Dezembro, 2010

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 210

Computação da Seleção Genômica Ampla (GWS)

Marcos Deon Vilela de Resende
Márcio Fernando Ribeiro Resende Júnior
Aurelio Mendes Aguiar
Jupiter Israel Muro Abad
Alexandre Alves Missiaggia
Carolina Sansaloni
Cesar Petroli
Dario Grattapaglia

Embrapa Florestas
Colombo, PR
2010

Exemplares desta publicação podem ser adquiridos na:

Embrapa Florestas
Estrada da Ribeira, Km 111, Guaraituba,
83411-000, Colombo, PR - Brasil
Caixa Postal: 319
Fone/Fax: (41) 3675-5600
www.cnpf.embrapa.br
sac@cnpf.embrapa.br

Supervisão editorial: Patrícia Póvoa de Mattos
Revisão de texto: Mauro Marcelo Berté
Normalização bibliográfica: Francisca Rasche
Editoração eletrônica: Mauro Marcelo Berté

1ª edição
1ª impressão (2010): sob demanda

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Florestas

Computação da Seleção Genômica Ampla (GWS) [recurso eletrônico] / Marcos Deon Vilela de Resende ... [et al.]. Dados eletrônicos - Colombo : Embrapa Florestas, 2010.
CD-ROM. - (Documentos / Embrapa Florestas, ISSN 1679-2599 ; 210)

1. Melhoramento vegetal 2. Melhoramento animal. 3. Seleção genotípica. 4. Seleção fenotípica. 5. Software. I. Resende, Marcos Deon Vilela de. II. Resende Júnior, Márcio Fernando Ribeiro. III. Aguiar, Aurelio Mendes. IV. Abad, Jupiter Israel Muro. V. Missiaggia, Alexandre Alves. VI. Sansaloni, Carolina. VII. Petroli, Cesar. VIII. Grattapaglia, Dario. IX. Série.

CDD 631.52 (21. ed.)

Autores

Marcos Deon Vilela de Resende
Estatístico, Doutor,
Pesquisador da Embrapa Florestas
deon@cnpf.embrapa.br

Márcio Fernando Ribeiro Resende Júnior
Engenheiro Florestal, Mestre,
Doutorando na Universidade da Flórida
mresende@ufl.edu

Aurelio Mendes Aguiar
Engenheiro Agrônomo, Doutor,
Pesquisador da Fibria
aurelio.aguiar@fibria.com.br

Jupiter Israel Muro Abad
Engenheiro Florestal, Doutor,
Pesquisador da Fibria
jmuro@fibria.com.br

Alexandre Alves Missiaggia
Engenheiro Florestal, Doutor,
Pesquisador da Fibria
amissiaggia@fibria.com.br

Carolina Paola Sansaloni
Bacharel em Genética, Mestre
Doutorando em Biologia Molecular
Universidade de Brasília
carosansaloni@hotmail.com

Cesar Daniel Petroli
Bacharel em Genética, Mestre
Doutorando em Biologia Molecular
Universidade de Brasília
petrolic@hotmail.com

Dario Grattapaglia
Engenheiro Florestal, Doutor,
Pesquisador da Embrapa Recursos
Genéticos e Biotecnologia
dario@cenargen.embrapa.br

Apresentação

A seleção genômica ampla é uma metodologia que, de forma pioneira, integra as tecnologias genômicas e as ferramentas da genética quantitativa e do melhoramento, propiciando um grande salto qualitativo nos sistemas de avaliação genética. Esta nova abordagem experimental vem rapidamente mudando os paradigmas do melhoramento genético de animais domésticos e plantas, causando uma verdadeira revolução na nossa capacidade de prever fenótipos e, com isso, aumentar a acurácia seletiva em idade precoce, maximizando o ganho genético por unidade de tempo.

A presente publicação aborda os métodos estatísticos, as estratégias computacionais e softwares para a implementação prática da seleção genômica ampla no melhoramento genético. Elaborada por uma equipe interinstitucional, envolvendo a academia e a iniciativa privada, apresenta também um exemplo prático de sua aplicação no melhoramento do eucalipto. Assim, trata-se de uma contribuição à difusão, ensino e aplicação operacional da seleção genômica nos programas de melhoramento de plantas e animais em andamento no País.

Helton Damin da Silva
Chefe-Geral

Sumário

Computação da Seleção Genômica Ampla (GWS)

Marcos Deon Vilela de Resende
Márcio Fernando Ribeiro Resende Júnior
Aurelio Mendes Aguiar
Jupiter Israel Muro Abad
Alexandre Alves Missiaggia
Carolina Sansaloni
Cesar Petroli
Dario Grattapaglia

Métodos para GWS

A seleção genômica ampla (GWS) ou seleção genômica (GS) foi proposta por Meuwissen et al. (2001) como uma forma de aumentar a eficiência e acelerar o melhoramento genético. A GWS enfatiza a predição simultânea (sem o uso de testes de significância para marcas individuais) dos efeitos genéticos de milhares de marcadores genéticos de DNA (SNP, DArT, microssatélites) dispersos em todo o genoma de um organismo, de forma a capturar os efeitos de todos os locos (tanto de pequenos quanto de grandes efeitos) e explicar toda a variação genética de um caráter quantitativo. A condição fundamental para isso é que haja desequilíbrio de ligação, em nível populacional, entre alelos dos marcadores e alelos dos genes que controlam o caráter. A predição dos efeitos genéticos é realizada com base em dados genotípicos e fenotípicos de indivíduos pertencentes a uma amostra da população de seleção.

Esses efeitos genéticos dos marcadores sobre fenótipos de caracteres quantitativos são somados e usados na predição de valores genéticos de indivíduos apenas genotipados, candidatos à seleção em programas de melhoramento genético. A predição

e a seleção podem ser realizadas em fases muito juvenis de plantas e animais, acelerando assim o processo de melhoramento genético. Adicionalmente, a própria predição tende a ser mais acurada por considerar o real parentesco genético dos indivíduos em avaliação, em detrimento do parentesco médio esperado matematicamente (RESENDE, 2007). A GWS propicia uma forma de seleção precoce direta (SPD), pois, atua precocemente sobre genes expressos na idade adulta. Ao contrário, a seleção precoce tradicional é indireta, pois atua (via avaliação fenotípica) sobre genes ativados na idade precoce, esperando que esses informem parcialmente sobre genes expressos na idade adulta. Assim, a SPD propiciada pela GWS é especialmente importante para o melhoramento de organismos perenes como animais, espécies florestais, fruteiras (e outras frutíferas), forrageiras, cana-de-açúcar, dentre outras.

Em resumo, a superioridade da GWS sobre a seleção baseada em fenótipos pode ser atribuída a quatro fatores: (i) uso da matriz de parentesco real e própria de cada caráter, fato que aumenta a acurácia seletiva; (ii) viabilização da SPD, que aumenta o ganho genético por unidade de tempo; (iii) permissão da avaliação repetida de cada alelo (propicia repetição experimental) sem o uso de testes clonais e de progênies, fato que aumenta a acurácia seletiva; (iv) uso de maior número de informações, combinando três tipos de informação (fenotípica, genotípica e genealógica) para corrigir e desregressar os dados e fazer a análise genômica, fato que aumenta a acurácia.

A GWS é um produto do terceiro milênio. Após a proposição da GWS em 2001, o procedimento permaneceu discreto até 2007, quando vários trabalhos abordaram o método e sua acurácia no melhoramento animal e vegetal (FERNANDO et al., 2007; GODDARD; HAYES, 2007; MEUWISSEN, 2007; BERNARDO; YU, 2007; RESENDE 2007). Outros trabalhos relatam que a GWS é o novo paradigma em genética quantitativa (RESENDE,

2008; GIANOLA et al., 2009), melhoramento de gado de leite (HAYES et al., 2009; VAN RADEN, 2008; VAN RADEN et al., 2009), de aves (GONZALEZ-RECIO et al., 2009), de plantas anuais (HEFFNER et al., 2009), de espécies florestais (RESENDE et al. 2008; GRATTAPAGLIA; RESENDE, 2010) e de outras plantas perenes (DIAS; RESENDE, 2009).

Vários métodos de predição de valores genéticos genômicos foram propostos: quadrados mínimos (LS), BLUP/GWS, BayesA e BayesB (MEUWISSEN et al., 2001), regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA et al., 2006), aprendizado de máquina (LONG et al., 2007), regressão RKHS (*Reproducing Kernel Hilbert Spaces*) (GIANOLA et al., 2008), LASSO Bayesiano (PARK; CASELLA, 2008; CAMPOS et al., 2009), Bayes B Acelerado (MEUWISSEN, 2009), Bayes C, BayesCπ, BayesD, BayesDπ (HABIER et al., 2010), Regressão via Quadrados Mínimos Parciais (PLSR) (SOLBERG et al., 2009) e Regressão via Componentes Principais (PCR) (SOLBERG et al., 2009). Os métodos BLUP (regressão aleatória), LASSO (*Least Absolute Shrinkage and Selection Operator*) e Bayes A e B pertencem à classe de regressão explícita. Por outro lado, o método RKHS pertence à classe de regressão implícita e é um método semiparamétrico (GIANOLA; VAN KAAM, 2008; GIANOLA; CAMPOS, 2009).

Conforme Resende (2007, 2008), essas abordagens diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos; o AM assume que existe um número limitado de genes e de SNPs a serem ajustados; o método BayesB é intermediário entre esses dois, assumindo poucos genes de grandes efeitos e muitos genes com pequenos efeitos. No método BayesB muitos efeitos de marcadores são assumidos como zero, a priori. Isso reduz o tamanho do genoma por meio da concentração nas partes do mesmo onde existem QTLs. O

melhor método é aquele que reflete melhor a natureza biológica do caráter poligênico em questão, em termos de efeitos gênicos.

O método quadrados mínimos (regressão fixa) é ineficiente devido a: impossibilidade de estimar todos os efeitos simultaneamente, pois o número de efeitos a estimar é maior do que o número de dados; estimando um efeito de cada vez e verificando a sua significância, conduz a superestimativas dos efeitos significativos; a acurácia do método é baixa; somente QTLs de grande efeito serão detectados e usados e, consequentemente, nem toda a variação genética será capturada pelos marcadores.

O método de quadrados mínimos assume distribuição a priori para os QTLs, com variância infinitamente grande, fato que é incompatível com a conhecida variância genética total. O BLUP/GWS assume os efeitos de QTL com distribuição normal com variância constante para todos os segmentos cromossômicos. Esse método contorna, por meio da estimação simultânea dos efeitos de todos os marcadores, a questão da necessidade de estimação de um grande número de efeitos a partir de um tamanho amostral restrito e, adicionalmente, a questão do fato de que muitos efeitos mostram colinearidade advinda do desequilíbrio de ligação entre os próprios marcadores.

A distribuição dos efeitos de QTL é conhecida em poucos caracteres e espécies. Em gado bovino leiteiro, Goddard e Hayes (2007) relatam a presença de 150 QTLs para o caráter produção de leite e estimaram a distribuição de seus efeitos como aproximadamente exponencial (HAYES; GODDARD, 2009). Com distribuição exponencial e não muitos efeitos com valor zero, o melhor estimador dos efeitos alélicos é denominado LASSO (TIBSHIRANI, 1996). Entretanto, com muitos efeitos com valor zero, o LASSO não é adequado. Usai et al. (2009) compararam o LASSO com BLUP e Bayes A empregando 156

SNPs significativos. As acurácias obtidas foram das ordens de 0,89, 0,75 e 0,84, respectivamente. Assim, o LASSO é uma boa opção quando se usa um número limitado de marcadores.

Comparações entre os métodos de predição de valores genéticos genômicos têm sido realizadas. Meuwissen et al. (2001) concluíram pela superioridade teórica do método Bayes B, o qual mostrou-se ligeiramente superior ao BLUP. Hayes et al. (2009) avaliaram a efetividade prática da seleção genômica em gado de leite nos Estados Unidos, Austrália e Nova Zelândia. Concluíram que o método BLUP mostrou-se aproximadamente igual a outros métodos mais complexos, em termos de acurácia. Isso ocorre para caracteres em que o modelo infinitesimal (muitos genes de pequenos efeitos) se aplica. Adicionalmente, o método BLUP é vantajoso porque a única informação a priori necessária é uma estimativa da variância genética aditiva do caráter. Os autores relataram também a importância da inclusão do efeito poligênico no modelo de avaliação genética, como forma de capturar e selecionar QTLs de baixa frequência não capturados pelos marcadores.

Habier et al. (2007) compararam os métodos de quadrados mínimos (denominado por eles como regressão fixa ou FR-LS), BLUP (denominado por eles como regressão aleatória ou RR-BLUP) e Bayes B, em termos de acurácia seletiva na seleção em longo prazo, após várias gerações depois da predição dos efeitos genéticos dos marcadores. Nessa situação, a acurácia tende a diminuir devido à modificação das relações de parentesco (em relação ao parentesco na geração de estimação dos efeitos genômicos), mas há um componente persistente da acurácia devido ao LD. Os resultados mostraram que o decréscimo na acurácia devido à modificação das relações de parentesco é maior no método RR-BLUP. Inicialmente, os métodos RR-BLUP e Bayes B apresentaram acurácia similar, mas, após 11 gerações, o método Bayes B superou o RR-BLUP.

Os métodos bayesianos estão associados a sistemas de equações não lineares e as predições não lineares podem ser melhores quando os efeitos de QTL não são normalmente distribuídos, devido à presença de genes de efeitos maiores. As predições lineares associadas ao RR-BLUP assumem que todos os marcadores contribuem igualmente para a variação genética (ausência de genes de efeitos maiores).

