4.

Uma vez comprovado que os modelos são acurados, estes podem ser utilizados para predição dos valores genético-genômicos dos indivíduos em gerações futuras da população de melhoramento, sem que haja a necessidade de realizar-se a fenotipagem deles para as características de interesse. A acurácia desse processo é equivalente à acurácia estimada na população de validação (GODDARD; HAYES, 2007; RESENDE, 2008).

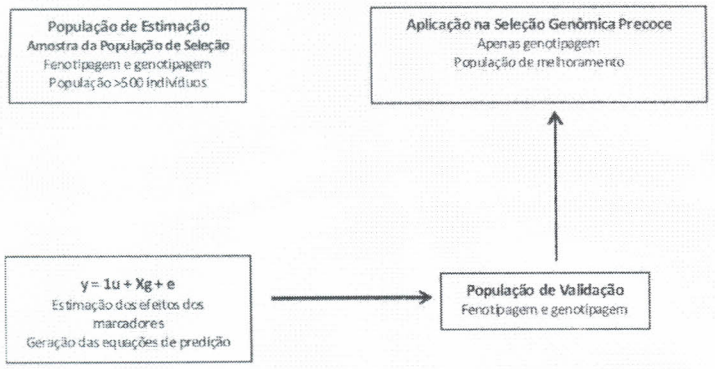


Figura 2 - Esquema de aplicação da seleção genômica ampla.

Validação cruzada

Em alguns casos, em geral devido a limitações na genotipagem de grande número de indivíduos, apenas uma população de tamanho intermediário (~1.000 indivíduos) é genotipada e fenotipada. Nesse caso, a mesma população tem de ser usada para estimação e validação dos efeitos dos marcadores. Nessa situação, para que os efeitos estimados dos marcadores não sejam superestimados devido à estimação e validação na mesma amostra, uma técnica de validação cruzada pode ser adotada (Figura 3). Nesse caso, após a obtenção dos dados genotípicos e fenotípicos, uma parcela da população é utilizada para estimar o modelo, enquanto a parcela restante é utilizada para sua validação. Esse processo, a separação de uma parcela para estimação do modelo e de outra para validação, é repetida n vezes (Ex: 10 vezes), de modo que no final do processo ter-se-ão n validações independentes. Esse processo é conhecido como validação cruzada (LEGARRA et al., 2008; USAI et al., 2009), e o número n de validações independentes pode ser tão pequeno quanto dois (metade da população usada na estimação e a outra metade na validação) ou tão grande quanto o número total de indivíduos. No último caso, essa estratégia fica conhecida como Jacknife ou leave-one-out. Quando aplicada, a estimação dos efeitos genéticos associados aos marcadores é feita com base em n-1 indivíduos, e valida-se o modelo predizendo

o valor genético genômico do indivíduo deixado de fora da população de estimação. Esse processo é repetido n vezes (sendo n o número de indivíduos que compõem a população total), de modo que o número de validações independentes é muito alto (se a população de treinamento/validação consistir de 1.000 indivíduos, haverão 1.000 validações independentes).

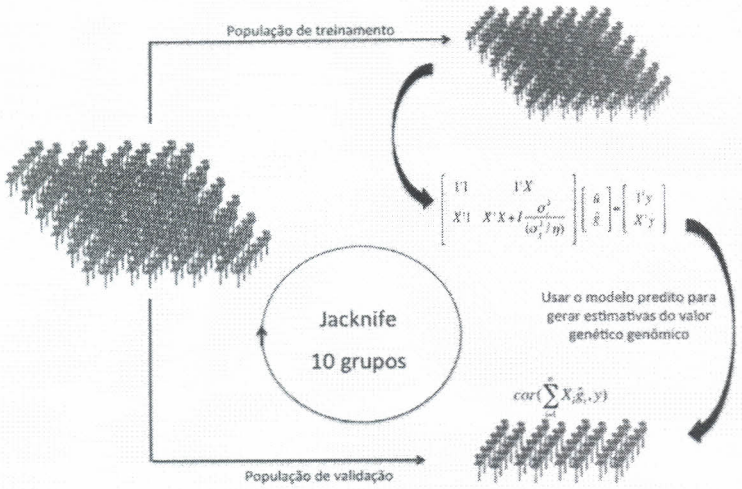


Figura 3 - Operacionalmente, a GWS faz uso de duas populações: (i) a população de treinamento e (ii) a população de validação. Essas podem fazer parte de uma única população submetida à seleção, e nesse caso há a necessidade de implementação de estratégias de validação cruzada, e.g. Jacknife.

Fatores que afetam a Seleção Genômica Ampla

A acurácia dos modelos preditivos estabelecidos com base na genotipagem e fenotipagem da população de treinamento depende de alguns fatores, a saber: (i) tamanho da população de treinamento, (ii) tamanho efetivo populacional (Ne), (iii) densidade de marcadores utilizada, (iv) herdabilidade da característica e (v) número de QTLs envolvidos no controle da(s) característica(s)-alvo (GODDARD, 2009;

GRATTAPAGLIA; RESENDE, 2011; RESENDE, 2008). É interessante salientar que os três primeiros fatores podem ser controlados pelo melhorista. O quarto fator (herdabilidade), embora possa ser mais bem estimado em função de um delineamento experimental bem planejado, não pode ser controlado. O mesmo acontece com o número de QTLs envolvidos no caráter, característica essa inerente à arquitetura genética do caráter em questão. Esses fatores são a seguir discutidos em mais detalhes, bem como o seu impacto na GWS.

No que diz respeito ao primeiro fator apontado (tamanho da população de treinamento), tem-se demonstrado via simulações que populações de treinamento de pequeno tamanho (Ex: menos de 500 indivíduos) não permitem a estimação adequada dos efeitos dos marcadores (RESENDE, 2008). Isso provavelmente se deve ao fato de que, nessas condições, a amplitude dos efeitos alélicos amostrados não é adequada, e os efeitos dos genes/alelos importantes para a correta determinação do caracter podem não estar sendo estimados com precisão (GRATTAPAGLIA; RESENDE, 2011). É interessante destacar, no entanto, que esse fator não é o principal determinante da qualidade dos modelos preditivos, uma vez que, embora exista incremento na acurácia dos modelos, à medida que o tamanho da população de treinamento aumenta, esse incremento é relativamente pequeno após o número de 1.000 indivíduos. Usando densidade elevada de marcadores (que garante que cobertura adequada do genoma), populações com menos de 1.000 indivíduos permitem atingir acurácia acima de 0,80. No caso de um número reduzido de marcadores ser utilizado, o tamanho populacional deve ser aumentado para cerca de 2.000 (GRATTAPAGLIA; RESENDE, 2011).

O tamanho efetivo da população (Ne) possui impacto muito maior que o tamanho absoluto da população de treinamento. Isso porque o tamanho efetivo determina, ao menos parcialmente, a extensão do DL na população (SVED, 1971). Em populações de maior Ne, a extensão do DL tende a ser bastante limitada e o número de alelos em dado locos, maior. Entretanto, em populações com Ne reduzido a extensão do DL é consideravelmente maior e, de maneira geral, o número de alelos presentes na população para cada locos é menor. Isso implica que população de alto Ne e maior número de marcadores deverão ser necessariamente genotipados na população

de treinamento, a fim de garantir que ao menos um deles esteja em DL com cada um dos QTLs que controlam a característica de interesse. Como o Ne é uma característica da população de melhoramento, pode-se manejar a população de treinamento, de modo a manter o Ne em níveis intermediários. Isso pode ser feito, restringindo-se o número de genitores que são intercruzados para gerar as famílias segregantes (RESENDE, 2002). Embora limitação extrema do número de genitores possa ser interessante do ponto de vista de manter um DL extenso, essa mesma redução pode ter impactos negativos significativos para o programa de melhoramento, uma vez que, ao restringir-se o Ne, reduz-se também a variabilidade genética da população, o que, consequentemente, pode diminuir os ganhos esperados em gerações futuras. Tamanhos efetivos na faixa de 10 a 50 são suficientes para o melhoramento de populações-elite por várias gerações (WHITE et al., 2007).

A densidade de marcadores é também um dos fatores que mais afetam a acurácia dos modelos preditivos. Essa dedução é fácil de ser entendida, ao relembrar que um dos fatores que afetam o desequilíbrio de ligação entre dois locos, além do Ne, é a taxa de recombinação entre dois locos (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 grande número de marcadores, espera-se encontrar um marcador mais próximo do QTL e, consequentemete, com menor taxa de recombinação. Em geral, no entanto, uma densidade de cerca de 2-3 marcadores por cM são eficientes se o Ne for mantido abaixo de 60. Resende (2008) apresenta tabelas com o DL como função do Ne e da densidade de marcadores. Desse modo, o número de marcadores necessários dependerá, em última instância, do tamanho total em cM do genoma recombinante da espécie de interesse e do Ne da população de melhoramento.

No que concerne ao impacto da herbabilidade das características sob seleção, estudos têm demonstrado que os modelos preditivos funcionam relativamente bem, mesmo para características de baixa herdabilidade. Grattapaglia e Resende (2011), por exemplo, demonstraram que a acurácia cresceu apenas 10-20% à medida que a herbabilidade aumentou de 0,2 para 0,6, independentemente do tamanho da população. Isso indica que, ao contrário da MAS, a

GWS é eficiente para selecionar indivíduos superiores, mesmo para características de baixa herdabilidade(CALUS et al., 2008). Ao considerar o fator número de QTLs envolvidos no controle genético das características-alvo da GWS, Grattapaglia e Resende (2011) verificaram que esse possui impacto significativo sobre a acurácia dos modelos preditivos apenas quando se faz uso de baixa densidade de marcadores moleculares. Isso, provavelmente, se deve ao fato de que nessa situação nem todos os QTLs estão em DL com pelo menos um marcador. Nesse caso, com baixa densidade de marcadores, a acurácia dos modelos preditivos é maior se a característica for controlada por número menor de QTLs.