Gonzalez-Recio et al. (2008) compararam métodos não paramétricos ou semiparamétricos (RKHS), regressão bayesiana e BLUP/GWS em termos de eficiência na seleção genômica. Concluíram que o método da regressão RKHS (*Reproducing Kernel Hilbert Spaces*) apresentou melhor capacidade preditiva do que os demais. Esse método equivale ao BLUP modelo animal com a matriz de parentesco substituída pelos *kernels* (RESENDE, 2008). O método não paramétrico RKHS parece ter maior capacidade preditiva quando aplicado a dados reais (GIANOLA et al., 2009). O Lasso Bayesiano é também interessante pois usa amostragem de Gibbs, uma vez que se conhece a distribuição condicional completa. Outros métodos foram avaliados por Solberg et al. (2009): Regressão via quadrados mínimos parciais (PLSR) e Regressão via componentes principais (PCR). Concluíram que esses são mais simples e rápidos computacionalmente, porém menos acurados que o Bayes B, com acurácias da ordem de 0.68 (PLSR e PCR) e 0.84 (Bayes B).

Gianola et al. (2009) fazem uma análise crítica dos métodos associados a modelos hierárquicos bayesianos (Bayes A e B) especificamente em relação às suas formulações em termos dos hiperparâmetros que propiciam variâncias específicas para cada marcador. Segundo os autores, nenhum dos métodos permite o aprendizado bayesiano sobre essas variâncias para prosseguir para longe das prioris. Em outras palavras, os hiperparâmetros a priori para essas variâncias sempre terão influência na extensão do *shrinkage* produzido nos efeitos dos marcadores. O usuário

do método pode controlar a quantidade de *shrinkage* apenas arbitrariamente, por meio da variação nos parâmetros υ e S (associados à distribuição qui-quadrado invertida assumida como priori). Segundo os autores, o método Bayes B não é bem formulado no contexto bayesiano. Isto porque designar a priori que $\sigma^2_g = 0$, não conduz necessariamente a $g_i = 0$, conforme intenção original de Meuwissen et al. (2001), em que g_i é o efeito genético do loco i. Sugerem então que o estado zero seja especificado ao nível dos efeitos e não ao nível das variâncias. Assim, à probabilidade de mistura π poderia ser atribuída uma distribuição a priori Beta. Surge então, o método Bayes C que é vantajoso e permite especificar uma distribuição a priori para π, permitindo a modelagem da distribuição dupla exponencial. Os métodos bayesianos para a GWS são tratados com mais detalhe em Resende (2008).

Vários outros métodos bayesianos foram propostos (Bayes Cπ, Bayes D, Bayes Dπ, conforme Habier et al., 2010), todos eles com o propósito de facilitar a aplicação do método Bayes B que é conceitualmente ideal, mas computacionalmente lento. Habier et al. (2010) relataram que nenhum dos métodos bayesianos são claramente superiores dentre eles; entretanto, o Bayes B, Bayes Cπ e Bayes Dπ apresentam a vantagem de propiciar informação sobre a arquitetura genética do caráter quantitativo e identificar as posições de QTL por modelagem da frequência de SNP.

Os métodos bayesianos são superiores quando a distribuição dos efeitos dos QTL é leptocúrtica (curtose positiva), devido à presença de genes de grandes efeitos. Com distribuição normal dos efeitos dos QTL, o método RR-BLUP é igualmente eficiente. Provavelmente, isso se aplica para a maioria dos caracteres quantitativos, pois, genes de grandes efeitos tendem a ser fixados ou eliminados e não mantidos em nível polimórfico nas populações. O RR-BLUP assume iguais e pequenas herdabilidades por loco, causando uma alta regressão em QTL de grandes efeitos, os quais deveriam receber herdabilidades mais altas.

O uso de um único $\lambda = \sigma_e^2 / \sigma_{gi}^2 = \sigma_e^2 /(\sigma_g^2 / n)$ para todos os marcadores não conduz nessariamente ao mesmo *shrinkage* para todos, pois, a variação nas frequências alélicas entre marcadores podem conduzir a diferentes graus de *shrinkage*.

Teoria dos métodos de regressão

Em um problema de regressão tem-se que a variável dependente y é dada como função de uma variável preditora (x) e vetor de erros aleatórios (e), segundo o modelo $y = \beta' x + e$. No contexto da seleção genômica, define-se x como um vetor de genótipos marcadores codominantes codificados como 0, 1 ou 2, de acordo com o número de cópias de um dos alelos do loco marcador. E β é definido como um vetor de coeficientes de regressão que contemplam os efeitos dos marcadores (via desequilíbrio de ligação com os genes) no caráter fenotípico y.

Usando esperança condicional, a equação de regressão é dada por:

$$\hat{y} = \hat{\beta}' x = E(y \mid x)$$

Isso implica

$$\hat{\beta} = E(\beta \mid x, y) = [\int \beta\, p(\beta) p(y \mid \beta, x)\, d\beta\,] / [\int p(\beta) p(y \mid \beta, x)\, d\beta\,]$$

em que $p(\beta)$ é a função densidade de probabilidade de β e

$p(y \mid \beta, x)$ é a função de verossimilhança de y.

Assim, a predição de y depende de $p(\beta)$, ou seja, da distribuição dos efeitos (via LD com os QTLs) dos marcadores. Essa distribuição pode ser tratada como informação ou distribuição a priori no contexto bayesiano ou como variável aleatória no contexto frequentista. Se $\beta \sim N(0, \sigma_\beta^2)$, $\hat{\beta}$ é

BLUP de β e $\hat{y}$ é BLUP de y. Isto implica que os efeitos de todos os marcadores são tomados da mesma distribuição. Alternativamente, pode ser assumido que $\beta_i \sim N(0, \sigma^2_{\beta_i})$, em que β_i é tomado de uma distribuição qui-quadrado invertida, segundo o enfoque bayesiano. Nesse caso, isso implica que grande número de marcadores apresenta efeitos pequenos e poucos marcadores apresentam efeitos grandes.

Esse método BLUP para os coeficientes de regressão é denominado regressão aleatória ou regressão de cumeeira (*Ridge Regression*) (RR-BLUP). Os coeficientes de regressão *ridge* são definidos como aqueles que minimizam a soma de quadrados penalizada dada por $(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\,\beta_i)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$

em que λ é o parâmetro de penalização (ou *shrinkage*) ou parâmetro *ridge*, n é o número de marcadores e N é o número de indivíduos. O primeiro termo da equação é a soma de quadrados dos resíduos da regressão e o segundo termo é a penalização, a qual depende da magnitude dos coeficientes de regressão via $\sum_{i=1}^{n} \beta_i^2$.

Outro método relacionado é o LASSO, que combina *shrinkage* (regularização) com seleção de variáveis e envolve o seguinte problema de otimização, via minimização de $(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \lambda \sum_{i=1}^{n} |\beta_i|$, em que $\sum_{i=1}^{n} |\beta_i|$ é a soma dos valores absolutos dos coeficientes de regressão. As soluções em que os coeficientes de regressão se distanciam de zero sofrem penalização. O componente $\lambda \sum_{i=1}^{n} |\beta_i|$

regulariza a regressão sem penalizar muito. O parâmetro de suavização λ controla a intensidade da regularização. O LASSO pode ser implementado também via abordagem bayesiana, em que λ controla a precisão da distribuição a priori atribuída aos coeficientes de regressão.

Computação do método *Random* (*Ridge*) *Regression* BLUP (RR-BLUP/ GWS)

O método RR-BLUP/GWS usa preditores do tipo BLUP, mas os efeitos de marcadores não são ajustados como variáveis classificatórias e sim como variáveis explicativas ou explanatórias. Assim, são variáveis regressoras e são ajustadas como covariáveis de efeitos aleatórios, ou seja, os fenótipos são regressados com base nessas covariáveis. O fato de serem covariáveis e não variáveis classificatórias, conduz a diferentes matrizes de incidência e consequentemente diferentes algoritmos computacionais em relação ao BLUP tradicional. O nome mais apropriado é regressão aleatória (*random regression*) do tipo BLUP (RR-BLUP) aplicado à seleção genômica ampla (RR-BLUP/ GWS). A técnica da regressão aleatória é um tipo especial da regressão de cumeeira (*ridge regression*).

Os estimadores associados à regressão aleatória e regressão de cumeeira promovem *shrinkage* ditado por uma função da quantidade λ. Quando λ não é conhecido, a escolha arbitrária do mesmo leva ao método de regressão *ridge regression* (RR). Se o parâmetro de regressão for associado a $\lambda = \sigma_e^2 / \sigma_{gi}^2 = \sigma_e^2 / (\sigma_g^2 / n)$ tem-se a regressão aleatória BLUP para o efeito do segmento cromossômico i, em que σ_{gi}^2 é a variância genética associada ao loco ou segmento i e σ_g^2 e σ_e^2 são a variância genética do caráter e variância residual, respectivamente. A quantidade n é desconhecida a priori, mas pode ser inferida conforme descrito adiante. O parâmetro de penalização λ pode também ser determinado por via iterativa ou sintonia, escolhendo-se aquele que maximiza a correlação entre valor fenotípico e valor genético predito na validação cruzada. Whittaker et al. (2000) e Meuwissen et al. (2001) foram pioneiros em propor a predição simultânea dos efeitos dos marcadores, sem o uso de testes de significância para marcas individuais. Isto contrasta com

o método da MAS proposto por Lande e Thompson (1990) e discutido por Gianola et al. (2003).

A distinção entre regressão fixa, regressão *ridge* e regressão aleatória está associada ao parâmetro de penalização λ^*, o qual é dado por $\lambda^* = (1-h^2)/h^2$. Valores pequenos de λ^* já são suficientes para reduzir o impacto da multicolinearidade entre as covariáveis presente na matriz Z´Z, que é aproximadamente singular. Valor de λ^* igual a zero (valor de h^2 igual a 1) caracteriza a regressão fixa. Valores de λ^* pequenos (0,01 a 1) caracterizam a regressão *ridge* e valores altos de λ^* (maiores que 0,1) caracterizam a regressão aleatória (Tabela 1). No caso do método RR-BLUP, um mesmo parâmetro de penalização é aplicado para todas as marcas, ao passo que o método Bayes B permite definir λ^* específicos para cada marcador.

Tabela 1. Classificação dos tipos de regressão em função do parâmetro de penalização.

Tipo de Regressão	**Penalização** $\lambda^* = (1-h^2)/h^2$	**Herdabilidade Individual** $h^2 = 1(1+\lambda^*)$
Fixa	0,00	1
Ridge; Aleatória	0,11	0,9
Ridge; Aleatória	0,25	0,8
Ridge; Aleatória	0,43	0,7
Ridge; Aleatória	0,67	0,6
Ridge; Aleatória	1,00	0,5
Aleatória	1,50	0,4
Aleatória	2,33	0,3
Aleatória	4,00	0,2
Aleatória	9,00	0,1
Aleatória	99,0	0,01
Aleatória	999,0	0,001

A predição via RR-BLUP/GWS é descrita a seguir com base em Resende (2007a; 2008). O seguinte modelo linear misto geral é ajustado para estimar os efeitos dos marcadores:

$y = Xb + Zm + e,$

em que *y* é o vetor de observações fenotípicas, *b* é o vetor de efeitos fixos, *m* é o vetor dos efeitos aleatórios de marcadores e *e* refere-se ao vetor de resíduos aleatórios. *X* e *Z* são as matrizes de incidência para *b* e *m*.

A matriz de incidência *Z* contém os valores 0, 1 e 2 para o número de alelos do marcador (ou do suposto QTL) em um indivíduo diplóide. Outra forma equivalente de codificar é usar os valores -1, 0 e 1. Esse é o modelo genotípico. Se células haplóides (sêmen) são usadas, tem-se o modelo gamético, e a matriz *Z* contém apenas os valores 0 e 1.

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

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

O valor genético genômico global do indivíduo *j* é dado por $VGG = \hat{y}_j = \sum_i Z_i \hat{m}_i$,

em que Z_i equivale a 0, 1 ou 2 para os genótipos aa, Aa e AA, respectivamente, para marcadores bialélicos e codominantes como os SNPs. Para marcadores dominantes, AA e Aa ficam confundidos e deveriam receber um peso médio de 1,33 na matriz Z, dado por 0,33 x 2 + 0,66 x 1.

As equações de predição apresentadas acima assumem a priori que todos os locos explicam iguais quantidades da variação

genética. Assim, a variação genética explicada por cada loco é dada por σ_g^2 / n, em que σ_g^2 é a variação genética total e *n* é o número de locos (quando cada loco está perfeitamente marcado por uma só marca). A variação genotípica σ_g^2 pode ser estimada por REML sobre os dados fenotípicos da maneira tradicional ou pela própria variação entre os marcadores ou segmentos cromossômicos de QTL, conforme descrito adiante.

Verifica-se que não há necessidade de uso da matriz de parentesco. A matriz de parentesco baseada em pedigree usada no BLUP tradicional é substituída por uma matriz de parentesco estimada pelos marcadores. Essa matriz de parentesco é função da própria matriz Z'Z presente nas equações de modelo misto. Esse procedimento é superior ao uso do pedigree, pois efetivamente captura a matriz de parentesco realizada para cada caráter e não uma matriz de parentesco médio associada ao pedigree. Por exemplo, a correlação genética aditiva entre dois irmãos completos, baseada em pedigree é 0,5. Mas os marcadores podem indicar que o valor verdadeiro é uma fração entre 0 e 1. O valor 0,5 é esperado em média. Mas a correlação pode ser 0, 0,5 ou 1,0, em cada loco, em função do número de alelos idênticos compartilhados entre os dois irmãos.

A matriz de parentesco realizada pode ser também computada à parte e incorporada nas equações de modelo misto do BLUP tradicional, conforme o modelo (jii) descrito a seguir. Nesse caso, ela é dada por $A = (Z^* Z^{*\prime}) / [\, 2 \sum_i^n p_{\ i} (1 - p_i) \,]$

(para SNPs sob modelo genotípico) em que p_i é a frequência de um dos alelos do loco i e Z* refere-se à matriz Z corrigida para suas médias em cada loco ($2p_i$). Para garantir A como uma matriz positiva definida, pode-se obter $A^p = A + 10^{-6} I$, em que I é uma matriz identidade. O coeficiente de endogamia genômico para o indivíduo i é dado por $A_{ii} - 1$. Outra forma de obter A é via $A = Z^* D Z^{*\prime}$, em que *D* é diagonal com D_{ii} dado

por $D_{ii} = 1/\{n\ [\ 2\ p_i\ (1-p_i)\ ]\ \}$, em que n é o número de marcadores (VAN RADEN, 2008).

A diagonal da matriz ZZ' contempla o parentesco de um indivíduo com ele mesmo e os elementos fora da diagonal mostram o número de alelos compartilhados por parentes. A correlação de Wright entre parentes pode ser obtida dividindo esses elementos fora da diagonal pelo produto das raízes quadradas dos respectivos elementos da diagonal. Por outro lado, a diagonal da matriz Z'Z mostra quantos indivíduos herdaram cada alelo e elementos fora da diagonal indicam quantas vezes dois alelos diferentes foram herdados pelo mesmo indivíduo. Usando métodos genômicos, o conceito de endogamia em um loco neutral não é mais válido, pois são consideradas medidas de parentesco nos locos do próprio caráter sob seleção. As medidas tradicionais de endogamia baseadas em pedigree resultam em perda de diversidade muito mais variáveis.

A predição de valores genéticos genômicos via BLUP pode ser computada via três métodos equivalentes (VAN RADEN, 2008):

Via RR-BLUP, conforme especificado acima, em que:

$\hat{g} = Z\,\hat{m} = Z\,(Z'R^{-1}Z + I\,\lambda)^{-1}\,Z'R^{-1}(y - X\,\hat{b})$, visto que $\hat{m} = (Z'R^{-1}Z + I\,\lambda)^{-1}Z'R^{-1}(y - X\,\hat{b})$. R é uma matriz diagonal de pesos para ponderar y com diferentes confiabilidades. Com confiabilidades altas e homogêneas (maiores que 0,85), pode-se considerar R = I e o sistema simplifica para $\hat{m} = (Z'Z + I\,\lambda)^{-1}Z'(y - X\,\hat{b})$.

Via BLP ou índice de seleção (com A genômica e $\hat{b}$ estimado via quadrados mínimos generalizados, o que é garantido quando y contem valores genéticos desregressados), em que:
$\hat{g} = A[A + R(\sigma_e^2/\sigma_g^2)]^{-1}(y - X\,\hat{b})$

.

Se necessário os efeitos dos marcadores podem ser obtidos por $\hat{m} = \{Z'/[2\sum_{i}^{n} p_i(1-p_i)]\}[A + R(\sigma_e^2/\sigma_g^2)]^{-1}(y - X\hat{b})$.

Via BLUP Modelo Equivalente, em que:

$\hat{g} = [R^{-1} + A^{-1}(\sigma_e^2/\sigma_g^2)]^{-1}R^{-1}(y - X\hat{b})$

Para implementação do procedimento RR-BLUP/GWS são necessários: X, Z, y e $\lambda = \sigma_e^2/\sigma_{gi}^2 = \sigma_e^2/(\sigma_g^2/n)$. O vetor y refere-se a fenótipos corrigidos; a matriz Z refere-se à contagem de doses dos marcadores moleculares; X é um vetor conhecido composto de valores 1; λ depende de componentes de variância (herdabilidade ou confiabilidade da seleção) e do número de segmentos cromossômicos. A seguir são descritos cada um desses elementos, os quais são tratados em Meuwissen et al. (2001), Resende (2008), Garrick et al. (2009), Gianola et al. (2009), Goddard (2009) e Hayes et al. (2009).

Fenótipos corrigidos

Os fenótipos devem ser corrigidos visando eliminar os efeitos dos genitores e desregressar os valores genéticos. Esses devem ser desregressados por três motivos: (i) não pode haver duas regressões: uma baseada em pedigree e outra baseada em marcadores; (ii) a matriz A baseada em pedgree é menos precisa que a ZZ´ baseada em marcas; (iii) influência de genes de grande efeito presentes em um dos genitores.

Adicionalmente, devem ser corrigidos para os efeitos genéticos dos genitores, trabalhando-se basicamente com o efeito da "segregação mendeliana desregressada", já que o dado ideal para a população de treinamento deve ser o "mérito genético verdadeiro de indivíduos não aparentados". E o efeito da segregação mendeliana proporciona isso: análise da associação de alelos de marcas e de QTLs, ou seja, desequilíbrio de ligação (LD) livre de genealogia. As ferramentas genômicas propiciam

uma inspeção direta da segregação mendeliana ao nível do cromossomo.

Outra forma explícita de se fazer isso, parcialmente, é a consideração do pedigree via ajuste de g*, o vetor de efeitos poligênicos. Sem a correção mencionada acima ou o ajuste de g*, os marcadores podem estar capturando apenas o parentesco entre os indivíduos e não necessariamente o desequilíbrio de ligação com os genes propriamente ditos. Nesse caso, a acurácia da validação em uma amostra independente (indivíduos de outras famílias) da população e, também, em indivíduos de outras gerações poderá ser baixa, ao contrário do que teria sido predito em uma validação em amostra dependente.

O procedimento de obtenção dos valores fenotípicos desregressados e corrigidos para os efeitos genéticos dos genitores envolve os seguintes passos:

(i) Definição do sistema de equações associado à predição do valor genético de um indivíduo i ($\hat{g}_i$) e do valor genético médio de seus genitores j e k ($\hat{g}_{gm} = (\hat{g}_j + \hat{g}_k)/2$):

$$\begin{bmatrix} Z'_{gm}Z_{gm} + 4\lambda^* & -2\lambda^* \\ -2\lambda^* & Z'_i Z_i + 2\lambda^* \end{bmatrix} \begin{bmatrix} \hat{g}_{gm} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} y_{gm} \\ y_i \end{bmatrix}$$, onde:

$\lambda^* = (1-h^2)/h^2$, em que h^2 é a herdabilidade ao nível de indivíduo.

$Z'_{gm}Z_{gm}$: conteúdo de informação associado à média dos genitores.

$Z'_i Z_i$: conteúdo de informação associado ao indivíduo (mais informações de seus descendentes ou clones).

y_{gm} *e* y_i: informação fenotípica corrigida para os efeitos fixos associada à média dos genitores e ao indivíduo, respectivamente.

(ii) Obtenção da quantidade desconhecida $Z'_{gm} Z_{gm}$ ***:***

$Z'_{gm} Z_{gm} = \lambda^* (0.5\,\alpha - 4) + 0.5\,\lambda^* (\alpha^2 + 16/\delta)^{1/2}$, em que:

$\alpha = 1/(0.5 - r^2_{gm})$

$\delta = (0.5 - r^2_{gm})/(1 - r^2_i)$

$r^2_{gm} = (r^2_{gj} + r^2_{gk})/4$: confiabilidade associada ao valor genético médio predito dos genitores j e k.

r^2_i: confiabilidade associada ao valor genético predito do indivíduo.

(iii) Obtenção da quantidade desconhecida $Z'_i Z_i$***:***

$Z'_i Z_i = \delta\, Z'_{gm} Z_{gm} + 2\lambda^* (2\delta - 1)$

(iv) Obtenção da quantidade desconhecida y_i***:***

Resolução para y_i, do sistema

$$\begin{bmatrix} Z'_{gm} Z_{gm} + 4\lambda^* & -2\lambda^* \\ -2\lambda^* & Z'_i Z_i + 2\lambda^* \end{bmatrix} \begin{bmatrix} \hat{g}_{gm} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} y_{gm} \\ y_i \end{bmatrix}.$$

Assim, $y_i = (-2\lambda^*)\hat{g}_{gm} + (Z'_i Z_i + 2\lambda^*)\hat{g}_i$, o qual representa a informação do indivíduo, agora corrigida para o valor genético médio de seus genitores.

(v) Obtenção do valor genético desregressado $\hat{g}^*_i$***:***

$\hat{g}^*_i = y_i / (Z'_i Z_i)$.

Assim, para obtenção de $\hat{g}^*_i$, necessita-se da herdabilidade h^2, das confiabilidades (quadrado da acurácia) das avaliações dos três indivíduos (r^2_{gj}, r^2_{gk} e r^2_i) e dos efeitos genéticos preditos dos três indivíduos ($\hat{g}_i$, $\hat{g}_j$ e $\hat{g}_k$).

Considere um caráter com h^2 de 0,20 e a avaliação genética de três indivíduos onde foram obtidos os seguintes resultados: $\hat{g}_i$ = 18, $\hat{g}_j = 13$ e $\hat{g}_k$ = 5; r_i^2 = 0,70; r_{gj}^2 = 0,90 e r_{gk}^2 = 0,80. Assim, são obtidos:

$$r_{gm}^2 = (r_{gj}^2 + r_{gk}^2)/4 = (0,90 + 0,80)/4 = 0,425;$$

$$\hat{g}_{gm} = (\hat{g}_j + \hat{g}_k)/2 = (13 + 5)/2 = 9;$$

$$\lambda^* = (1 - h^2)/h^2 = 0,8/0,2 = 4;$$

$$\alpha = 1/(0,5 - r_{gm}^2) = 1/(0,5 - 0,425) = 13,3333;$$

$$\delta = (0,5 - r_{gm}^2)(1 - r_i^2) = (0,5 - 0,425)/(1 - 0,70) = 0,25.$$

Com base nesses valores e seguindo o passo (ii), calcula-se $Z_{gm}^{'} Z_{gm}$:

$$Z_{gm}^{'} Z_{gm} = \lambda^*(0,5\alpha - 4) + 0,5\lambda^*(\alpha^2 + 16/\delta)^{1/2} = 4(0,5 \ \ 13,3333 - 4) + 0,5 \ \ 4(13,3333^2 + 16/0,25)^{1/2} = 41,765$$

A seguir calcula-se o $Z_i^{'} Z_i$ seguindo o passo (iii):

$$Z_i^{'} Z_i = \delta \ Z_{gm}^{'} Z_{gm} + 2\lambda^*(2\delta - 1) = 0,25 \ 41,765 + 2 \ \ 4(2 \ \ 0,25 - 1) = 6,4412.$$

Computa-se agora, seguindo o passo (iv), a quantidade

$$y_i = (-2\lambda^*)\hat{g}_{gm} + (Z_i^{'} Z_i + 2\lambda^*)\hat{g}_i = (-2 \ \ 4) \ 9 + (6,4412 + 2 \ \ 4) \ 18 = 187,9423$$

E finalmente calcula-se o valor corrigido e desregressado, seguindo o passo (v):

$$\hat{g}_i^* = y_i/(Z_i^{'} Z_i) = 187,9423/6,4412 = 29,1780$$

Esse é o valor do indivíduo a ser usado na análise genômica integrando o vetor y. Tal quantidade é equivalente a $\hat{g}_i^* = (\hat{g}_{i-gm})/r_i^{2*}$, ou seja, ao valor genético individual corrigido para a média de seus genitores e desregressado pela quantidade $r_i^{2*} = 1 - \lambda^*/(Z_i^{'} Z_i + \lambda^*) = 1 - 4^*/(6,4412 + 4) = 0,6169$.

Em caso de testes de progênie em uma só geração, o valor individual corrigido para o valor genético médio de seus genitores e desregressado são dados pela expressão $\hat{g}_i^* = (y - X\hat{b} - W\hat{c} - 0,5\ \hat{g}_j - 0,5\ \hat{g}_k)$, em que $\hat{b}$ e $\hat{c}$ são os efeitos estimados de blocos e de parcelas.

Apenas desregressar por r_i^2 captura LD e parentesco. Seria necessário ajustar o efeito poligênico para remover a estruturação devida ao parentesco. Regressar por r_i^{2*} e corrigir para efeito dos genitores captura apenas LD, eliminando a correlação intraclasse entre os valores genéticos preditos. Por esse motivo, o valor genético genômico dos indivíduos na população de validação (visando ao cômputo da acurácia) são dados por $u + \hat{g}_i = u + Z\hat{m}^*$. Não se deve somar $\hat{g}_{gm}$. Por outro lado, na população de estimação, visando à seleção, deve-se computar $u + \hat{g}_i = u + \hat{g}_{gm} + Z\hat{m}^*$ ou fazer a predição de m usando os valores genéticos desregressados, mas não corrigidos para os efeitos dos genitores e usar diretamente $u + \hat{g}_i = u + Z\hat{m}$. Na população de seleção propriamente dita (onde apenas os genótipos dos marcadores estão disponíveis), a seleção precoce deve basear-se diretamente em $u + \hat{g}_i = u + Z\hat{m}$, mas a acurácia da seleção é calculada com base em $u + \hat{g}_i = u + Z\hat{m}^*$, em que $\hat{m}^*$ é o vetor de efeitos preditos dos marcadores, obtido via $\hat{g}_i^*$.