Fica claro, portanto, que (i) a acurácia da GWS é altamente dependente do tamanho efetivo populacional e da densidade de marcadores utilizados, (ii) que a GWS requer populações de treinamento relativamente grandes e (iii) a herdabilidade das características e o número de QTLs que controlam essas características tem impactos menores sobre a acurácia da GWS.

4. Métodos Estatísticos na Seleção Genômica Ampla

A predição usando informações genômicas baseia-se em marcadores espalhados por todo o genoma. Assim, as metodologias estatísticas usadas para essas predições devem ser capazes de, estimar de forma acurada os efeitos dos marcadores. Assim, uma dificuldade do ponto de vista estatístico é a escolha de um método capaz de utilizar muitos marcadores em um único modelo. Na maioria desses casos, o número n de parâmetros (marcadores) utilizados no modelo é, em geral, maior que o número de observações fenotípicas, N, criando um paradoxo estatístico, em que n>N.

A análise de regressão linear multivariada assume n variáveis ***X***1, ***X***2 ,....., ***X n***, em que ***Xi*** é um vetor 1 x N. O modelo para a variável resposta y (fenótipos), em sua forma mais simples, pode ser expresso como:

y=Xg+e

em que X é uma matrix de incidência de dimensões N x n, é o vetor de coeficientes de regressão e *e* é o erro aleatório, em que é assumida distribuição normal: $e \sim N(0, I\sigma_e^2)$. Quando o vetor de coeficientes é tratado como de efeitos fixos, este pode ser estimado pela teoria clássica

de regressão como $g = (X'X)^{-1}X'y$. No entanto, para que tenha solução única, essa equação requer que o número de observações N seja maior que n. Como a teoria de seleção genômica requer grande número de marcadores, na maioria dos casos a condição $n < N$ não se aplica, de maneira que os efeitos dos marcadores não podem ser tratados como fixos no modelo.

Vários foram os métodos propostos para solucionar esse problema, dos quais se podem destacar o RR-BLUP, BayesA e Bayes B (MEUWISSEN et al., 2001), LASSO Bayesiano (DE LOS CAMPOS et al., 2009; LEGARRA et al., 2011) e Bayes Cπ (HABIER et al., 2011). De maneira geral, uma forma de classificar os diferentes tipos de modelos são aqueles que resolvem o paradoxo de n > por meio de métodos de regularização ou shrinkage, como o BLUP, e métodos que assumem um critério para seleção de covariáveis e reduzem a dimensão do problema. Além disso, várias são as pressuposições de cada método, de maneira que na próxima seção esses métodos são discutidos em maiores detalhes. Nesta secção, não serão discutidas as metodologias de regressão *stepwise* por mínimos quadrados, uma vez que os efeitos dos marcadores não são estimados simultâneamente, o que reduz a acurácia do modelo. Maiores informações sobre essa metodologia podem ser encontrados em Meuwissen et al. (2001) e Moser et al. (2009).

4.1. RR-BLUP

A abordagem mais simples para modelar o efeito dos marcadores como aleatório é o uso do BLUP (*Best Linear UnbiasedPrediction*), pormeio de uma regressão ridge ou aleatória (RR-BLUP). Característica importante que diferencia RR-BLUP dos outros métodos é o fato de que o este assume que os efeitos dos marcadores apresentam distribuição normal com variância constante. Essa pressuposição é, na verdade, equivalente ao modelo infinitesimal proposto por Fisher. Quando muitos QTLs controlam a característica de interesse e nenhum deles é de grande efeito, o RR-BLUP torna-se boa alternativa. Assim como o modelo de regressão linear multivariada exposto na seção anterior, a forma mais simples de modelar os efeitos dos marcadores através de RR-BLUP é pelo modelo:

$$y = Xg + e$$

em que $g \sim N(0, I\sigma_g^2)$..

Como os efeitos dos marcadores são assumidos como aleatórios, eles podem ser estimados resolvendo a equação:

$$g = (X'R^{-1}X + I\sigma_g^2)^{-1}X'R^{-1}y$$

em que R é uma matrix diagonal que pode conter pesos relativos (variâncias residuais específicas) associados às acurácias dos valores fenotípicos desregressados (y) utilizados na predição. Em geral, quando essa informação não é disponível, a matriz diagonal R é tida como $R = I\sigma_e^2$, e a equação pode ser simplificada: $g = (X'X + I\lambda)^{-1}X'y$, em que $\lambda = k = \sigma_e^2 / \sigma_g^2$ é constante para todos os marcadores.

A variância dos marcadores é constante, e em geral *k* deve ser considerada como função da variância genética aditiva estimada anteriormente por métodos tradicionais de genética quantitativa (e.g. REML – *RestrictedMaximumLikelihood*). Em casos de uso da matriz de marcadores X padronizada, deve-se dividir a variância aditiva pelo número total de marcadores usados no modelo (MEUWISSEN et al., 2009). Caso contrário, essa fração é corrigida pelo somatório da heterozigozidade dos locos individuais i, conforme demonstrado por Habier et al. (2008). Assim, deve ser assumido como ou $\sigma_g^2 = \sigma_a^2 / [2\sum_{i=1}^{n} p_i(1-p_i)]$ para X padronizada e para X não padronizada, respectivamente, em que σ_a^2 é a variância genética aditiva do caráter. Alternativamente, pode ser assumido como desconhecido e o valor de *k*, ser definido iterativamente, através de amostragem de Gibbs (*Gibbs sampling*). Esse procedimento é chamado de Regressão Ridge Bayesiana (PÉREZ et al., 2010).

Outros modelos mais complexos podem ser considerados ainda no contexto do RR-BLUP. Assim, dependendo do objetivo do usuário, pode-se ajustar um modelo que estima os efeitos devido à dominância de cada marcador. De maneira similar, caso o número de marcadores utilizados seja baixo, alguns QTLs podem não estar em desequilíbrio de ligação com nenhum marcador. Nesse caso, é recomendado que se incorpore um efeito aleatório poligênico, que contém a fração da variância genética não capturada pelos marcadores. Em geral, todas

as análises assumem também efeito fixo, conforme demonstrado no exemplo a seguir. Em geral, esse vetor de efeitos fixos contém apenas a média geral, embora possa conter mais efeitos.

Exemplo:

Para exemplificar a análise pelo método RR-BLUP, considerou-se a genotipagem e fenotipagem de seis indivíduos e 10 marcadores, conforme exposto a seguir. Neste exemplo, nenhuma análise de validação será aplicada, embora esta seria recomendada como demonstrado anteriormente, neste capítulo.

$$Z = \begin{bmatrix} 2 & 2 & 1 & 0 & 1 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 & 2 & 2 & 1 & 2 & 0 & 2 \\ 2 & 2 & 1 & 1 & 2 & 0 & 1 & 1 & 2 & 1 \\ 0 & 1 & 0 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 2 & 1 & 2 & 0 & 1 & 2 & 0 & 2 & 2 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 & 2 & 1 & 0 & 1 \end{bmatrix} \text{ e}$$

$$y = \begin{bmatrix} 5{,}23 \\ 5{,}12 \\ 4{,}64 \\ 5{,}02 \\ 4{,}91 \\ 4{,}88 \end{bmatrix}$$

O modelo linear misto equivale a:

$$y = 1u + Zg + e$$

em que y é um vetor coluna de fenótipos, 1 é um vetor coluna que contém o número 1, N (número de indivíduos) vezes, X é a matriz de incidência que aloca o genótipo de cada loco marcador a cada indivíduo, u é um escalar contendo o efeito fixo da média geral, g é um vetor contendo os efeitos de um dos alelos de cada loco marcador e e é um vetor de erros aleatórios.

As equações de modelo misto equivalem a:

$$\begin{bmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + I\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1'y \\ Z'y \end{bmatrix}$$

Considerando $\lambda = \frac{\sigma_e^2}{\sigma_g^2} = 1$ e resolvendo a equação anterior, obtêm-se os valores da média estimada em 5,32, e o restante do vetor solução contém os efeitos estimados das marcas:

$$\hat{g} = \begin{bmatrix} -0{,}04 \\ -0{,}002 \\ -0{,}04 \\ -0{,}07 \\ -0{,}07 \\ 0{,}02 \\ -0{,}05 \\ 0{,}02 \\ -0{,}11 \\ 0{,}007 \end{bmatrix}$$

O método RR-BLUP é equivalente à substituição da matriz de parentesco (matriz A) pela matriz de parentesco genômico (matriz G) nas equações de modelos mistos (BLUP tradicional, convencionalmente usado em análises quantitativas) (HABIER et al., 2008). A diferença dos dois métodos é que, no RR-BLUP, estima-se o efeito individual de cada marcador, que é subsequentemente utilizado em conjunto para gerar o valor genético genômico. No caso do uso de G-BLUP, o termo predito é diretamente o valor genético genômico, via parentesco obtido pelos dados genômicos.

4.2. Bayes A

Outro método proposto por Meuwissen et al. (2001) é denominado Bayes A. Sob determinadas distribuições *a priori*, este método equivale ao método BLUP com variâncias genéticas heterogêneas entre locos, pois as variâncias dos segmentos cromossômicos diferem em cada segmento e são estimadas sob esse modelo, considerando-se a informação combinada dos dados e da distribuição *a priori* para essas variâncias. Nesse caso, o modelo é ajustado por meio de uma abordagem bayesiana com estrutura hierárquica em dois níveis. Os efeitos dos marcadores são assumidos como amostras de uma distribuição normal com média zero e variância de cada marcador dada por uma distribuição qui-quadrada inversa e escalonada conforme a seguir:

$$g_i|\sigma_g^2 \sim N(0,\sigma_g^2)$$

$$\sigma_g^2 \sim \chi^{-2}(\nu_g, S_g^2)$$

em que ν_g é o número de graus de liberdades, e S_g^2 é o parâmetro de escala da distribuição. Assim, tem-se que a distribuição *a priori* dos efeitos genéticos dos marcadores, $g_i|\nu_g, S_g^2$, tem distribuição t de Student univariada, ou seja, $g_i|\nu_g, S_g^2 \sim t(0,\nu_g, S_g^2)$. Assim, essa formulação resulta na modelagem dos efeitos dos marcadores como amostras de uma distribuição t de Student.