Frequências alélicas, variância dos marcadores e herdabilidade

Marcadores codominantes (SNP) – Modelo genotípico

Variâncias e padronizações

Nesse caso, a matriz de incidência *Z* contém os valores 0, 1 e 2 para o número de alelos do marcador (ou do suposto QTL) em um indivíduo diplóide.

Com marcadores codominantes, a média e variância da variável Z associada à matriz de incidência são dadas por:

Média da variável $Z = 0 \times p^2 + 1 \times 2p(1-p) + 2 \times (1-p)^2 = 2p$

Variância da variável $Z = Var(Z) = Var(Z_i) = (0 - 2p)^2 \times p^2 + (1 - 2p)^2 \times 2p(1-p) + (2 - 2p)^2 \times (1-p)^2 = 2p(1-p)$

Verifica-se que a variância da variável Z equivale ao grau de heterozigose ou frequência de heterozigotos na população. A raiz quadrada dessa variância pode ser usada para padronizar os dados dos marcadores na matriz Z, da seguinte forma para cada elemento Z_i da matriz, referente ao loco i:

$Z_i = 0$ se o indivíduo é homozigoto para o primeiro alelo (mm);

$Z_i = 1/(Var(Z_i))^{1/2}$ se o indivíduo é heterozigoto (Mm);

$Z_i = 2/(Var(Z_i))^{1/2}$ se o indivíduo é homozigoto para o segundo alelo no loco marcador (MM);

$Z_i = 2p_i /(Var(Z_i))^{1/2}$ se o genótipo do marcador é um dado perdido, em que $2p_i$ vem do valor esperado $E(Z_i = 2 \text{ ou } Z_i = 1) = 2 \times p_i^2 + 1 \times 2p_i(1-p_i) = 2p_i$.

A quantidade p_i é a frequência do segundo alelo do marcador.

Dessa forma, a variância de Z com Z_i ajustado é igual a 1. Alternativamente, os numeradores de Z_i podem ser subtraídos pela média de Z (via 0-2p, 1-2p e 2-2p, respectivamente) obtendo-se uma variável com média zero e variância unitária.

Sendo m o efeito do marcador na população, a variância devida ao marcador é dada por $Var(Z_i m) = Var(Z_i)\ Var(m)$. Com a transformação acima, $Var(Z_i) = 1$ e portanto, $Var(Z_i m) = Var(m)$. Em outras palavras, modelando a variância do efeito do marcador, modela-se diretamente a variância do marcador, independentemente de sua frequência. Mas, a padronização não é estritamente necessária.

Relação entre variância genética e variância dos marcadores

A relação entre variância genética aditiva e variância dos efeitos dos marcadores é essencial na predição genômica. Do exposto acima, segue que $Var(g_i) = Var(Z_i m) = Var(Z_i)\ Var(m) = 2p_i(1-p_i)\ Var(m_i) = 2p_i(1-p_i)\ m_i^2$ equivale à variância genética devida ao loco i. Para vários locos, a variância genética aditiva total é dada por $\sigma_g^2 = \sum_i^n 2\,p_i(1-p_i)\,m_i^2$, a qual pode ser expressa também por

$\sigma_g^2 = \sum_i^n U_i V_i$, em que $U_i = 2\,p_i(1-p_i)$ e $V_i = m_i^2$

A covariância entre U e V, denominada C_{UV} é dada por

$C_{UV} = (\sum_i^n U_i V_i)/n - (\sum_i^n U_i/n)(\sum_i^n V_i)/n)$.

Rearranjando essa expressão tem-se

$\sum_i^n U_i V_i = n\,C_{UV} + (\sum_i^n U_i)(\sum_i^n V_i)/n)$, de forma que

$\sigma_g^2 = \sum_i^n U_i V_i = n\,C_{UV} + [\sum_i^n 2\,p_i(1-p_i)](\sum_i^n m_i^2)/n$.

Sendo $(\sum_i^n m_i^2)/n = \sigma_m^2$, tem-se

$\sigma_g^2 = [2\sum_i^n p_i(1-p_i)\sigma_m^2] + n\,C_{UV}$.

Assim, a variância entre marcadores (σ_m^2) obtida por REML, as frequências alélicas e os efeitos dos marcadores preditos por BLUP podem ser usados na obtenção da variância genética aditiva total. Desse modo, a variância genética entre marcas, estimada pelo software Selegen Genômica-REML/BLUP/GWS, deve ser multiplicada pelo número de marcas e por 2p(1-p) e também acrescida de nC_{UV}, para cômputo da variância genética aditiva total.

Em alguns casos, C_{UV} assume o valor zero (quando a média dos m equivale a zero) ou muito baixo. Em outros casos, a

quantidade m_i^2 é substituída por σ_m^2, pois a esperança de m_i^2 é a variância do efeito do marcador, ou seja, $E(m_i^2) = \sigma_m^2$. Assim, muitas das aplicações usam $\sigma_g^2 = [2\sum_i^n p_i(1-p_i)\sigma_m^2]$ e a variância entre marcadores dada por

$\sigma_m^2 = (\sigma_g^2 - n\,C_{UV}) / [2\sum_i^n p_i(1-p_i)]$ é simplificada para $\sigma_m^2 = \sigma_g^2 / [2\sum_i^n p_i(1-p_i)]$. Um exemplo completo é apresentado no tópico Exemplo Aplicado no Melhoramento do Eucalipto.

Matriz de parentesco genômico

Sem padronização prévia dos elementos de Z, tem-se

$$A = (Z^* Z^{*\prime}) / [2\sum_i^n p_i(1-p_i)]$$

em que p_i é a frequência de um dos alelos do loco i e $Z^* = Z - P$ onde P é uma matriz com elementos $2p_i$ na coluna i. Com padronização prévia dos elementos de Z e centrando a média em zero tem-se $A = ZZ'$.

Marcadores dominantes (DArT) - Modelo genotípico

Variâncias e padronizações

Nesse caso, a matriz de incidência *Z* contém os valores 0 ou 1 para ausência ou presença de um dos alelos do marcador (ou do suposto QTL) em um indivíduo diplóide.

Com marcadores dominantes, a média e variância da variável Z associada à matriz de incidência são dadas por:

Média da variável Z: 1 x p + 0 x (1-p) = p, em que p é a frequência do código 1, que contempla MM e Mm de forma confundida.

Var (Z) = Var (Z_i) = Variância da variável Z: $(1 - p)^2$ x p + $(0 - p)^2$ x (1-p) = p(1-p)

Assim, a variável Z tem distribuição Bernoulli com média p e variância p(1-p). A raiz quadrada dessa variância pode ser usada para padronizar os dados dos marcadores na matriz Z, da seguinte forma:

Z_i = 0 se a banda é ausente no indivíduo.

$Z_i = 1/(Var(Z_i))^{1/2}$ se a banda está presente no indivíduo.

$Z_i = p_i /(Var(Z_i))^{1/2}$ se o genótipo do marcador é um dado perdido.

A quantidade p_i é a frequência do código 1.

Relação entre variância genética e variância dos marcadores

No caso de marcadores dominantes, tem-se que $Var(g_i) = Var(Z_i m) = Var(Z_i) Var(m) = p_i(1-p_i) Var(m_i) = p_i(1-p_i) m_i^2$, que equivale à variância genética devida ao loco i. Para vários locos $\sigma_g^2 = [\sum_i^n p_i (1-p_i) m_i^2]$.

Expressa diretamente em função da variância dos efeitos dos marcadores (σ_m^2) tem-se que $\sigma_g^2 = [\sum_i^n p_i (1-p_i) \sigma_m^2] + n C_{UV}$. Ignorando nC_{UV}, tem-se que

$$\sigma_m^2 = \sigma_g^2 / [\sum_i^n p_i (1-p_i)].$$

A quantidade C_{UV} é dada por

$$C_{UV} = (\sum_i^n U_i V_i)/n - (\sum_i^n U_i / n)(\sum_i^n V_i)/n),$$

em que $U_i = p_i (1-p_i)$ e $V_i = m_i^2$. Assim, a variância entre marcadores (σ_m^2) obtida por REML, as frequências alélicas e os efeitos dos marcadores preditos por BLUP podem ser usados na obtenção da variância genética aditiva total. Assim, a variância genética entre marcas, estimada por REML, deve ser multiplicada pelo número de marcas e por p(1-p) e também acrescida de nCUV, para cômputo da variância genética aditiva total. Isso

pode ser feito por meio do software Selegen Genômica RR-BLUP, conforme apresentado no tópico Exemplo Aplicado ao Melhoramento do Eucalipto.

Matriz de parentesco genômico

Sem padronização prévia dos elementos de Z, tem-se

$$A = (Z^* Z^{*\prime}) / [\sum_{i}^{n} p_i(1-p_i)]$$

em que p_i é a frequência associada à presença de marca no loco i e $Z^* = Z - P$ onde P é uma matriz com elementos p_i na coluna i. Com padronização prévia dos elementos de Z e centrando a média em zero tem-se $A = Z Z'$.

Marcadores codominantes (SNP) – Modelo gamético ou alélico

Nesse caso, a matriz de incidência *Z* contém os valores 0 ou 1 para ausência ou presença de um dos alelos do marcador em uma célula haplóide.

A média e variância da variável Z associada à matriz de incidência são dadas por:

Média da variável Z = 1 x p + 0 x (1-p) = p, em que p é a frequência do código 1, que contempla M.

Variância da variável Z = Var (Z) = Var (Z_i) = $(1 - p)^2$ x p + $(0 - p)^2$ x (1-p) = p(1-p)

Assim, a variável Z tem distribuição Bernoulli com média p e variância p(1-p).

Relação entre variância genética e variância dos marcadores

Do exposto anteriormente e computando o efeito do alelo duas vezes para se ter o g de um indivíduo diplóide, segue que Var(g_i)

= 2Var(Z_i2m) = 2 Var(Z_i) Var(2m) = 2[p_i(1-p_i)] Var(2m_i) = [2p_i(1-p_i)] 4 m_i^2 equivale à variância genética devida ao loco i. Para vários locos $\sigma_g^2 = 4[2\sum_i^n p_i(1-p_i)m_i^2]$.

Expressa diretamente em função da variância dos efeitos dos marcadores (σ_m^2), tem-se que $\sigma_g^2 = 4[2\sum_i^n p_i(1-p_i)\sigma_m^2]$

Portanto, $\sigma_m^2 = \sigma_g^2 / \{4[2\sum_i^n p_i(1-p_i)]\} = \sigma_g^2 / (4n\overline{H})$,

em que $\overline{H} = (1/n)[2\sum_i^n p_i(1-p_i)]$ é a heterozigose média dos marcadores. Com freqüência alélica p = 0.5 em todos os locos marcadores, tem-se que $\sigma_m^2 = \sigma_g^2 / (2n)$. Sob modelo gamético, a quantidade 2n advém do fato que cada marcador afeta o fenótipo duas vezes, via alelo de origem paterna e materna.

Matriz de parentesco genômico

Sem padronização prévia dos elementos de Z, tem-se

$A = (Z^* Z^{*\prime}) / [\sum_i^n p_i(1-p_i)]$

em que p_i é a frequência de um dos alelos do loco i e $Z^* = Z - P$ onde P é uma matriz com elementos p_i na coluna i. Com padronização prévia dos elementos de Z e centrando a média em zero tem-se $A = ZZ'$.

Herdabilidade

A variância genética e a herdabilidade (h^2) podem ser computadas via dados fenotípicos ou via dados de marcadores e fenotípicos conforme descrito anteriormente no cômputo de σ_g^2. A h^2 a ser usada no RR-BLUP deve ser a herdabilidade ajustada ou dos dados corrigidos ($h_{aj}^2 = \sigma_g^2 / \sigma_{yaj}^2$), em que σ_{yaj}^2 é a variância fenotípica ajustada. Se y é corrigido para a média dos genitores o numerador de h_{aj}^2 deve conter apenas a variância genética devida à segregação mendeliana, ou seja, $h_{aj}^{2*} = (1/2)\sigma_g^2 / \sigma_{yaj}^2$ ou $h_{aj}^{2*} = (3/4)\sigma_g^2 / \sigma_{yaj}^2$ quando se conhece

os dois genitores (famílias de irmãos germanos) ou apenas um dos genitores (famílias de meios irmãos), respectivamente. Essas herdabilidades podem ser expressas também em função da herdabilidade individual h^2, por meio das expressões $h_{aj}^{2*} = (1/2\ h^2)/(1/2\ h^2 + (1-h^2))$ para progênies de irmãos germanos e $h_{aj}^{2*} = (3/4\ h^2)/(3/4\ h^2 + (1-h^2))$ para progênies de meios-irmãos. Essas fórmulas mostram que o denominador de h_{aj}^{2*} também contempla apenas a variância genética devida à segregação mendeliana e não a variância genética total. Outra forma de expressar h_{aj}^{2*} é usar diretamente a confiabilidade r_i^{2*}, apresentada no quarto tópico. Para cômputo do RR-BLUP e da acurácia da GWS, h_{aj}^{2*} pode ser tomada como a média dos r_i^{2*} dos indivíduos em análise.

Número de marcadores com efeitos significativos

Na predição RR-BLUP/GWS, necessita-se da quantidade $\lambda = \sigma_e^2 / \sigma_{gi}^2 = \sigma_e^2 / (\sigma_g^2 / n)$, em que n é o número de locos controlando o caráter (assumindo que cada loco está perfeitamente marcado), o qual é desconhecido a priori.

A variância genética contribuída por cada loco é dada por $\sigma_{gi}^2 = 2\,p_i\,(1-p_i)\,a_i^2$ em que p_i é a frequência de um dos alelos do loco i e a_i é o efeito de substituição alélica (FALCONER, 1989). A variância genética total do caráter é dada pelo somatório das variâncias nos locos individuais, ou seja, $\sigma_g^2 = [2\sum_i^n p_i\,(1-p_i)\,a_i^2]$ em que o somatório estende para todos os n locos. Com variâncias de magnitudes iguais em todos os locos, tem-se $\sigma_g^2 = n\,\sigma_{gi}^2$. Conforme tópico anterior, na GWS, σ_{gi}^2 é dada aproximadamente por $\sigma_{gi}^2 = \sigma_g^2 / [2\sum_i^n p_i\,(1-p_i)]$

em que o somatório estende para todos os n locos marcadores codominantes e p_i refere-se à frequência de um dos alelos de cada loco marcador, considerando todos os locos marcadores

ajustados no modelo. A quantidade $V(Z_i) = 2\,p_i\,(1-p_i)$ é a variância da variável de incidência Z no loco marcador i.

Assim, na expressão $\lambda = \sigma_e^2 \,/\, \sigma_{gi}^2 = \sigma_e^2 \,/(\sigma_g^2/n)$, n pode ser tomado como $[\,2\sum_i^n p_i\,(1-p_i)\,]$. Alternativamente, λ pode ser expresso como $\lambda = \sigma_e^2 \,/\, \sigma_{gi}^2 = \sigma_e^2 \,/(\sigma_g^2/n) = (1-h^2)/(h^2/n)$

e, portanto $\lambda = (1-h^2)/(h^2/n) = (1-h^2)/\{(h^2/[\,2\sum_i^n p_i\,(1-p_i)\,]\}$.

Expresso de outra forma,

$\lambda = n(1-h^2)/h^2 = [\,2\sum_i^n p_i\,(1-p_i)\,](1-h^2)/h^2$

Assim, de posse de h^2 e das frequências alélicas nos locos marcadores, obtém-se λ para uso nas equações de modelo misto. É importante notar que h^2 refere-se à herdabilidade ajustada ou, em alguns casos, à confiabilidade ($r_{\hat{g}g}^2$) da predição.

Recomenda-se então analisar inicialmente todo o conjunto de marcadores codominantes em todos os indivíduos fenotipados (população de estimação completa), usando n como o número total de marcadores ponderados por $V(Z_i) = 2\,p_i\,(1-p_i)$ ou simplesmente usar n como o número total de marcadores. Esse procedimento visa identificar os marcadores com maiores efeitos em módulo, objetivando rodar análises com subgrupos menores de marcadores e determinar quantos e quais marcadores maximizam a acurácia seletiva. O número ótimo de marcadores é um compromisso entre maior informatividade (maior acurácia, pela maior captura de genes) e menor precisão (menor acurácia, pelo menor tamanho amostral por efeito estimado) com o aumento do número de marcadores. Posteriormente, a validação deve ser realizada usando apenas a fração de marcadores que maximiza a acurácia, usando n como o somatório $[\,2\sum_i^n p_i\,(1-p_i)\,]$ nesse subconjunto de marcadores. Esse procedimento é recomendável, pois tende a produzir

acurácia mais alta, similar à obtida pelo método Bayes B. Dessa forma, ambos os métodos assumem que muitos dos marcadores apresentam efeitos zero. Isto o faz também o método do aprendizado de máquina (AM).

Outra abordagem para inferir sobre n é usar o seu valor esperado, dado o tamanho efetivo (Ne) da população e o tamanho L do genoma da espécie. Com base no tamanho efetivo populacional (Ne), pode-se calcular o número efetivo de locos ou segmentos cromossômicos (Me) devidos à ligação (segundo esse conceito, para dois gametas quaisquer, o genoma é quebrado em Me segmentos de tamanho igual). Nesse caso, n é dado por $n = Me\ V(q) = Me\ k$, sendo V(q) a heterozigose média de todos os segmentos cromossômicos independentes, ou seja, $V(q) = 2p(1-p)$, em que p é a frequência alélica média. V(q) é análogo a $V(Z_i)$, sendo que q refere-se aos locos gênicos e Z refere-se aos locos marcadores.

Segundo Goddard (2008), e conforme apresentado por Resende (2008), a quantidade Me é dada por $Me = (2NeL)/[Ln(4NeL)]$, em que L é o tamanho total do genoma em Morgans. Entretanto, Hayes et al. (2009) relatam que o valor mais apropriado para Me situa-se entre 4NeL e $(2NeL)/[Ln(4NeL)]$, sendo uma boa aproximação usar $Me = 2NeL$, ou seja, assumir o número efetivo de locos como 2NeL. Esse número efetivo de locos deve ser ponderado por uma função da frequência alélica do gene (via frequência do marcador), que está implícita em V(q). O valor de n é dado então por $n = Me\ V(q) = Me\ k$, em que $V(q) = k$ é dado por $k = 1/[Ln(2Ne)]$. Dessa forma, $n = 2NeL\ 1/[Ln(2Ne)]$. A quantidade Me V(q) refere-se ao número esperado de marcas com efeitos significativos. Isso é confirmado pelos resultados práticos com eucalipto, associado ao $Ne = 100$ e $L = 13,2$ na Tabela 2, em que consta $n = 502$, o qual é coerente com o número de marcadores (500 a 750) que maximiza a acurácia com a GWS.

Para a predição BLUP, as alternativas que podem ser adotadas visando inferir sobre n são:

(i) Ajustar os efeitos de cada SNP individualmente, avaliando suas significâncias e, posteriormente, ajustar simultaneamente todos os SNPs (locos) com efeitos significativos, usando n como o número de SNPs significativos. Idealmente, esse n deve ser ponderado por uma função da frequência alélica do marcador.

(ii) Computar n via n = Me V(q) = Me k.

(iii) Usar todos os marcadores, sem teste de significância e, computando $n = [2\sum_{i}^{n} p_i(1-p_i)]$.

Geralmente o número de SNPs significativos é maior do que o número de locos pois cada SNP rastreia um grande segmento cromossômico e então o efeito de cada segmento cromossômico é dividido em muitos SNPs. Em gado de leite, o número de SNPs com efeitos significativos variou de 3 mil a 4 mil entre caracteres, dentre cerca de 40 mil marcadores usados (HAYES et al., 2009).

A melhor opção é adotar a estratégia (iii) seguida de escolha de subconjuntos menores de marcadores, com base no módulo dos maiores efeitos de todos os marcadores estimados inicialmente. Esse ponto distingue a GWS da GWAS (*Genome Wide Association Studies*), a qual procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses, visando detectar efeitos com significância estatística. A GWAS sofre com a alta taxa de falsos negativos devido ao uso de pontos de corte muito rigorosos, visando evitar a ocorrência de falsos positivos. A GWS equivale à GWAS tradicional aplicada sobre todos os locos simultaneamente e baseando-se em estimação e predição em vez de teste de hipótese. Dessa forma, consegue explicar parte muito maior da variabilidade genética e evitar a chamada herdabilidade faltante ou perdida (*missing*

heritability, conforme Maher (2008)), típica dos estudos de análise de ligação e de associação.

O número máximo de SNPs com efeitos significativos é limitado pelo Ne. Com Ne mais baixo, menor é n. O número real de segmentos cromossômicos total é 4NeL, ou seja, 120 mil em bovinos, que é bem maior que o número efetivo de segmentos. Em eucalipto (genoma com 13,2 Morgans), tem-se a Tabela 2.

O valor de n computado via $n = Me\ V(q) = Me\ k$, pode ser usado em estudos de simulação da acurácia seletiva, visando inferir sobre o número de locos ou segmentos a compartilhar a variância genética total. Esse número, em uma população de cruzamentos ao acaso, depende apenas do Ne e do tamanho total do genoma da espécie. Na Tabela 2 são apresentados valores de n para bovinos (genoma com $L = 30$ Morgans) e eucalipto (genoma com $L = 13,2$ Morgans), para diferentes valores de Ne.

Tabela 2. Número efetivo de segmentos cromossômicos (Me) e de locos (n) em função do tamanho efetivo (Ne) e do comprimento do genoma (L) em bovinos e eucalipto.

Ne	Ltot	Me = 2NeL	k = 1/*Ln*(2Ne)	n = Me k
Bovinos				
15	30	900	0,29	261
30	30	1.800	0,24	432
50	30	3.000	0,22	660
100	30	6.000	0,19	1.140
200	30	12.000	0,17	2.040
500	30	30.000	0,14	4.200
1.000	30	60.000	0,13	7.800
Eucalipto				
15	13,2	396	0,29	115
30	13,2	792	0,24	190
50	13,2	1.320	0,22	290
100	13,2	2.640	0,19	502
200	13,2	5.280	0,17	898
500	13,2	13.200	0,14	1.848
1.000	13.2	26.400	0.13	3.432

Cômputo da acurácia esperada

Resende (2008) e Resende et al. (2008) apresentaram uma abordagem para cômputo da acurácia esperada com a GWS, a qual foi empregada por Grattapaglia e Resende (2010). A acurácia esperada é dada por

$$r_{q\hat{q}} = (r^2_{m\hat{m}}\, r^2_{mq})^{1/2} = \sqrt{(N\, h^2_m\, r^2_{mq})/[1+(N-1)\, h^2_m]}$$

com $h^2_m = (h^2_{aj}\, r^2_{mq}/n)/(h^2_{aj}\, r^2_{mq} + (1-h^2_{aj}))$.

A magnitude do desequilíbrio de ligação é quantificada por (SVED, 1971):

$$r^2_{mq} = E(r^2) = \frac{1}{4\,N_e\,L+1}$$

Valores de r^2_{mq} (proporção da variação do gene explicada pelo marcador) para diferentes valores de Ne e espaçamento entre marcadores são apresentados na Tabela 3. Verifica-se que para uma população com Ne igual a 100, são necessários dez marcadores por cM para se conseguir um r^2_{mq} de 0,71. Com esse r^2_{mq} e um caráter controlado por 100 locos e com herdabilidade individual de 0,30, avaliando-se N = 1.000 indivíduos na população de estimação, a acurácia esperada é de 0,70 (RESENDE, 2008), valor esse muito interessante do ponto de vista prático.

Daetwyler et al. (2008) assumiram $\sigma^2_e = 1$ e $r^2_{mq} = 1$, obtendo

$$r_{m\hat{m}} = \sqrt{(N h^2_{aj}/n)/[1+(N h^2_{aj}/n)]} = \sqrt{(\omega h^2_{aj})/[1+(\omega\, h^2_{aj})]}$$

e mostrando a importância da quantidade $\omega = N/n$, a qual equivale ao número de indivíduos N usados para estimar o efeito de cada loco na população de estimação. Resende (2008) obteve uma expressão mais geral, não assumindo $\sigma^2_e = 1$ e $r^2_{mq} = 1$, ou seja, mantendo esses dois elementos na fórmula.

Tabela 3. Valores de r^2_{mq} para diferentes valores de tamanho efetivo (Ne) e espaçamento entre marcadores (Lseg em Morgans), segundo a equação de Sved (1971).

Ne	L seg	r^2_{mq}	L seg	r^2_{mq}
15	0,01	0,63	0,001	0,94
30	0,01	0,45	0,001	0,89
50	0,01	0,33	0,001	0,83
100	0,01	0,20	0,001	0,71
200	0,01	0,11	0,001	0,56
500	0,01	0,05	0,001	0,33
1.000	0,01	0,02	0,001	0,20
15	0,005	0,77	0,0005	0,97
30	0,005	0,63	0,0005	0,94
50	0,005	0,50	0,0005	0,91
100	0,005	0,33	0,0005	0,83
200	0,005	0,20	0,0005	0,71
500	0,005	0,09	0,0005	0,50
1.000	0,005	0,05	0,0005	0,33
15	0,002	0,89	-	-
30	0,002	0,81	-	-
50	0,002	0,71	-	-
100	0,002	0,56	-	-
200	0,002	0,38	-	-
500	0,002	0,20	-	-
1.000	0,002	0,11	-	-

Em resumo, a acurácia da GWS depende de cinco fatores: (i) da herdabilidade do caráter; (ii) do número de locos controlando o caráter e da distribuição de seus efeitos; (iii) do número de indivíduos na população de descoberta; (iv) do tamanho efetivo populacional; (v) do espaçamento entre marcadores, o qual depende do seu número e do tamanho do genoma.

Esses dois últimos fatores determinam r^2_{mq}. Os dois primeiros fatores não estão sobre o controle do melhorista. Os três últimos podem ser modificados pelo melhorista visando aumentar a acurácia da GWS. Valores de acurácia esperada para várias situações foram tabelados por Resende (2008). Na Tabela 4 são apresentados resultados da acurácia seletiva da GWS para um caráter controlado por 100 locos e com herdabilidade individual no sentido restrito igual a 0,30. Verifica-se que, para uma população de eucalipto com tamanho efetivo 100 (r^2_{mq} = 0,7, conforme a Tabela 3), a acurácia seletiva esperada com a GWS é de 0,79, para um tamanho amostral de N = 4.000 indivíduos. Esse valor supera a acurácia máxima (0,70) para a seleção de indivíduos pelo BLUP tradicional na idade adulta. Isto atesta o grande potencial da GWS.