O valor de S_g^2 pode ser derivado com base no valor esperado de uma variável aleatória com distribuição qui-quadrado invertida escalonada (HABIER et al., 2011). Essa esperança matemática é dada por $E(\sigma^2) = \frac{S^2\nu}{\nu-2}$. Assim, o parâmetro de escala é dado por $S^2 = \frac{E(\sigma^2)(\nu-2)}{\nu}$. Então, para os efeitos genéticos dos marcadores, tem-se $E(\sigma_g^2) = \frac{S_g^2\nu_g}{\nu_g-2}$ e $S_g^2 = \frac{E(\sigma_g^2)(\nu_g-2)}{\nu_g}$. A esperança $E(\sigma_g^2)$ equivale $E(\sigma_g^2) = \frac{\sigma_a^2}{\sum_{i=1}^{n} 2p_i(1-p_i)}$. Assim, $S_g^2 = \frac{\sigma_a^2}{\sum_{i=1}^{n} 2p_i(1-p_i)} \frac{(\nu_g-2)}{\nu_g}$, em que ν_g = 4.2, conforme Meuwissen et al. (2001), σ_a^2 é a variância genética aditiva do caráter, e p_i é a frequência alélica do marcador i.

Para os efeitos residuais, tem-se $E(\sigma_e^2) = \frac{S_e^2\nu_e}{\nu_e-2}$ e $S_e^2 = \frac{E(\sigma_e^2)(\nu_e-2)}{\nu_e}$. A esperança $E(\sigma_e^2)$ equivale a $E(\sigma_e^2) = \tilde{\sigma}_e^2$. Assim, $S_e^2 = \tilde{\sigma}_e^2 \frac{(\nu_e-2)}{\nu_e} = \tilde{\sigma}_e^2 \frac{(4{,}2-2)}{4{,}2}$, em que $\tilde{\sigma}_e^2$ é um valor *a priori* de σ_e^2.

Para obtenção da informação combinada da distribuição *a priori* e da verossimilhança dos dados, ou seja, para obtenção da distribuição *a posteriori* dos efeitos genéticos dos marcadores, adota-se o procedimento de simulação estocástica (método Monte Carlo cadeias de Markov – MCMC), denominado amostragem de Gibbs.

Em termos mais simples, o algoritmo da amostragem de Gibbs pode ser apresentado de forma resumida, conforme Resende (2008) e Meuwissen et al. (2001):

1. Fornecer os valores iniciais dos parâmetros de locação e dispersão do modelo. Esses valores iniciais podem ser calculados através de procedimentos-padrão, como a estimação de componentes de variância por REML ou quadrados mínimos. Considerando a média geral $\bar{y}$ como único efeito fixo, pode-se calcular $\bar{y}$ como a média aritmética das observações. O vetor dos efeitos de marcadores devem ser inicializados com um número positivo de pequena magnitude.
2. Atualizar σ^2_{gi} para o i-ésimo marcador, amostrando-o da distribuição condicional completa $P(\sigma^2_{gi}|g_i) = \chi^{-2}(\upsilon_g + n), S^2_g + g_i' g_i$ com $\upsilon_g = 4.2$ e S^2_g calculado conforme a expressão anterior.
3. Dados g_i e $\bar{y}$, calcular os valores de e via $e = (y - 1\bar{y} - Xg)$, em que X = [X_1 X_2 X_3] 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. Amostrar, de uma distribuição normal com média (1/N) $(y - Xg)$ e variância σ^2_e / N, a média geral dado a variância residual.
5. Amostrar, de uma distribuição com média $\frac{X_{ij}' y - X_{ij}' X g_{ij=0} - X_{ij}' 1_n u}{X_{ij}' X_{ij} + \sigma^2_e / \sigma^2_{gi}}$ e variância $\sigma^2_e / (X_{ij}' X_{ij} + \sigma^2_e / \sigma^2_{gi})$, todos os efeitos de marcadores g_{ij} dados a amostragem mais recente da média, σ^2_e e σ^2_{gi}, em que X_{ij} é o vetor coluna de X com efeitos g_{ij}. No caso, $g_{ij=0}$ equivale a g com efeito g_{ij} igualado a zero.

6. Repetir os passos de (2) a (5) até que se obtenha a convergência da cadeia.

4.3. Bayes B

Durante a modelagem dos efeitos dos marcadores, uma pressuposição possível, principalmente quando se usa número extremamente elevado de marcadores, é a de que muitos marcadores estão em regiões genômicas que não contêm nenhum QTL, tendo, assim, efeito igual a zero. Baseado nessa pressuposição, Meuwissen et al. (2001) propuseram o método Bayes B. Essa abordagem assume que um número de marcadores (com proporção π) tem efeito zero, e o restante dos marcadores, com proporção 1-π, é amostrado com uma variância individual para cada marcador, considerando as mesmas *prioris* usadas no método Bayes A. Assim, uma estrutura hierárquica semelhante ao Bayes A pode ser descrita:

$$g_i|\sigma_{g_i}^2 \sim N(0, \sigma_{g_i}^2)$$

$$\sigma_{g_i}^2 = 0 \; com \; probabilidade \; \pi$$

$$\sigma_{gi}^2 \sim \chi^{-2}(\nu_g, S_g^2) \text{ com probabilidade } 1 - \pi.$$

Um dos problemas desse método é a dependência de uma definição do valor de π. Assim, para que a análise seja eficiente, esse método requer um conhecimento *a priori* sobre a característica analisada, para que a predeterminação de π seja coerente com a arquitetura genética da característica fenotípica. Caso um valor inconsistente para π seja escolhido, isso refletirá negativamente nas acurácias das predições dos valores genéticos genômicos. A alternativa para esse problema foi chamada de Bayes Cπ e apresentada por Habier et al. (2011). Esses autores propuseram pequena modificação no método, de maneira que o valor de π fosse iterativamente amostrado, sob distribuição *a priori* uniforme (π ~ uniforme(0,1)). (Uma vez que o MCMC converge, o parâmetro é definido como a média da distribuição *a posteriori*, e a análise roda mais uma vez para estimar o efeito dos marcadores. De maneira geral, π é definido para refletir a proporção esperada de marcadores em desequilíbrio de ligação com o QTL relativa ao número total de marcadores.

4.4. LASSO Bayesiano

Conforme mencionado anteriormente, as duas abordagens bayesianas propostas por Meuwissen et al. (2001) assumem efeitos de marcadores amostrados de uma distribuição t de Student. Outra possibilidade que reflete a distribuição de efeitos dos marcadores é a distribuição exponencial dupla. Esta tem caudas mais longas, quando comparadas com a distribuição t, no entanto contém número maior de efeitos pequenos (diferentes de zero) (Figura 4). A abordagem conhecida como LASSO (*Least absolute shrinkage and selection operator*) Bayesiano (DE LOS CAMPOS et al., 2009; LEGARRA et al., 2011) usa essa distribuição para modelar os efeitos dos QTLs, formulada num contexto bayesiano. As estimativas LASSO podem ser derivadas, a exemplo da moda, da distribuição *posteriori* bayesiana ao considerar a distribuição a *priori* como uma exponencial dupla independente (TIBSHIRANI, 1996). De maneira equivalente aos modelos anteriores, o BLASSO pode ser implementado num contexto hierárquico, em que os marcadores são amostrados de uma distribuição normal com variância amostrada de uma distribuição exponencial dupla.

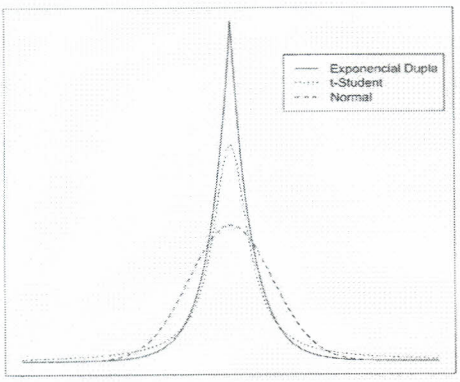


Figura 4 - Função densidade de probabilidade da distribuição exponencial dupla quando comparada com a distribuição normal e t de Student.

Uma vez que se obtêm estimativas dos efeitos de cada marcador, é possível, como demonstrado anteriormente, estimar o Valor Genético Genômico de cada indivíduo genotipado na população. Caso o valor genético paramétrico (real) dos indivíduos estivesse disponível, a acurácia de predição dos modelos de GWS poderia ser obtida pela simples correlação de Spearman entre os VGG e o valor genético real. Assim, caso os valores genéticos estimados através de seleção fenotípica (VGF) tenham alta acurácia (acima de 0,95), é possível assumir que esses são uma aproximação dos valores genéticos paramétricos e estimar a acurácia do modelo pela correlação entre VGG e VGF. Como em grande parte das vezes isso não acontece, o vetor y deve consistir de valores fenotípicos desregressados (VFD) (GARRICK et al., 2009; RESENDE et al., 2010), e, assim, a correlação entre VGG e VFD mede a capacidade preditiva de fenótipos. A acurácia de predição dos valores genéticos pode ser obtida ao dividir a correlação entre VGG e VFD pela raiz quadrada da herdabilidade da característica analisada (LEGARRA et al., 2011; RESENDE, 2008).

5. Estimação dos efeitos genéticos genômicos na população de seleção

Assim, como visto nos tópicos anteriores, a geração de um modelo para aplicação da Seleção Genômica Ampla depende de duas informações: dados fenotípicos e genotípicos. Essas duas informações são usadas em conjunto para atingir o objetivo final da GWS, que é a geração de um modelo de predição capaz de usar apenas informações genotípicas para predizer fenótipos futuros. As variáveis preditoras são o conjunto de marcadores, o que requer a estimação, em um etapa inicial, da contribuição (efeito) de cada marcador em explicar o fenótipo. Essas estimativas são utilizadas no modelo de predição e, em conjunto, compõem o Valor Genético Genômico (VGG) do indivíduo. Assim, uma vez que as estimativas dos efeitos de marcadores estão disponíveis, o VGG é predito da seguinte maneira:

$$VGG = \sum_{i}^{n} X_i \hat{g}_i$$

em que n é o número de marcadores dispostos no genoma, é a linha da matriz de incidência que aloca o genótipo do i-ésimo marcador para cada indivíduo e é o efeito estimado do i-ésimo marcador.

Exemplo:

Como exemplo da estimação do valor genético em uma pequena e hipotética "população" (seis indivíduos), considere a estimativa de efeito de 10 marcadores (g_1, g_2, ..., g_{10}). As únicas informações disponíveis são os genótipos de cada indivíduo (X) e o vetor de efeitos anteriormente estimados para cada marcador:

$$X = \begin{bmatrix} 2 & 21 & 01 & 00 & 20 & 0 \\ 1 & 02 & 02 & 21 & 20 & 2 \\ 2 & 21 & 12 & 01 & 12 & 1 \\ 0 & 10 & 20 & 12 & 01 & 2 \\ 2 & 12 & 01 & 20 & 22 & 0 \\ 1 & 11 & 21 & 02 & 10 & 1 \end{bmatrix} \text{ e } \hat{g} = \begin{bmatrix} 0.23 \\ 0.12 \\ -0.36 \\ 0.02 \\ -0.09 \\ -0.12 \\ 0.34 \\ 0.29 \\ -0.19 \\ -0.13 \end{bmatrix}$$

Valores negativos dos efeitos dos marcadores indicam a contribuição para o decréscimo do fenótipo e, de maneira semelhante, valores positivos indicam efeito positivo que leva ao acréscimo do fenótipo. O Valor Genético Genômico dessa "população" é, então, predito como:

$$VGG = \sum_{i}^{n} X_i \hat{g}_i = \begin{bmatrix} 0.83 \\ -0.25 \\ 0.30 \\ 0.27 \\ -0.27 \\ 0.78 \end{bmatrix}$$

Caso um valor genético positivo fosse o atributo de interesse do melhorista, os melhores indivíduos seriam o indivíduo número 1 e o indivudo número 6. Entretanto, se esse fenótipo de interesse fosse negativo, os dois melhores indivíduos para dar continuidade ao melhoramento seriam os indivíduos 2 e 5. Um exemplo deste último

caso seria a suscetibilidade a doenças, em que os indivíduos que são **menos** suscetíveis são preferidos para o melhoramento genético.

O VGG de um indivíduo é uma estimativa do valor genético real, ou o valor paramétrico da contribuição genética daquele indivíduo para determinada característica. Dessa forma, um fator importante a ser determinado é a acurácia de predição dos valores genéticos por cada método. Existem basicamente duas formas de calcular essa acurácia para a seleção genômica ampla, conforme detalhado adiante. A importância prática da estimação da acurácia de predição reside no cálculo do ganho de seleção, principal medida calculada pelo melhorista em um programa de melhoramento. O ganho genético pode ser calculado pela equação (FALCONER, 1989):

$$\Delta G = \frac{Intensidade\ de\ seleção\ x\ \boldsymbol{Acurácia\ da\ seleção}\ x\ desvio-padrão\ genético}{\boldsymbol{Duração\ do\ ciclo\ de\ melhoramento}}$$

Dessa forma, fica claro que a seleção genômica tem influência em dois termos dessa equação quando comparados com a seleção fenotípica tradicional. Um modelo preditor que usa informações genéticas tende a ser mais acurado na predição do mérito genético de cada indivíduo. Além disso, para algumas espécies em que o ciclo de melhoramento é longo, como o caso de plantas perenes e animais, a seleção ultraprecoce via marcadores pode reduzir drasticamente a duração do ciclo de melhoramento, o que aumenta o ganho de seleção por unidade de tempo. Resende et al. (2012b) e Resende et al. (2012c) avaliaram esse impacto no melhoramento de *Pinus taeda*. Programas de melhoramento convencional de árvores do gênero *Pinus* podem ter um ciclo de melhoramento de até 20 anos. Esses autores utilizaram 4.825 marcadores do tipo SNP para gerar modelos de predição de valores genéticos com uma acurácia que variou entre 0,65 e 0,75, dependendo da característica. Para comparar a eficiência da seleção genômica com a seleção fenotípica tradicional na mesma população, a fórmula citada anteriormente de ganho genético pode ser reduzida para:

$$\Delta G = \frac{\boldsymbol{Acurácia\ da\ seleção}}{\boldsymbol{Duração\ do\ ciclo\ de\ melhoramento}}$$

uma vez que os outros dois parâmetros, intensidade de seleção

e variância genética, estão relacionados com as decisões do melhorista e com as características da população, respectivamente, independentemente da escolha do método de seleção. Os respectivos autores analisaram o impacto da GWS considerando essas acurácias e uma redução pela metade no ciclo de melhoramento. Ao comparar essas acurácias com as obtidas na mesma população caso a seleção fenotípica fosse realizada, observaram eficiência da GWS estimada como 53-120% superior à seleção fenotípica, dependendo da característica.

Exemplo

Simulação de uma população em desequilíbrio de ligação (DL)

Para ilustrar o uso da seleção genômica ampla na predição de valores genéticos, foi considerada uma população F_1, derivada do cruzamento entre duas populações genitoras P_1 e P_2, todas obtidas por simulação. Inicialmente, foram simulados 100 indivíduos de cada população que foram genotipados em relação a 50 locos que expressam dois alelos codominantes em cada loco. Apesar dos números reduzidos de indivíduos e de locos estudados, o exemplo ilustra bem a aplicabilidade da seleção genômica ampla para fins de melhoramento genético, sem perda da possibilidade de generalização para outras situações em que uma dimensionalidade maior seria recomendável e, provavelmente, utilizada.

Os dados de simulação foram obtidos por meio do aplicativo GENES, e os resultados preliminares do *status* população em relação à condição de equilíbrio de Hardy-Weinberg e desequilíbrio de ligação são apresentados na Tabela 1. O número de locos em equilíbrio foi avaliado por meio da estatística qui-quadrado, confrontando os resultados observados com os esperados iguais a p^2, 2pq e q^2 para AA, Aa e aa, respectivamente. Para o cálculo do desequilíbrio de ligação, consideram-se dois locos, com dois alelos cada, ou seja, A e a, e B e b, com frequências alélicas p_A,q_a, p_B e q_b, respectivamente, resultando nas frequências gaméticas π_{AB}, π_{Ab}, π_{aB} e π_{ab} para cada possibilidade. O componente básico para o cálculo de desequilíbrio é a diferença entre a frequência esperada e a observada dos gametas, dada por:

$$\Delta = (\pi_{AB} - p_A p_B)$$

Uma medida do desequilíbrio, denotada r, é fornecida por:

$$r = \frac{\Delta}{p_A q_a p_B q_b}$$

É conveniente considerar r^2 como o quadrado do coeficiente de correlação entre dois locos. Entretanto, ao menos que os dois locos tenham frequências alélicas idênticas, o valor da correlação igual a 1 não é possível de ser obtida.

Uma medida de desequilíbrio alternativa é a estatística D', calculada conforme descrito a seguir:

$$|D'| = \frac{\Delta}{\min(p_A p_B, q_a q_b)} \quad \textbf{para } \Delta < 0$$

$$|D'| = \frac{\Delta}{\min(p_A q_b, q_a p_B)} \quad \textbf{para } \Delta > 0$$

O valor de D' é obtido a partir das frequências alélicas observadas e irá variar entre 0 e 1, se as frequências alélicas diferirem entre os locos. D' poderá ser menor do que 1 apenas se todos os quatro possíveis gametas forem observados, assumindo, consequentemente, que eventos de recombinação ocorreram entre os locos.

As estatísticas r e D' refletem diferentes aspectos do desequilíbrio de ligação, e comportam-se diferentemente sob condições variadas. No exemplo considerado, todas as estatísticas estimadas para os genitores foram de pequena magnitude, indicando que os pares de locos na população não se encontram em desequilíbrio. Também, constata-se que as populações genitores encontram-se, como esperado no processo de simulação, em equilíbrio de Hardy-Weinberg. Em alguns poucos locos, a hipótese de equilíbrio foi rejeitada, podendo ser atribuído a erros de amostragem e tamanho reduzido da população, impossibilitando a análise acurada.

A geração F1 manifesta considerável desequilíbrio em consequência de não ser resultante de acasalamento ao acaso, mas derivada da hibridação entre P_1 e P_2. Os valores de r e D' obtidos para essa geração estão representados graficamente a seguir e permitem

visualizar as situações em que se verifica maior quantidade de pares de locos em desequilíbrio de ligação.