Tabela 4. Aumento da acurácia da GWS em função do aumento do tamanho da população de estimação. Caráter controlado por 100 locos e com herdabilidade individual no sentido restrito igual a 0,30.

Número de indivíduos	r^2_{mq} = 0,1	r^2_{mq} = 0,3	r^2_{mq} = 0,5	r^2_{mq} = 0,7	r^2_{mq} = 0,9
100	0,06	0,18	0,27	0,36	0,44
200	0,09	0,24	0,36	0,47	0,57
500	0,13	0,33	0,48	0,61	0,72
1.000	0,17	0,40	0,57	**0,70**	0,81
2.000	0,21	0,46	0,62	**0,76**	0,87
4.000	0,25	0,50	0,66	**0,79**	0,91
8.000	0,28	0,52	0,68	0,81	0,93

*Acurácia máxima para a seleção de indivíduos pelo BLUP tradicional na idade adulta = 0,70.

Ganhos adicionais podem ser conseguidos por unidade de tempo, conforme a Tabela 5. Verifica-se que ganho da ordem de 126% pode ser conseguido com a redução, de 4 para 2 anos do tempo necessário para completar um ciclo de seleção.

Tabela 5. Eficiência da GWS por unidade de tempo.

Acurácia fenotípica (AF)	Acurácia genômica (AG)	Tempo fenotípica (TF)	Tempo genômica (TG)	Eficiência (AG TF)/ (AF TG)	Superioridade %
0,70	0,79	4	4	1,13	13
0,70	0,79	4	3	1,50	50
0,70	0,79	4	2	2,26	126
0,70	0,79	4	1	4,51	351
0,70	0,79	4	0,5	9,03	803

Goddard et al. (2009) e Hayes et al. (2009) apresentaram outra abordagem para cômputo da acurácia esperada na estimação do efeito de um marcador com a GWS. Nessa abordagem, a acurácia é dada por:

$$r_{m_i \hat{m}_i} = \sqrt{(NV(Z_i)/[NV(Z_i)+\lambda]} = \sqrt{(NV(Z_i)/[NV(Z_i)+(\sigma_e^2/\sigma_m^2)]}$$

em que:

$$\lambda = \sigma_e^2 / \sigma_m^2$$

$$\sigma_m^2 = \sigma_g^2/[M\,e\,V(q)] = \sigma_g^2/(M\,e\,k)$$

Assumindo que a variância residual apresenta magnitude próxima da variância fenotípica, tem-se $\lambda = (M\,e\,k)/h^2$. Assim, de posse do Ne, L, h^2 e N, tem-se a acurácia esperada na estimação do efeito de cada marcador.

A confiabilidade da seleção baseada na soma dos efeitos de todos os marcadores equivale à média ponderada das confiabilidades dos efeitos preditos de cada marcador. Os fatores de ponderação são os $V(Z_i)$, ou seja, a heterozigose (variância) de cada marcador.

Assim $V(Z_i)$ participa duas vezes da expressão da acurácia, em $r_{m\hat{m}}$ e na ponderação. Quanto maior a heterozigose, maior a confiabilidade e maior o peso e, portanto, maior

a acurácia. A confiabilidade global da predição é dada por $r^2_{m\hat{m}} = [1-\lambda/(2N\alpha^{1/2})*\log((1+\alpha+2\alpha^{1/2})/(1+\alpha-2\alpha^{1/2}))]$, em que $\alpha = 1+2\lambda/N$

Isso assume r^2_{mq} = 1. A quantidade r^2_{mq} depende do r² (medida tradicional de LD) de cada marcador ponderados por uma função das frequências alélicas e variâncias devidas aos locos gênicos. De posse de r^2_{mq}, obtém-se a acurácia seletiva dada por $r_{q\hat{q}} = (r^2_{m\hat{m}}\, r^2_{mq})^{1/2}$.

O r^2_{mq} é uma média ponderada do r² de cada par marcador-QTL. O r² é o quadrado da correlação entre alelos ou genótipos presentes no loco marcador e no loco do QTL (Tabela 6). As duas formas de cálculo do r² produzem resultados iguais se os genitores femininos e masculinos são cruzados aleatoriamente. O método baseado na correlação entre genótipos é mais fácil computacionalmente, pois não requer haplotipagem.

Tabela 6. Cálculo do desequilíbrio de ligação entre marcador e QTL.

Indivíduo	Genitor	Genótipos loco marcador	Genótipos loco QTL	Núm. alelos loco marcador	Núm. alelos loco QTL
1	Feminino Masculino	0 0	0 0	0	0
2	Feminino Masculino	1 1	0 1	2	1
3	Feminino Masculino	1 0	1 0	1	1
4	Feminino Masculino	0 1	0 0	1	0
5	Feminino Masculino	1 1	1 0	2	1
Correlação r	r^2	r = 0,53	r^2 = 0,29	r = 0,76	r^2 = 0,58

A relação entre efeitos genéticos do marcador e do QTL pode ser melhor entendida segundo os modelos abaixo.

Modelo para fenótipo via efeito genético do QTL (g_{QTL})

$$y = u + g_{QTL} + e$$

Modelo para fenótipo via efeito genético do marcador (g_m)

$$y = u + g_{QTL} + e = u + Z g_m + e$$

A quantidade g_m é uma regressão dada por

$$g_m = Cov(y,Z)/Var(Z) = Cov(g_{QTL},Z)/Var(Z)$$

$$= r[Var(g_{QTL})/Var(Z)]^{1/2} = r\{Var(g_{QTL})/[2p(1-p)]\}^{1/2}$$

A quantidade da variação no QTL explicada pelo marcador é dada por:

$$Var(Z g_m) = g_m^2 Var(Z) = r^2 [Var(g_{QTL})/Var(Z)] Var(Z) = r^2 Var(g_{QTL})$$

Assim, surge o conceito de r^2 como a proporção da variação do QTL explicada pelo marcador.

Populações de estimação, validação e seleção

Na prática da seleção genômica ampla, três populações podem ser definidas: população de estimação, validação e seleção. Essas podem: (i) ser fisicamente distintas (três populações diferentes); (ii) exercer duas funções ao mesmo tempo (uma só população usada para estimação e validação); (iii) exercer três funções ao mesmo tempo (uma só população usada para estimação, validação e seleção). Em geral, as estratégias (i) e (ii) são mais usadas. A Figura 1 ilustra a estratégia (ii).

População de Estimação. Também denominada população de descoberta, de treinamento ou de referência. Esse conjunto de dados contempla um grande número de marcadores avaliados em um número moderado de indivíduos (1.000 a 2.000,

dependendo da acurácia desejada, conforme relatado no sexto tópico), os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Equações de predição (regressão múltipla aleatória) de valores genéticos genômicos são obtidas para cada caráter de interesse. Essas equações associam a cada marcador ou intervalo o seu efeito (predito por RR-BLUP) no caráter de interesse. Nessa população são descobertos, via marcadores, os marcadores que explicam os locos que controlam os caracteres, bem como são estimados os seus efeitos.

População de Validação. Quando fisicamente disjunta da população de estimação, esse conjunto de dados é menor do que aquele da população de descoberta e contempla indivíduos avaliados para os marcadores SNPs e para os vários caracteres de interesse. As equações de predição de valores genéticos genômicos são testadas para verificar suas acurácias nessa amostra independente. Para computar essa acurácia, os valores genéticos genômicos são preditos (usando os efeitos estimados na população de estimação) e submetidos à análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos marcadores, os erros dos valores genéticos genômicos e dos valores fenotípicos são independentes e a correlação entre esses valores é predominantemente de natureza genética e equivale à capacidade preditiva ($r_{y\hat{y}}$) da GWS em estimar os fenótipos, sendo dada pela própria acurácia seletiva ($r_{g\hat{g}}$) multiplicada pela raiz quadrada da herdabilidade individual (h), ou seja, $r_{y\hat{y}} = r_{g\hat{g}}\, h$ conforme demonstrado em tópico posterior. Assim, para estimação da própria acurácia, deve-se obter $r_{g\hat{g}} = r_{y\hat{y}} / h$. Isso é válido quando são usados os valores fenotípicos brutos para cômputo da correlação. Quando são usados valores genotípicos preditos com base nos fenótipos em vez dos valores fenotípicos brutos, a herdabilidade deve ser substituída pela confiabilidade. De maneira geral, adota-se a estratégia (ii), segundo um esquema *Jacknife* de validação cruzada, conforme descrito no tópico seguinte.

População de Seleção. Esse conjunto de dados contempla apenas os marcadores avaliados nos candidatos à seleção. Essa população não necessita ter os seus fenótipos avaliados. As equações de predição derivadas na população de descoberta são então usadas na predição dos valores genéticos genômicos (VGG) ou fenótipos futuros dos candidatos à seleção. Mas a acurácia seletiva associada refere-se àquela calculada na população de validação.

A seguinte estratégia e sequência de análise envolvendo as populações de estimação e validação podem ser indicadas:

- compute a predição dos valores genéticos genômicos (VGG) usando todos os marcadores e calcule a correlação $r_{VGG,y}$ entre VGG e y.

- Ordene os marcadores por maiores módulos dos efeitos estimados dos marcadores.

- Crie arquivos com subconjuntos dos marcadores com maiores módulos dos efeitos estimados dos marcadores (100, 250, 500, 1.000, 1.500, 2.000, ...).

- Analise todos esses arquivos e compute as correlações $r_{VGG,y}$ e escolha o arquivo ótimo que maximiza a $r_{VGG,y}$.

- Faça a validação nesse arquivo ótimo com $k = 2$ no processo Jacknife descrito a seguir.

- Faça a validação nos outros arquivos menores que o ótimo e um maior para ver tendências (usar $k = 2$).

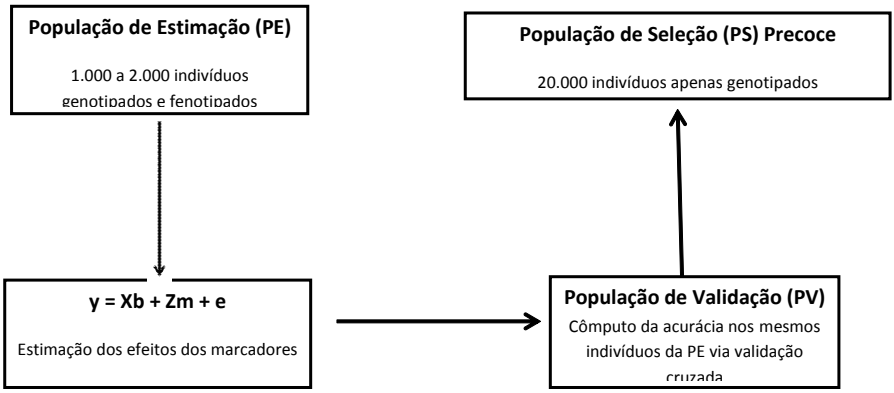

Figura 1. Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético.

População de validação e *Jacknife*

Na estimação de um parâmetro q a partir de uma amostra ou conjunto de dados com n observações, o procedimento *Jackknife* para a estimação da variância do estimador $\hat{\theta}$ consiste na omissão de cada uma das n observações, uma em cada reamostragem. A metodologia generalizada do *Jackknife* baseia-se na divisão do conjunto de *N* dados amostrais em g grupos de tamanho igual a *k*, de forma que N = gk. Em geral, k é tomado como 1, mas, pode ser tão grande quanto *N/2*. O estimador $\hat{\theta}_i$ corresponde àquele baseado em amostras de tamanho (g -1) k, onde o i-ésimo grupo de tamanho k foi removido. Com $k = 1$, $N = g$ e $(g - 1)k = g - 1 = N-1$, de forma que $\hat{\theta}_i$ refere-se à amostra em que foi omitida a observação i (RESENDE, 2002).

As validações realizadas no tópico sobre o **s**oftware Selegen Genômica, com k = 1 e k = 2, conduziram aos mesmos valores de acurácia na população de validação. Assim, não há necessidade de usar k = 1, sendo que valores maiores são também suficientes para a validação cruzada.

Correlação e regressão entre valores genéticos preditos e fenótipos na população de validação

Os coeficientes de correlação e regressão envolvendo valores observados e preditos são medidas práticas da capacidade dos métodos predizerem de forma acurada e não viesada, respectivamente. A correlação fornece a capacidade preditiva, a qual equivale ao produto da acurácia pela raiz quadrada da herdabilidade. O coeficiente de regressão equivale algebricamente a 1.

Coeficientes de regressão abaixo de 1 indicam que os valores genéticos são superestimados e apresentam variabilidade além da esperada e, acima de 1, indicam que os valores genéticos estimados apresentam variabilidade aquém da esperada. Não vício é importante quando a seleção envolve indivíduos de muitas gerações usando efeitos dos marcadores estimados em uma só geração. Coeficientes de regressão próximos de 1 indicam que as avaliações são não viesadas e são efetivas em predizer as reais magnitudes das diferenças entre os indivíduos em avaliação. A seguir são apresentadas algumas definições paramétricas importantes envolvendo os valores fenotípicos corrigidos (y_c) e os valores genéticos genômicos preditos na população de validação ($\hat{g}_V$).