Tabela 1 - Número de locos em equilíbrio de Hardy-Weinberg e estimativas do desequilíbrio de ligação em duas populações genitoras P_1 e P_2 e a geração F_1

Pop.	nº locos EHW	D'											Pares de
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Locos
P_1	2	405	270	144	89	73	41	27	15	30	117	14	1225
P_2	3	437	280	133	98	78	44	32	29	25	65	4	1225
F1	21	377	311	176	137	78	47	43	24	13	19	0	1225
		r											
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
P_1	2	818	326	60	14	5	2	0	0	0	0	0	1225
P_2	3	807	316	7	17	7	1	0	0	0	0	0	1225
F1	21	589	361	133	86	33	16	7	0	0	0	0	1225

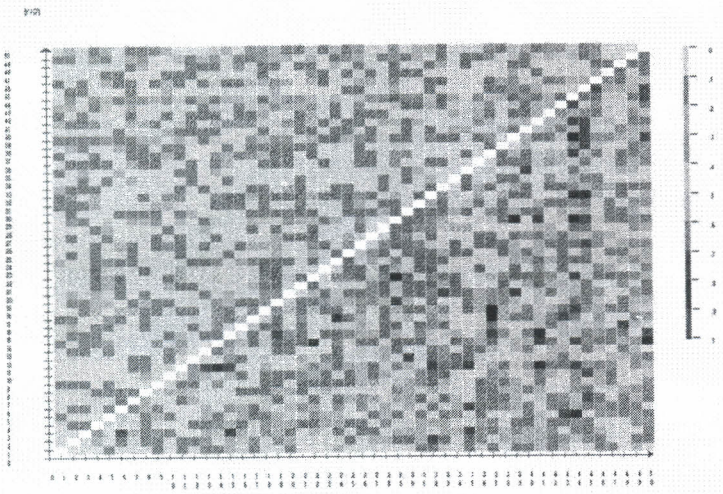


Gráfico 1 - Valores de r e D' obtidos para esta geração F_1.

Simulação de características quantitativas

Foi simulado um caráter quantitativo controlado por 20 dos 50 locos estudados, considerando ação aditiva entre os genes. O valor genético total expresso por determinado indivíduo pertencente à população F_1 foi estimado a partir da expressão:

$$G_i = \mu + a_i + d_i$$

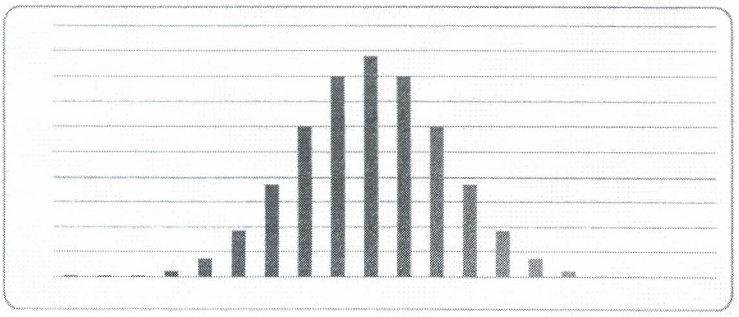
em que:

$$a_i = \sum_{j=1}^{g} p_j \alpha_j$$

$$d_i = 0$$

sendo α_j o efeito do alelo favorável no loco j, considerado igual a 1, 0 ou -1 para as classes genotípicas AA, Aa e aa, respectivamente, e p_j é a contribuição do loco j para a manifestação da característica considerada,

neste exemplo, como tendo distribuição binomial. Para facilidade de interpretação no exemplo em consideração, foi estabelecido que os 20 primeiros locos genotipados são os controladores da característica, e as suas importâncias relativas com a magnitude podem ser ilustradas a seguir.



O valor fenotípico do indivíduo i foi estabelecido a partir do efeito genotípico (G_i) e do efeito ambiental (E_i), de forma que se tenha:

$$F_i + G_i + E_i$$

Os efeitos ambientais foram gerados segundo a distribuição normal com média zero e variância compatível com uma herdabilidade individual, que neste exemplo foi estabelecida como igual a 60%.

Os valores genotípicos e fenotípicos de cada indivíduo estão apresentados na Tabela 2. Verifica-se que a correlação entre esses valores foi igual a 0,768371 e seu quadrado (r^2) igual a 0,5904, cujo valor é próximo ao da herdabilidade estabelecida na simulação.

Predição de valores genéticos EGBV

Para fins de ilustração, são consideradas três estratégias de seleção. A primeira refere-se à seleção com base nos valores fenotípicos dos indivíduos avaliados, cuja acurácia do processo expressa pela herdabilidade do caráter é estimada em 60%, significando certa dificuldade de selecionar genótipos de fato superiores ou de descartar aqueles de desempenho genético não favorável. A segunda possibilidade é ideal, mas não possível de ser praticada em situações

reais, referente à seleção direta sobre o valor genético simulado e desprovido da influência ambiental. E, por fim, a seleção praticada sobre valores genéticos preditos, que levam em consideração os valores fenotípicos, mas agregam consideráveis informações a partir da genotipagem realizada.

Na predição dos valores genéticos foram utilizadas as abordagens RR-BLUP, G-BLUP e LASSO, fornecendo os resultados demonstrados na Tabela 3.

Tabela 2 - Valores fenotípicos (Vf), genotípicos (Vg) e preditos pelas técnicas RR-Blup, G-Blup e Lasso numa população F_1

Ind	Vf	Vg	rr/GBlup	Lasso	Ind	Vf	Vg	rr/GBlup	Lasso
1	99.804	99.730	-0.1316	99.793	51	99.979	100.343	0.0555	99.988
2	99.892	99.771	-0.1002	99.798	52	99.940	100.117	-0.2003	99.697
3	100.015	100.074	0.1655	100.080	53	99.429	99.859	-0.3857	99.498
4	100.052	99.905	0.0115	99.956	54	100.117	99.522	0.2361	100.179
5	99.912	99.960	0.0307	99.937	55	99.644	100.105	-0.2662	99.643
6	99.674	99.736	-0.3224	99.573	56	99.509	99.594	-0.3636	99.510
7	99.998	100.119	0.0115	99.940	57	99.906	99.517	-0.1310	99.796
8	100.081	99.995	0.0658	100.009	58	99.783	99.864	-0.0331	99.883
9	100.142	100.060	0.0673	100.043	59	99.944	99.885	0.0082	99.934
10	99.787	99.649	-0.1670	99.740	60	100.172	99.846	0.2165	100.190
11	100.018	99.990	0.0789	100.003	61	100.180	100.064	0.1367	100.059
12	99.560	99.739	-0.2236	99.683	62	100.295	100.179	0.2490	100.201
13	100.158	100.019	0.1888	100.118	63	99.708	100.047	-0.1263	99.779
14	100.230	99.773	0.0506	99.946	64	99.964	99.781	0.1028	100.027
15	100.143	99.754	-0.0274	99.872	65	99.690	99.969	-0.3321	99.577
16	100.283	100.248	0.1449	100.118	66	99.564	99.660	-0.2509	99.644
17	99.840	99.708	-0.1103	99.803	67	99.962	99.697	-0.0233	99.911
18	99.901	100.075	-0.0390	99.909	68	99.929	99.913	-0.0340	99.859
19	100.257	100.034	0.1729	100.124	69	99.992	99.757	0.1507	100.096
20	99.852	99.834	0.0076	99.927	70	99.584	99.994	-0.3874	99.506
21	100.337	100.383	0.4264	100.401	71	100.154	99.584	0.1295	100.032
22	100.171	100.028	0.0692	100.005	72	100.154	99.900	0.1772	100.115

23	99.548	99.549	-0.3033	99.584	73	99.620	100.060	-0.2523	99.659
24	100.233	100.112	0.2322	100.204	74	99.212	99.578	-0.3779	99.490
25	99.728	99.738	-0.3187	99.616	75	99.905	99.643	0.0857	100.000
26	99.979	100.060	0.0350	99.959	76	100.001	99.723	-0.0553	99.872
27	99.741	99.850	-0.1078	99.826	77	99.546	99.802	-0.1455	99.748
28	99.914	100.055	0.0628	99.979	78	100.138	99.760	0.0806	100.038
29	99.742	99.579	-0.2847	99.609	79	99.459	100.004	-0.2526	99.641
30	100.085	100.069	0.1187	100.079	80	99.665	99.662	-0.2781	99.630
31	99.616	99.567	-0.3593	99.553	81	99.650	99.808	-0.2047	99.675
32	99.730	99.946	-0.0710	99.873	82	100.342	99.629	0.1908	100.144
33	100.074	99.753	0.0056	99.915	83	99.589	100.103	-0.1674	99.764
34	99.777	99.599	-0.1808	99.744	84	100.178	99.965	0.2042	100.162
35	100.532	100.322	0.3476	100.310	85	99.516	100.115	-0.1430	99.754
36	99.772	99.728	-0.2078	99.742	86	99.701	99.635	-0.2512	99.653
37	99.963	100.052	0.0485	99.984	87	99.731	99.654	-0.0906	99.830
38	100.088	100.156	0.2255	100.180	88	99.579	99.654	-0.1770	99.738
39	99.369	99.568	-0.3871	99.495	89	99.474	99.773	-0.2182	99.687
40	100.267	100.462	0.3723	100.353	90	99.846	99.693	-0.0279	99.852
41	99.576	100.022	-0.0560	99.869	91	99.829	99.874	-0.0807	99.850
42	99.593	100.067	-0.0757	99.837	92	99.687	99.821	-0.2336	99.686
43	99.851	99.920	-0.0506	99.859	93	99.935	99.661	-0.0915	99.830
44	100.139	99.999	-0.0084	99.921	94	99.652	99.787	0.0122	99.952
45	100.310	100.114	0.2125	100.162	95	99.294	99.928	-0.5264	99.339
46	99.679	99.863	-0.1678	99.756	96	99.886	99.401	0.0558	99.986
47	99.748	99.925	-0.0173	99.923	97	100.032	99.830	0.1370	100.067
48	99.926	99.986	-0.0779	99.862	98	99.897	99.997	0.0228	99.977
49	99.559	100.086	-0.1372	99.788	99	99.799	100.000	-0.0608	99.861
50	100.491	99.730	0.3502	100.339	100	100.132	99.737	0.1369	100.061

Nessa ilustração deve ser considerado que os valores genotípicos são paramétricos e, portanto, conhecidos apenas por tratar-se de ilustração em que se consideraram dados provenientes de simulação em que todas as informações genéticas eram previamente conhecidas. Para fins de melhoramento genético, o pesquisador teria

à sua disposição, em condições de campo, apenas as informações fenotípicas resultantes de sua mensuração que é afetada, em maior ou menor intensidade, pelas causas não genéticas. Neste exemplo, o quadrado da correlação entre os valores fenotípicos e genotípicos foi próximo de 60% (herdabilidade da característica), indicando que o ambiente seria um agente perturbador do processo seletivo de considerável magnitude, reduzindo os ganhos e permitindo que genótipos não tão superiores, favorecidos pelo ambiente, contribuíssem para a próxima geração e que genótipos superiores, com desempenho prejudicado pelo ambiente, fossem descartados. O aumento da acurácia, ou seja, na capacidade de o pesquisador inferir sobre o valor genético do indivíduo a partir de valores mensurados, passa a ser indispensável, maximizando os ganhos e reduzindo o custo, tempo e mão de obra despendidos na experimentação. Pode-se verificar, por meio dessa ilustração, que a inclusão das informações moleculares associadas ao uso de técnicas biométricas apropriadas propiciou informações de grande valia aos pesquisador.Na Tabela 3, pode-se verificar que o uso do valor genômico predito é critério de seleção mais eficiente, pois proporciona acurácia acima do que seria obtido com o uso dos valores fenotípicos. A predição a partir da abordagem do Lasso mostrou-se ligeiramente superior em razão, provavelmente, da ação diferencial dos locos sobre a característica quantitativa simulada.

Tabela 3 - Estimativas do quadrado da correlação entre valores fenotípicos (Vf), genotípicos (Vg) e genômicos (Vgen) preditos

Procedimentos	r^2(Vf,Vg)	r^2(Vgen,Vg)
RR-BLUP	0.5904	0.767338
G-BLUP	0.5904	0.767338
LASSO	0.5904	0.788665

Efeito gênico

Os efeitos de cada loco estudado sobre a característica podem ser avaliados por meio das abordagens de GWS utilizadas. Deve ser ressaltado que, neste exemplo, foram simulados 50 locos e estabelecido que apenas 20 deles tinham importância direta sobre as características estudadas, com efeito previamente conhecido. A

análise genômica, como pode ser verificado na Figura 5, identificou os locos mais importantes como aqueles eleitos como os diretamente envolvidos no controle gênico do caráter. Também, verifica-se, nessa figura, distribuição observada similar à distribuição binomial estabelecida no processo de simulação.

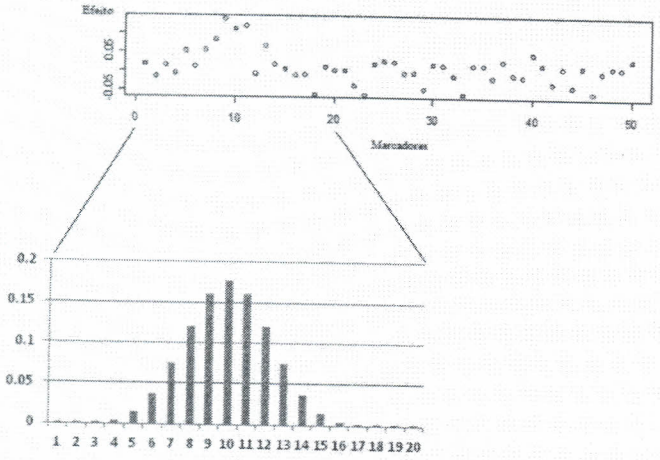


Figura 5 - Estimativa do efeito de cada marcador utilizado.

6. Aplicação da Seleção Genômica Ampla em Diferentes Organismos

A aplicabilidade da seleção genômica no melhoramento de animais e plantas tem sido demonstrada por meio de uma série de trabalhos de simulação e, mais recentemente, de prova de conceito. Nesta seção foram abordados, por limitação de espaço, apenas alguns desses trabalhos, inicialmente relacionados à área animal e, posteriormente, àqueles ligados ao melhoramento de plantas.

6.1. Aplicações da Seleção Genômica no Melhoramento Animal

Schaeffer (2006), primeiramente, demonstrou que a GWS poderia ser aplicada ao melhoramento de gado leiteiro de modo economicamente viável. Segundo dados desse autor, para testar 500 touros ao ano, por meio do esquema tradicional de testes de progênies, há um gasto de 25 milhões de dólares, e o tempo total do processo é de 64 meses da concepção à prova. Se apenas 20 dos 500 touros retornarem ao serviço (e.g. forem selecionados e, por isso, não imediatamente abatidos), o custo por touro selecionado é de 1,25 milhão de dólares. É evidente que o melhor touro, entretanto, pode render milhões de dólares de retorno ao longo dos anos. Ao considerar os ganhos genéticos por ciclos, segundo esse autor ocorre incremento médio de 0,215 desvio-padrão ao ano. O custo por um desvio-padrão, portanto, é de cerca de 116 milhões de dólares. Ao incorporar-se, no entanto, a GWS ao esquema de melhoramento, segundo esse autor, podem-se economizar cerca de 92%. Isso porque os custos de genotipagem de 2.500 animais da população de treinamento, 2.000 fêmeas e 500 touros (além dos custos de manutenção dos touros selecionados por três anos) é de apenas 3,2 milhões de dólares. O custo por um desvio-padrão também pode ser drasticamente reduzido, uma vez que no caso de utilização da GWS o valor é de apenas 4,17 milhões (no esquema tradicional, o custo é de 116 milhões de dólares). Isso considerando-se que os valores genéticos genômicos sejam preditos com uma acurácia de 0,75.

Dados experimentais posteriormente apresentados por Luan et al. (2009) para várias características de interesse ao melhoramento animal (entre elas produção de leite e várias características relacionadas à saúde) indicam que a acurácia dos VGG obtidos por G-BLUP e Bayes B variam entre 0,12 e 0,62. Os menores valores de acurácia se aplicam a caracteres de baixa herdabilidade. Esses autores atribuíram a magnitude das acurácias ao pequeno tamanho da população de treinamento por eles utilizadas, indicando que populações de maiores tamanhos podem proporcionar maiores acurácias. Este trabalho também indicou que as acurácias preditas por G-BLUP são, em geral, maiores que aquelas preditas por Bayes B. Como o G-BLUP assume que todas as regiões contribuem igualmente para a determinação do

caráter, ao passo que o Bayes B considera que alguns QTLs podem ter maiores efeitos, outra parcela contribui com o restante da variação, e há indicações experimentais de que: (i) a distribuição de efeitos verdadeiros está suficientemente distribuído entre os loci, (ii) há pouco benefício em se ajustarem modelos mais complexos, (iii) ao menos a maioria dos SNPs explica pequenas porções da variância genética e (iv) considerar alguns SNPs outliers (que explicam substancialmente mais da variância, como no Bayes B) não melhora as estimativas dos VGG. Segundo esses autores, esse último ponto é provavelmente devido ao fato de que há muito poucos SNPs outliers que a variância genética explicada por estes é muito pequena relativamente àquela explicada por todos os SNPs de pequeno efeito. No entanto, conforme sugerido pelos referidos autores para características cuja variância genética possa ser explicada por pequeno número de genes, Bayes B deve proporcionar resultados mais promissores que o G-BLUP.

Van Raden et al. (2009) também demonstraram a confiabilidade das predições genômicas para o melhoramento animal. Esses autores relataram acurácias dos VGG de cerca de 0,75 para o índice de mérito total, usando 38.416 SNPs genotipados em 3.576 touros testados via testes de progênies. Esse autor também concluiu que, ao contrário de muitas simulações, apenas poucos QTLs de grande efeito e muitos de pequeno efeito contribuem para a variação genética. Hayes et al. (2009) também apresentaram, na forma de revisão, os resultados de acurácia dos modelos preditivos baseados em GWS em experimentos realizados nos Estados Unidos, Austrália, Nova Zelândia e Holanda. Segundo os citados autores, os experimentos usaram populações de referência entre 650 e 4.500 touros testados via testes de progênies, e genotipados com aproximadamente 50.000 SNPs. As confiabilidades dos VGG para touros jovens sem resultados de testes de progênies variaram entre 20 e 67%, dependendo da herdabilidade das características, do número de touros nas populações de treinamento e do método estatístico utilizado para predizer os efeitos dos SNPs na população. Um achado comum em pelo menos três países (EUA, Austrália e Nova Zelândia) é que o método do G-BLUP proporcionou confiabilidades tão elevadas quanto métodos mais complexos. Nesse caso, houve a inclusão de efeito poligênico (representado pelo valor de melhoramento médio dos genitores). Essa inclusão, segundo Hayes et

al. (2009), é recomendada para capturar qualquer variância genética não associada aos marcadores e para impor pressão de seleção sobre QTLs de baixa frequência que podem não ser capturados pelo marcadores. As acurácias dos VGGs foram, em vários experimentos, superiores à acurácia do valor de melhoramento médio dos genitores, que é atual critério para selecionar jovens touros para entrar em testes de progênies. Segundo Hayes et al. (2009), isso foi suficiente para grandes companhias de melhoramento de gado começarem a comercializar touros com base em seu VGG aos 2 anos de idade. Essa estratégia deve dobrar as taxas de ganhos genéticos na indústria.