Covariância

$$Cov(\hat{g}_V, y_c) = Cov[\hat{g}_V, y_c] = Cov[\hat{g}_V, (g+e)] = Cov(\hat{g}_V, g)$$

Variâncias

$$Var(\hat{g}_V) = \sigma^2_{\hat{g}_V}$$

$$Var(y_c) = \sigma^2_{y_c} = \sigma^2_g + \sigma^2_e = \sigma^2_g / h^2$$

Correlação

$$r_{gf} = Cor(\hat{g}_V, y_c) = Cov(\hat{g}_V, y_c)/(\sigma_{\hat{g}_V}\sigma_{y_c}) = Cov(\hat{g}_V, g)/(\sigma_{\hat{g}_V}\sigma_{y_c})$$
$$= Cov(\hat{g}_V, g)/[\sigma_{\hat{g}_V}(\sigma_g^2/h^2)^{1/2}] = Cov(\hat{g}_V, g)/[\sigma_{\hat{g}_V}(\sigma_g/h)] = r_{\hat{g}g}h$$

Regressão de y_c em $\hat{g}_V$

$$b_{y\hat{g}} = Reg(y_c/\hat{g}_V) = Cov(\hat{g}_V, y_c)/(\sigma^2_{\hat{g}_V}) = Cov(\hat{g}_V, g)/(\sigma^2_{\hat{g}_V}) = \sigma^2_{\hat{g}_V}/\sigma^2_{\hat{g}_V} = 1$$

Acurácia

$$r_{\hat{g}g} = r_{gf}/h$$

Confiabilidade ou determinação

$$r^2_{\hat{g}g} = (r_{gf}/h)^2$$

O erro padrão da estimativa da acurácia pode ser computado por $s(r_{\hat{g}g}) = [(1-r^2_{\hat{g}g})/(N-2)]^{1/2}$

No tópico do Selegen Genômica, para o arquivo com 500 marcas, a correlação entre os valores genéticos genômicos nas populações de estimação e de validação equivaleu a 0,96 e o coeficiente de regressão dos primeiros nos últimos equivaleu a 1,03. Esses resultados confirmam que existe parentesco genealógico entre os indivíduos da população, ou seja, que as duas populações não são totalmente independentes.

Análise de associação na GWAS

A GWAS (*Genome Wide Association Studies*) procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses, visando detectar efeitos com significância estatística.

O seguinte modelo de regressão em marcas simples pode ser empregado visando à associação entre marcador e QTL em uma população panmítica, conforme Resende (2008).

$y = 1u + Xm + e,$

em que *y* é o vetor de observações fenotípicas, 1 é um vetor com valores 1, *u* é o escalar referente à média geral, *m* é o efeito fixo do marcador, *e* refere-se ao vetor de resíduos aleatórios e *X* é a matriz de incidência para *m*. A dimensão de *m* é igual ao número de alelos do marcador.

Esse modelo assume que o marcador afetará o caráter apenas se ele estiver em LD com o suposto QTL.

A estrutura de médias e variâncias é definida como:

$$E(y) = 1u + Xm$$

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

em que σ_e^2 é a variância residual.

Como exemplo, considere a avaliação de 12 indivíduos para um caráter e para um marcador do tipo SNP. Os dados referentes aos genótipos e fenótipos dos indivíduos são apresentados na Tabela 7.

Tabela 7. Avaliação de 12 indivíduos para um caráter e para um marcador do tipo SNP.

Indivíduo	Fenótipo	Primeiro alelo do SNP1	Segundo alelo do SNP1
1	9,87	A	a
2	14,48	A	A
3	8,91	A	a
4	14,64	A	A
5	9,55	A	a
6	7,96	a	a
7	16,07	A	A
8	14,01	A	a
9	7,96	a	a
10	21,17	A	A
11	10,19	A	a
12	9,23	A	A

A matriz de incidência X associa os números de cada alelo do SNP aos fenótipos. É suficiente ajustar o efeito de apenas um dos alelos. Assim, a matriz X terá apenas uma coluna para o efeito de um dos alelos do SNP, por exemplo, o A. Essa coluna contém o número de cópias do alelo A que os indivíduos possuem. Portanto, contém os valores 0, 1 ou 2 para um indivíduo diplóide. O número de linhas dessa matriz é igual ao número de indivíduos.

A matriz 1 inclui uma coluna para a média geral. As matrizes 1 e X (número de alelos A) são dadas por

$$1_{(12x1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{e} \quad X_{(12x1)} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 0 \\ 2 \\ 1 \\ 0 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

As equações de quadrados mínimos para a estimação dos efeitos da média geral e do SNP equivalem a:

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

em que y é o vetor de fenótipos. Resolvendo-se esse sistema, obtém-se: $\begin{bmatrix} \hat{u} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} 7{,}2713 \\ 3{,}7856 \end{bmatrix}$

A hipótese da nulidade, ou seja, de que o marcador não apresenta qualquer efeito sobre o caráter, pode ser avaliada pelo teste F. A hipótese nula é rejeitada se $F > F(a, v_1, v_2)$, em que F é a estatística de Snedecor calculada dos dados, *a* é o nível de significância e v_1 e v_2 são os graus de liberdade associado à distribuição F tabelada. A hipótese alternativa é de que o marcador afeta o caráter, ou seja, devido ao fato de que marcador e QTL encontram-se em desequilíbrio de ligação.

O valor da estatística F é calculado, conforme Resende (2008), via.

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

No presente exemplo, o valor calculado de F foi de 9,74. Tal valor pode ser comparado com o valor tabelado de F ao nível de significância de 5% e graus de liberdade 1 e 10, o qual equivale a 4,96. Assim, o efeito do SNP é significativo. Isso era esperado, pois, associados aos maiores valores fenotípicos estão os alelos A do SNP, conforme se vê claramente na tabela dos dados.

O nível de significância a ser adotado em estudos de associação genômica ampla demanda sérias considerações. Isto porque milhares de marcadores estarão sendo testados e, portanto, existe o problema de múltiplos testes de hipótese. Nesse caso, o nível nominal de significância adotado para cada teste não corresponde àquele realizado em todo o experimento. Com um nível de significância de 5%, espera-se 5% dos resultados como falsos positivos. Com 20 mil marcadores, o número de falsos positivos esperados é de 1.000. A correção de Bonferroni poderia aliviar isso. Entretanto, ela não leva em consideração que os testes no mesmo cromossomo não são independentes, pois os marcadores podem estar em desequilíbrio de ligação entre eles e também com o QTL.

Uma boa alternativa é usar o conceito da taxa de descobertas falsas (FDR), definida como a proporção esperada de QTLs detectados que são falsos positivos. A FDR pode ser calculada como FDR = m Pmax/n, em que Pmax é o maior Pvalor de QTL que excede o nível de significância, *n* é o número de QTLs que excedem o nível de significância e *m* é o número de marcadores testados. Com 10 mil SNPs testados, nível de significância (Pvalor) de 0,001 e 80 SNPs declarados como significativos, a FDR = 10.000 x 0,001/80 = 0,125. Essa magnitude (12,5%) de taxa de falsa descoberta pode ser considerada aceitável (não muito acima de 10%).

Software Selegen Genômica: *Random (Ridge) Regression* BLUP: RR-BLUP/GWS

Considere o pequeno exemplo da Tabela 8, referente à avaliação de cinco indivíduos para o caráter diâmetro e genotipagem para sete marcas, em que são apresentados o número de um dos alelos de cada loco marcador.

Tabela 8. Avaliação de cinco indivíduos para o caráter diâmetro e genotipagem para sete marcas.

Indivíduo	Diâmetro	Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
1	9,87	2	0	0	0	2	0	0
2	14,48	1	1	0	0	1	1	0
3	8,91	0	2	0	0	0	0	2
4	14,64	1	0	1	0	1	0	0
5	9,55	1	0	0	1	1	1	0

Os efeitos genéticos dos marcadores são obtidos resolvendo-se

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Tem-se as seguintes matrizes

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

$$y = \begin{bmatrix} 9.87 \\ 14.48 \\ 8.91 \\ 14.64 \\ 9.55 \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Efetuando-se as multiplicações e assumindo $\frac{\sigma_e^2}{(\sigma_g^2/n)} = 1$, tem-se

$X'X = [5]$;

X'Z = [5 3 1 1 5 2 2]

Z'X = *(X'Z)'* = [5 3 1 1 5 2 2]'

$$Z'Z + I = \begin{bmatrix} 8 & 1 & 1 & 1 & 7 & 2 & 0 \\ 1 & 6 & 0 & 0 & 1 & 1 & 4 \\ 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 & 1 & 0 \\ 7 & 1 & 1 & 1 & 8 & 2 & 0 \\ 2 & 1 & 0 & 1 & 2 & 3 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

X'y = [57,45]

$$Z'y = \begin{bmatrix} 58{,}4100 \\ 32{,}3000 \\ 14{,}6400 \\ 9{,}5500 \\ 58{,}4100 \\ 24{,}0300 \\ 17{,}8200 \end{bmatrix}$$

Assim, tem-se:

$$\begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} 5 & 5 & 3 & 1 & 1 & 5 & 2 & 2 \\ 5 & 8 & 1 & 1 & 1 & 7 & 2 & 0 \\ 3 & 1 & 6 & 0 & 0 & 1 & 1 & 4 \\ 1 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 1 & 1 & 0 \\ 5 & 7 & 1 & 1 & 1 & 8 & 2 & 0 \\ 2 & 2 & 1 & 0 & 1 & 2 & 3 & 0 \\ 2 & 0 & 4 & 0 & 0 & 0 & 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 57{,}4500 \\ 58{,}4100 \\ 32{,}3000 \\ 14{,}6400 \\ 9{,}5500 \\ 58{,}4100 \\ 24{,}0300 \\ 17{,}8200 \end{bmatrix}.$$

Os resultados são

$$\begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} 12{,}4519 \\ -0{,}3526 \\ 0{,}2761 \\ 1{,}4467 \\ -1{,}3701 \\ -0{,}3526 \\ 0{,}5436 \\ -1{,}63765 \end{bmatrix}$$

em que 12,4519 é a média geral e os demais valores são as estimativas dos efeitos dos marcadores.

O valor genético genômico dos indivíduos de uma população de seleção podem ser obtidos por $VGG = \hat{y}_j = \sum_i Z_i \hat{m}_i$. No caso, as predições são

$$VGG = \begin{bmatrix} -1,4104 \\ 0,1145 \\ -2,7230 \\ 0,7415 \\ -1,5317 \end{bmatrix}$$

Outro exemplo, baseado em haplótipos, é apresentado em detalhes por Resende (2008).

Um novo exemplo é apresentado a seguir, relatando a estimação e a validação em uma amostra com dez indivíduos fenotipados para o caráter diâmetro e genotipados para dez marcas. Esse conjunto de dados foi submetido à análise pelo software Selegen Genômica-RR/BLUP, em que foram obtidas as estimativas dos efeitos dos marcadores, a predição dos valores genéticos genômicos e realizada a validação.

O arquivo de resultados é apresentado a seguir e mostra cada passo do esquema *Jacknife* e, ao final, apresenta as estimativas e predições definitivas. Na validação cruzada via *Jacknife*, a capacidade preditiva obtida foi de 0,07 e a acurácia seletiva na população de seleção foi igual a 0,09, refletindo o baixo tamanho amostral (N = 10) para estimar os efeitos de marcadores. A herdabilidade usada foi 0,60.

SELEGEN-Genomica: Random (Ridge) Regression BLUP:RR-BLUP
Inicio : 2007 Versao Atual : Setembro 2010
Sistema de Selecao Genomica Ampla (GWS)
Modelos Lineares Mistos via REML/BLUP/GWS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
Universidade Federal de Viçosa Viçosa - MG
marcos.deon@gmail.com
marcos.deon@ufv.br

Arquivo : C:\Ano2010SelegenGenomica2010\Oficial-10M-10IND.txt
Modelo : 1
Número de Variáveis : 1
Variável Analisada : 1
N. Linhas Lidas : 10
Zeros significativos : Nao
Data : 29/11/2010
Hora : 18:37:58

Grupo 1

1. Componentes de Variância (REML Individual)
h2 = 0.600000
Média geral = 38.114018

Ef. Fixo	Media
1	38.1140

2. Componentes de Média (BLUP Individual)

Efeitos dos Marcadores Genéticos

Ordem	Marcador	e1	u + e1
1	10	1.4851	39.5991
2	2	1.1880	39.3021
3	3	0.4011	38.5151
4	4	0.0000	38.1140
5	8	-0.2775	37.8365
6	7	-0.3241	37.7900
7	9	-0.7143	37.3997
8	6	-0.8026	37.3115
9	5	-1.1719	36.9422
10	1	-2.7208	35.3932

Valores Genéticos Genômicos

População de Validação

Ordem	Indivíduo	g	u + g	f
1	4	-1.8460	36.2680	30.2160

Populacao de Estimação

Ordem	Indivíduo	g	u + g	f
1	9	1.6348	39.7488	44.3250
2	2	0.8748	38.9888	42.3460
3	7	-0.7732	37.3408	38.5860
4	8	-1.0949	37.0192	38.1550
5	3	-1.8420	36.2720	35.4010
6	10	-1.8460	36.2680	34.9160
7	5	-2.9726	35.1414	32.1460
8	6	-4.1141	33.9999	33.9460
9	1	-4.9310	33.1830	28.1410

Grupo 2

1. Componentes de Variância (REML Individual)

h2 = 0.600000
Média geral = 37.309216

Ef. Fixo Média
1 37.3092

2. Componentes de Média (BLUP Individual)

Efeitos dos Marcadores Genéticos

Ordem	Marcador	e1	u + e1
1	10	1.4171	38.7263
2	2	1.1198	38.4290
3	7	0.4969	37.8061
4	3	0.3209	37.6301
5	6	0.0557	37.3649
6	4	0.0000	37.3092
7	8	-0.1995	37.1097
8	9	-0.6039	36.7053
9	5	-1.1341	36.1751
10	1	-2.9844	34.3248

Valores Genéticos Genômicos

População de Validação

Ordem	Indivíduo	g	u + g	f
1	5	-1.5950	35.7142	32.1460

População de Estimação

Ordem	Indivíduo	g	u + g	f
1	9	2.4299	39.7391	44.3250
2	2	0.8368	38.1460	42.3460
3	7	0.4260	37.7352	38.5860
4	8	-0.3164	36.9929	38.1550
5	3	-1.1952	36.1140	35.4010
6	4	-2.1476	35.1616	30.2160
7	10	-2.1476	35.1616	34.9160
8	6	-3.4112	33.8980	33.9460
9	1	-4.2256	33.0836	28.1410

.
.
.