Segundo Habier et al. (2010), na maioria dos trabalhos publicados não foi demonstrada a dependência das acurácias dos VGGs com o parentesco dos indivíduos na população de treinamento. Segundo esse autor, a dependência das acurácias dos VGGs em relações genéticas aditivas, assim como a acurácia devida ao desequilíbrio de ligação, deve ser conhecida para desenvolver futuros programas de melhoramento, uma vez que parentes próximos que foram testados via testes de progênies para características quantitativas podem não estar disponíveis. Esse autor demonstrou então, com dados reais, que a acurácia dos VGGs preditos por Bayes B e G-BLUP decrescem com o decréscimo da máxima relação genética aditiva. Assim, a acurácia de futuros candidatos (a seleção) pode ser menor que aquelas relatadas em estudos anteriores (HAYES et al., 2009; Van RADEN et al., 2009), uma vez que a informação de parentes próximos não estará disponível quando a seleção com base em VGGs for aplicada.

Moser et al. (2010), posteriormente, demonstraram, também com dados reais, que avaliações genômicas acuradas de ampla população de touros e vacas podem ser realizadas com um único ensaio de genotipagem de cerca de 3.000 a 5.000 SNPs, espaçados uniformemente pelo genoma. Essa é uma estratégia interessante, pois ainda hoje, mesmo com as constantes reduções nos preços de genotipagem, genotipar elevado número de animais para dezenas de milhares de SNPs é fator limitante para a ampla aplicação da GWS ao melhoramento animal.

Outro trabalho interessante que demonstra a aplicabilidade da GWS ao melhoramento animal foi relatado por Legarra et al. (2008). Ao comparar três métodos de seleção (i. uso do pedigree e informação

fenotípica; ii. uso de marcadores cobrindo amplamente o genoma e informação fenotípica; e iii. a combinação de ambos) para quatro importantes características em uma população de camundongos, esses autores verificaram que o uso de marcadores distribuídos ao longo de todo o genoma aumentou a habilidade preditiva em 0,22. Isso sugere que a GWS possui acurácias melhores e habilidades preditivas que os modelos poligênicos clássicos (LEGARRA et al., 2008).

6.2. Aplicações da Seleção Genômica no Melhoramento de Plantas

Em relação ao melhoramento de plantas, Bernardo e Yu (2007) foram um dos primeiros a propor a aplicabilidade da GWS. Esses autores, via simulação, demonstraram as perspectivas da GWS para características quantitativas em milho. Ao simular a performance testcross de duplos-haploides em três ciclos de seleção, baseada em informações dos marcadores, para situações em que 20, 40 ou 100 QTLs estavam envolvidos no controle genético de características quantitativas (de diferentes herdabilidades), esses autores verificaram que a resposta à seleção foi 18-43% maior via GWS que a resposta via seleção recorrente assistida por marcadores moleculares (MARS – marker assisted recurrent selection). Segundo esses autores, esse esquema que minimiza fenotipagem e maximiza genotipagem é bastante favorável ao melhoramento da espécie, principalmente se o custo da genotipagem for bastante reduzido.

Wong e Bernardo (2008), posteriormente, demonstraram que a GWS é também aplicável a espécies alógamas perenes, como é o caso da palma de óleo (dendê). Esses autores demonstraram, via simulações, com tamanhos populacionais de 50 a 70, que respostas a GWS foram 4 a 25% superiores àquelas correspondentes com a seleção fenotípica, dependendo da herdabilidade da característica e do número de QTLs. Segundo esses autores, o custo por unidade de ganho foi 35 a 65% inferior com a GWS em comparação com a seleção fenotípica, quando o custo por Data Point foi considerado como US$0,15. Os citados autores demonstraram, ainda, que a GWS pode viabilizar quatro ciclos de seleção no mesmo tempo requerido normalmente para dois ciclos de seleção com base em dados fenotípicos (38 anos). Para uma espécie perene, essa é sem dúvida uma enorme vantagem.

Rutkoski et al. (2011), revisando os conceitos relacionados a GWS no melhoramento de plantas, propuseram também a implementação da GWS ao melhoramento de trigo, tendo como foco resistência durável à ferrugem do colmo causada por *Puccinia graminis*. Esses autores sugeriram a utilização de seleção recorrente recíproca (método não tradicionalmente utilizado no melhoramento de trigo) acoplado à seleção genômica como forma de se aumentar a resistência à ferrugem. Segundo sugeriram esses autores, o esquema proposto, quando comparado aos esquemas tradicionais, pode proporcionar redução de 2 vezes no tempo requerido para completar um ciclo de seleção, além de facilitar a piramidação de genes de resistência.

Albrecht et al. (2011) foram os primeiros a relatar um estudo experimental em larga escala de predições genômicas dos valores testcross de milho em avançado ciclo de melhoramento. Com base na genotipagem de 1.152 SNPs em 1,380 linha de duplos-haploides derivados de 36 cruzamentos, e nos dados fenotípicos para produção de grãos e conteúdo de matéria seca no grão em sete localidades, esses autores verificaram que os modelos baseados em GWS proporcionaram acurácias maiores que modelos baseados somente em informação de *pedigree*. A acurácia predita média baseada em dados genômicos foi elevada, mesmo para um caráter complexo como produção de grãos (0,72-0,74), quando o esquema de validação cruzada permitia alto nível de parentesco entre os *sets* de estimação e teste. Quando as predições foram realizadas em famílias distantemente relacionadas, as acurácias preditas decresceram significativamente (0,47-0,48).

Grattapaglia e Resende (2011) demonstraram a aplicação da GWS no melhoramento florestal. Ao simular, no contexto de um programa de melhoramento florestal, diferentes situações referentes a (i) tamanho da população de treinamento, (ii) herdabilidade da característica e número de QTLs na acurácia da GWS, esses autores demonstraram que a GWS tem o potencial de impactar radicalmente nos programas de melhoramento. A acurácia de referência da seleção baseada em BLUP fenotípico (0,68) pode ser igualada pela GWS mesmo em densidades consideradas baixas de marcadores (e.g. 2 marcadores/cM), quando o tamanho efetivo populacional é reduzido (e.g. < 30). Para maiores tamanhos efetivos populacionais, densidade maior de marcadores é requerida (~20 marcadores/cM). Esses autores

demonstraram, ainda, que o ciclo de melhoramento pode ser bastante reduzido e reduções da ordem de 50% podem levar a ganhos na eficiência de seleção superiores a 100%. Resende et al. (2012b), posteriormente, demonstraram, por meio de dados reais de Pinus, que as (i) acurácias dos modelos preditivos variam de 0,63 a 0,75, em razão da característica avaliada (altura de plantas e diâmetro) e (ii) a eficiência de seleção por unidade de tempo foi 53 a 112% superior usando a GWS, em comparação com a seleção fenotípica, assumindo uma redução de 50%). Uma vez que a população avaliada estava replicada clonalmente em quatro locais, esses autores puderam ainda testar a acurácia dos modelos preditivos estimados e validados em um local, quando utilizados em outro local. Foi verificado que as acurácias são elevadas quando os modelos preditivos são utilizados em uma mesma zona de melhoramento e que são consideravelmente reduzidas ao serem utilizadas em zonas distintas e distantes. Esses autores verificaram, ainda, que modelos gerados em idades precoces não têm boa performance para predizer os fenótipos em idade produtiva (e.g. 6 anos). Esse estudo foi, assim, o primeiro a demonstrar efetivamente a aplicabilidade da seleção genômica no melhoramento florestal, e os notáveis ganhos que podem ser obtidos ao utilizar modelos em idades relevantes e dentro das zonas de melhoramento no qual foram estimados. Resende et al. (2012a) demonstraram, também, a eficiência da aplicação da Seleção Genômica em uma população de eucalipto. Esses autores demonstraram, ainda, que as análises de seleção genômica capturaram proporções significativas da herdabilidade, que variaram de 74–97%, dependendo da característica.

7. Perspectivas

A seleção genômica ampla tem grande potencial para revolucionar a forma como o melhoramento de plantas e animais é realizado. Como demonstrado nas seções anteriores, resultados extremamente positivos e animadores foram obtidos em estudos-piloto em plantas e animais. Em escala operacional, o uso dessa tecnologia pode aumentar as acurácias de seleção, levando a populações melhoradas mais produtivas e proporcionando mais ganho. Em paralelo, sua aplicação pode reduzir, ou até eliminar, alguns ensaios

de campo, que são de caro estabelecimento e manutenção, além de ocupar áreas que poderiam ser destinadas à produção operacional. No melhoramento de espécies perenes e de animais, a seleção ultraprecoce deve proporcionar reduções drásticas na duração do ciclo de melhoramento e gerar ganhos por unidade de tempo nunca vistos.

Do ponto de vista científico, as pesquisas nessa área irão certamente proporcionar melhores entendimentos sobre o controle genético de caracterítsicas quantitativas. Perguntas há muito estudadas na genética quantitativa, como a sobredominância, *inbreeding* e vigor híbrido, podem-se beneficiar dessas novas informações que, agora, são baseadas nos QTLs em cada indivíduo, ao oposto de aproximações feitas em populações baseadas na segregação esperada, dado certo parentesco.

Paralelo aos avanços das metodologias estatísticas, os métodos de genotipagem evoluíram muito nos últimos anos. Métodos de redução de complexidade do genoma, aliados a plataformas de sequenciamento de última geração (Illumina HiSeq 2000) estão sendo usados para descoberta de SNPs (ELSHIRE et al., 2011; RESENDE et al., 2012D). Os dados do preço de sequenciamento projetam uma redução de custo drástica, de maneira que a genotipagem de milhares de marcadores em milhares de indivíduos irá, em breve, tornar-se acessível a todas as espécies. Assim, o uso de elevado número de marcadores permite a geração de modelos preditivos mais acurados e que são mantidos por número maior de gerações sem a necessidade de reestimação. Além disso, com número maior de marcadores, a tendência é de cada vez mais obter marcadores mais próximos do polimorfismo causador (QTN – *Quantitative Trait Nucleotide*). Assim, a seleção genômica ampla tende a gerar resultados idênticos àqueles gerados pelos estudos de associação(GWAS – *Genome Wide Association Studies*). O objetivo desses estudos é identificar os genes controladores de caracteristicas quantitativas. Embora esses estudos tenham objetivo diferente da Seleção Genômica, a perspectiva é de que, com número elevado de marcadores, essas duas técnicas tendem a convergir, ou seja, a GWS atuará sobre os genes propriamente ditos.

Por fim, a genotipagem de grande número de marcadores irá requerer o desenvolvimento de métodos analíticos de seleção de covariáveis que reduzem a dimensionalidade do espaço analisado e

que são eficientes do ponto de vista de processamento. Os autores deste capítulo esperam, nos próximos anos, um *boom* de publicações com diferentes métodos, de maneira semelhante ao que ocorreu com as metodologias de detecção de QTLs no final dos anos 1990.

Referências

ALBRECHT, T.; WIMMER, V.; AUINGER, H.J.; ERBE, M.; KNAAK, C. et al. Ggenome-based prediction of testcross values in maize. **Theoretical and Applied Genetics, v.** 123, p. 339-350, 2011.

BAIRD, N.A.; ETTER, P.D.; ATWOOD, T.S.; CURREY, M.C.; SHIVER, A.L. et al. **Rapid snp discovery and genetic mapping using sequenced rad markers**. PLOS ONE **3**. 2008.

BERNARDO, R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. **Crop Science**, N. 48, p. 1649-1664, 2008.

BERNARDO, R.; YU, J.M. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, v. 47, p. 1082-1090, 2007.

CALUS, M.P.L.; MEUWISSEN, T.H.E.; DE ROOS, A.P.W.; VEERKAMP, R.F. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 78, p. 553-561, 2008.

COOPER, M.; VAN EEUWIJK, F.A.; HAMMER, G.L.; PODLICH, D.W.; MESSINA, C. Modeling qtl for complex traits: detection and context for plant breeding. **Current Opinion in Plant Biology, v.** 12, p. 231-240, 2009.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, p. 375-385, 2009.

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

ELSHIRE, R.J.; GLAUBITZ, J.C.; SUN, Q.; POLAND, J.A.; KAWAMOTO, K. et al. **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species**. PLOS ONE 6, 2011.

FALCONER, D.S. **Introduction to quantitative genetics**. Wiley: Longman; Harlow: Burnt Mill; Essex, England; New York, 1989.

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

FRARY, A.; NESBITT, T.C.; GRANDILLO, S.; KNAAP, E.; CONG, B. et al. FW2.2: a quantitative trait locus key to the evolution of tomato fruit size. **Science**, v. 289, p. 85-88, 2000.

GARRICK, D.J.; TAYLOR, J.F.; FERNANDO, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics, Selection, Evolution**, GSE, v. 41, **p.** 55, 2009.

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

GODDARD, M.E.; HAYES, B.J. GENOMIC SELECTION. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.

GRATTAPAGLIA, D. Mapas genéticos e seleção assistida por marcadores moleculares. IN: BOREM, A. (Ed.). Biotecnologia Florestal. Viçosa, MG, 2007. p. 201-230.

GRATTAPAGLIA, D.; BRADSHAW, H.D. Nuclear-dna content of commercially important eucalyptus species and hybrids. **Canadian Journal of Forest Research**, v. 24, p. 1074-1078, 1994.

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

HABIER, D.; FERNANDO, R.L.; DEKKERS, J.C.M. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. **Genetics**, 2008.

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

HABIER, D.; TETENS, J.; SEEFRIED, F.R.; LICHTNER, P.; THALLER, G. The impact of genetic relationship information on genomic breeding values in german holstein cattle. **Genetics Selection Evolution**, v. 42, 2010.

HASTBACKA, J.; DELACHAPELLE, A.; KAITILA, I.; SISTONEN, P.; WEAVER, A. et al. Linkage disequilibrium mapping in isolated founder populations Diastrophic dysplasia in finland. **Nature Genetics**, v. 2, p. 204-211, 1992.

HAYES, B.J.; BOWMAN, P.J.; CHAMBERLAIN, A.J.; GODDARD, M.E. Invited review: genomic selection in dairy cattle: progress and challenges (v. 92, p. 433, 2009). **Journal of Dairy Science**, v. 92, p. 1313-1313, 2009.

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

JENKINS, S.; GIBSON, N. High-throughput snp genotyping. **Comparative and Functional Genomics**, v. 3, **p.** 57-66, 2002.

KIRST, M. Forest genomics: new approaches, challenges and perspectives. In: BOREM, A. (Ed.). **Biotecnologia Florestal.** Viçosa, MG, 2007. p. 231-252.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, 1990.

LEGARRA, A.; ROBERT-GRANIE, C.; CROISEAU, P.; GUILLAUME, F.; FRITZ, S. Improved lasso for genomic selection. **Genetics Research, v.** 93, p. 77-87, 2011.

LEGARRA, A.; ROBERT-GRANIE, C.; MANFREDI, E.; ELSEN, J.M. Performance of Genomic Selection in Mice. **Genetics**, v. 180, p. 611-618, 2008.

LUAN, T.; WOOLLIAMS, J.A.; LIEN, S.; KENT, M.; SVENDSEN, M. et al. The accuracy of genomic selection in norwegian red cattle assessed by cross-validation. **Genetics**, v. 183, p. 1119-1126, 2009.

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

MEUWISSEN, T.H.; SOLBERG, T.R.; SHEPHERD, R.; WOOLLIAMS, J.A. A fast algorithm for bayesb type of prediction of genome-wide estimates of genetic value. **Genetics, Selection, Evolution** GSE, v. 41, p. 2, 2009.

MOSER, G.; KHATKAR, M.S.; HAYES, B.J.; RAADSMA, H.W. Accuracy of direct genomic values in holstein bulls and cows using subsets of snp markers. **Genetics Selection Evolution**, v. 42, 2010.

MOSER, G.; TIER, B.; CRUMP, R.E.; KHATKAR, M.S.; RAADSMA, H.W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. Genetics, Selection, Evolution GSE, v. 41, p. 56, 2009.

PATERSON, A.H.; DAMON, S.; HEWITT, J.D.; ZAMIR, D.; RABINOWITCH, H.D. et al. Mendelian factors underlying quantitative traits in tomato Comparison across species, generations, and environments. **Genetics**, v. 127, p. 181-197, 1991.

PÉREZ, P.; DE LOS CAMPOS, G.; CROSSA, J.; GIANOLA, D. Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R. **The Plant Genome Journal**, v. 3, p. 106, 2010.

RESENDE, M.D.; RESENDE, M.F.; SANSALONI JR., C.P.; PETROLI, C.D.; MISSIAGGIA, A.A. et al. Genomic selection for growth and wood quality in *eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **The new phytologist**, 2012A.

RESENDE, M.D.V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília, 2002.

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

RESENDE, M.D.V.; RESENDE, M.F.R.; AGUIAR, A.M.; ABAD, J.I.M.; MISSIAGGIA, A.A. et al. **Computação da Seleção Genômica Ampla (GWS)**. Colombo, PR: Embrapa Florestas, 2010.

RESENDE, M.F.R.; MUÑOZ, P.; ACOSTA, J.J.; PETER, G.F.; DAVIS, J.M. et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytologist**, v. 193, p. 617-624, 2012B.

RESENDE, M.F.R.; MUÑOZ, P.; RESENDE, M.D.; GARRICK, V.D.; FERNANDO, R.L. et al. Accuracy of genomic selection methods in a standard dataset of loblolly pine (*Pinus taeda* L.). **Genetics**, 2012C.

RESENDE, M.F.R.; NEVES, L.G.; BALMANT, K.M.; DERVINIS, C.;VANDYK, D. et al. RAPID-SEQ – A novel approach to genotype by sequencing reduced genome representations. In: Plant and animal genome, 20., 2012D, San Diego. **Proceedings...** San Diego, 2012D.

RUTKOSKI, J.E.; HEFFNER, E.L.; SORRELLS, M.E. Genomic selection for durable stem rust resistance in wheat. **Euphytica**, v. 179, p. 161-173, 2011.

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

SMARAGDOV, M.G. Genomic selection as a possible accelerator of traditional selection. **Russian Journal of Genetics, v.** 45, p. 633-636, 2009.

SVED, J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. **Theoretical Population Biology**, v. 2, p. 125-141,1971.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society**, v. 58, **p.** 267-288, 1996.

USAI, M.G.; GODDARD, M.E.; HAYES, B.J. Lasso with cross-validation for genomic selection. **Genetics Research**, v. 91, p. 427-436, 2009.

VANRADEN, P.M.; VAN TASSELL, C.P.; WIGGANS, G.R.; SONSTEGARD, T.S.; SCHNABEL, R.D. et al. Invited review: reliability of genomic predictions for north american holstein bulls. **Journal of Dairy Science**, v. 92, p. 16-24, 2009.

WHITE, T.L.; ADAMS, W.T.; NEALE, D.B. **Forest genetics.** Wallingford: CABI Publishing,, 2007.

WONG, C.K.; BERNARDO, R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. **Theoretical and Applied Genetics, v.** 116, p. 815-824, 2008.

YANO, M.; KATAYOSE, Y.; ASHIKARI, M.; YAMANOUCHI, U.; MONNA, L. et al. HD1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the arabidopsis flowering time gene constans. **The Plant Cell, v.**12, p. 2473-2484, 2000.