Grupo 10

---

1. Componentes de Variância (REML Individual)

h2 = 0.600000
Média geral = 37.690191

Ef. Fixo	Média
1	37.6902

2. Componentes de Média (BLUP Individual)

Efeitos dos Marcadores Genéticos

Ordem	Marcador	e1	u + e1
1	10	1.6963	39.3865
2	2	0.9076	38.5978
3	3	0.1545	37.8447
4	7	0.1039	37.7941
5	4	0.0000	37.6902
6	8	-0.0473	37.6429
7	6	-0.3336	37.3566
8	9	-0.8176	36.8726
9	5	-1.0332	36.6570
10	1	-3.1187	34.5715

Valores Genéticos Genômicos
População de Validação

Ordem	Indivíduo	g	u + g	f
1	10	-2.8742	34.8160	34.9160

Populacao de Estimação

Ordem	Indivíduo	g	u + g	f
1	9	1.8902	39.5804	44.3250
2	2	0.2445	37.9347	42.3460
3	7	-0.5900	37.1002	38.5860
4	8	-0.7748	36.9154	38.1550
5	3	-1.7132	35.9770	35.4010
6	4	-2.8742	34.8160	30.2160
7	5	-3.1039	34.5863	32.1460
8	6	-4.1627	33.5275	33.9460
9	1	-4.8656	32.8246	28.1410

Populacao de Validação (Jacknife)

Ordem	Indivíduo	g	u + g	f
1	9	-0.2431	36.2147	44.3250
2	2	-0.4889	36.0389	42.3460
3	8	-1.2304	35.8311	38.1550
4	3	-1.3602	36.4229	35.4010
5	5	-1.5950	35.7142	32.1460
6	7	-1.6202	36.0155	38.5860
7	4	-1.8460	36.2680	30.2160
8	1	-2.2946	36.5328	28.1410
9	10	-2.8742	34.8160	34.9160
10	6	-4.3690	33.2465	33.9460

rgf = 0.0707
rgg = 0.0913

Média dos Efeitos dos Marcadores Genéticos

Ordem	Marcador	e1
1	10	1.5443
2	2	0.8517
3	3	0.1100
4	7	0.1064
5	4	0.0000
6	8	-0.0721
7	6	-0.3259
8	9	-0.7399
9	5	-0.9145
10	1	-2.8076

Valores Genéticos na População de Estimação

Ordem	Indivíduo	g	u + g
1	9	1.7625	39.2648
2	2	0.2219	37.7241
3	7	-0.5212	36.9811
4	8	-0.6980	36.8043
5	3	-1.5496	35.9527
6	4	-2.5858	34.9165
7	10	-2.5858	34.9165
8	5	-2.8053	34.6969
9	6	-3.8355	33.6668
10	1	-4.3555	33.1467

Exemplo aplicado ao melhoramento do eucalipto

Esse tópico relata a aplicação da GWS em dados reais de eucaliptos híbridos envolvendo as espécies *Eucalyptus urophylla*, *E. grandis* e *E. globulus*. Foram avaliados 920 indivíduos de uma população com tamanho efetivo 100, para o caráter diâmetro do tronco. Esses indivíduos foram também genotipados para 3.564 marcadores do tipo DArT. A seguir são apresentados resultados das análises empregando o software Selegen Genômica: Random (Ridge) Regression BLUP (RR-BLUP/GWS).

Inicialmente, foram computados os valores genéticos genômicos (VGG) usando todos os marcadores. Em seguida, os marcadores foram ordenados por maiores módulos dos efeitos estimados dos marcadores e foram criados arquivos com subconjuntos dos marcadores com maiores módulos dos efeitos estimados (250, 500, 1.000, 1.500, 2.000, 3.564 marcadores). Todos esses arquivos foram analisados para cômputo de componentes de variância, herdabilidade e valores genéticos genômicos nas populações de estimação e de validação. Os resultados referentes aos componentes de variância e herdabilidade são apresentados na Tabela 9.

Tabela 9. Componentes de variância e herdabilidade estimados pelo Selegen Genômica.

Número de Marcadores (M)	Frequência Alélica (p)	Variância p(1-p)	n = M p(1-p)	Variância Marcador REML	VResidual REML	Vf	h^2	nCuv	Vg	h^2 Corr
250	0,54	0,248	62	0,32	29,32	49,19	0,404	0,221	19,651	0,399
500	0,54	0,248	124	0,2104	25,71	51,84	0,504	0,157	25,975	0,501
750	0,54	0,248	186	0,1585	24,69	54,22	**0,545**	0,151	29,529	0,545
1.000	0,55	0,248	248	0,129	24,61	56,53	**0,565**	0,145	31,928	0,565
1.500	0,54	0,248	373	0,0885	25,89	58,86	0,560	0,146	32,975	0,560
2.000	0,53	0,249	498	0,0634	28,25	59,83	0,528	0,154	31,586	0,528
3.564	0,51	0,250	890	0,0312	34,36	62,13	0,447	0,193	27,772	0,447
Fenotípica							**0,600**			

A partir da variância entre marcadores (V_m) e variância residual (V_{res}) estimados via REML pelo Selegen-REML/BLUP/GWS, a herdabilidade é estimada por $h^2 = (nV_m)/(nV_m + V_{res})$. Verifica-se que a quantidade n C_{UV} apresenta valores desprezíveis. Com números crescentes de marcadores, verifica-se que a variância explicada por cada marcador diminui, mas a variância genética total (Vg) explicada aumenta. A herdabilidade multi-locos também aumenta até o número de 1.000 marcadores, atingindo o valor de 0,565, valor este próximo do valor 0,60, estimado via dados fenotípicos. Assim, 1.000 marcadores recuperam ou explicam 94% da herdabilidade, restando apenas 6% de "*missing heritability*". Maiores números de marcadores não contribuíram para recuperar mais da herdabilidade, havendo decréscimos a partir desse número. O valor de h^2 com base na análise com 750 marcadores de maiores efeitos, também já recupera elevada fração da herdabilidade.

Na Tabela 10 são apresentados valores da capacidade preditiva da GWS na população de estimação, associada aos arquivos com diferentes números de marcadores, em ordem daqueles de maior efeito. Verifica-se que a capacidade preditiva praticamente é maximizada na análise com os 750 ou 1.000 marcadores de maiores efeitos, confirmando a inferência realizada com base na recuperação da herdabilidade. A análise com 500 marcadores de maiores efeitos também já propicia alta capacidade preditiva.

Tabela 10. Capacidade preditiva da GWS na População de Estimação.

Núm. Marcas	Núm. Locos	Herdabilidade	Correlação
250	62	0,6	0,71
500	124	0,6	0,79
750	186	0,6	**0,82**
1000	248	0,6	0,83
1500	373	0,6	0,84
2000	498	0,6	0,84
3564	890	0,6	0,82

Na Tabela 11 são apresentados valores da capacidade preditiva e acurácia da GWS na população de validação, associada aos arquivos com diferentes números de marcadores, em ordem daqueles de maior efeito. Verifica-se que a capacidade preditiva e a acurácia praticamente são maximizadas nas análises com 500 e 750 marcadores de maiores efeitos. As capacidades preditivas da Tabela 9 são menores do que aquelas da Tabela 8, conforme esperado, devido à independência das amostras de estimação e de validação, por ocasião da implementação da validação.

O aumento do número de marcadores não aumenta linearmente a acurácia da GWS pelo método RR-BLUP, concordando com os resultados de Fernando et al. (2007). Assim, entre 500 a 1.000 locos são suficientes para maximizar a acurácia na população de validação. Esse número é coerente com a quantidade n = Me V(q), que se refere ao número esperado de marcas com efeitos significativos (ver tópico Frequências Alélicas, Variância dos Marcadores e Herdabilidade). Esse valor, associado ao Ne = 100 e L = 13,2 (tamanho do genoma do eucalipto) na Tabela 2, em que consta n = 502, é coerente com o número de marcadores (500 a 750) que maximiza a acurácia com a GWS (Tabela 9). O valor da acurácia esperada, associada a 100 locos (500 marcas na Tabela 11) e um r^2_{mq} de 0,33 (segundo a equação de Sved e dado o número de marcadores e o Ne = 100) é em torno de 0,50 (RESENDE, 2008), valor esse inferior ao valor 0,70 da Tabela 11. Isso pode indicar que o r^2_{mq} da população híbrida considerada é maior do que o esperado segundo a equação de Sved, devido à origem híbrida da mesma, envolvendo o cruzamento de três espécies.

O aumento ou diminuição da acurácia da GWS via RR-BLUP é um compromisso ou balanço entre acréscimo da quantidade de informação útil via uso de maior número de locos marcadores e diminuição do tamanho de amostra efetivo para estimar o efeito de cada loco, ou seja, menor número de indivíduos por loco a ser estimado (menor N/n).

Tabela 11. Capacidade Preditiva (Correlação) e Acurácia da GWS na População de Validação.

Número Marcas	Número Locos	Herdabilidade	Correlação	Acurácia	Regressão*	Regressão**
250	62	0,6	0,51	0,66	0,82	0,98
500	124	0,6	0,54	0,70	0,89	0,99
750	186	0,6	0,53	0,69	0,92	0,97
1.000	248	0,6	0,52	0,67	0,93	0,98
1.500	373	0,6	0,46	0,59	0,91	0,95
2.000	498	0,6	0,38	0,49	0,84	0,92
3.564	890	0,6	0,10	0,13	-	-

* Predição usando herdabilidade igual a 0,60; ** Predição usando as herdabilidades da Tabela 7.

O número reduzido de marcadores explicando grande parte da variação genética ou da acurácia máxima possível é muito interessante do ponto de vista prático. Nesse caso, arranjos de DNA com baixa densidade de marcadores previamente selecionados poderiam ser usados nas populações de seleção. Na Austrália, a acurada predição de valores genéticos genômicos em gado leiteiro pode ser realizada com chips de SNP contendo 1.000 (propiciando 85% da acurácia obtida com 42.500 SNP) a 5.000 (propiciando 95% da acurácia obtida com 42.500 SNP) SNP igualmente espaçados (MOSER et al., 2010).

Uma alternativa ao uso de marcadores previamente selecionados é o uso de marcadores igualmente espaçados e em maior número do que aqueles selecionados. Isso permite atender a vários caracteres e pode conduzir ao uso generalizado da GWS em várias espécies e países.

Na Tabela 12 são apresentados resultados referentes ao comportamento da capacidade preditiva em função da

herdabilidade usada na predição, para os arquivos com 10, 250, 500 e 750 marcadores.

Tabela 12. Comportamento da capacidade preditiva (rgf da população de validação) em função da herdabilidade usada na predição, para os arquivos com 10, 250, 500 e 750 marcadores.

h^2	10 Marcas	250 Marcas	500 Marcas	750 Marcas
	rgf	rgf	rgf	Rgf
0,05	0,265	0,466	0,467	0,441
0,1	**0,266**	0,486	0,496	0,473
0,2	0,266	0,501	0,521	0,504
0,3	0,266	0,503	0,531	0,518
0,4	0,266	**0,504**	0,534	0,525
0,5	0,265	0,499	**0,535**	0,526
0,6	0,265	0,493	0,530	**0,526**
0,7	0,265	0,485	0,521	0,513
0,8	0,265	0,476	0,508	0,495
0,9	0,265	0,462	0,484	0,456
1	0,265	0,432	0,375	0,181

Verifica-se que, em cada arquivo, a herdabilidade que maximiza a capacidade preditiva na validação cruzada é aquela estimada por REML, especificamente para cada arquivo (Tabela 9). Assim, esses pontos de máximo foram propiciados pelas herdabilidades de 0,10; 0,40; 0,50 e 0,55 para os arquivos com 10, 250, 500 e 750 marcadores, respectivamente, conduzindo a capacidades preditivas de 0,27; 0,50; 0,53 e 0,53, respectivamente (Tabela 12). Assim, entre diferentes arquivos, maiores herdabilidades verdadeiras (assinaladas em negrito na Tabela 12) propiciam maior capacidade preditiva na validação cruzada. O ponto de máximo da capacidade preditiva reflete a coerência interna e intrínseca dos dados em informar sobre o fenótipo. Em geral, o ponto de máximo global da capacidade preditiva situa-se na herdabilidade entre 0,5 e 0,6 e com número de marcas entre 500 e 750 (Tabela 12).

Os pontos de máximo da capacidade preditiva podem então ser usados de forma reversa para informar sobre a herdabilidade capturada pelos marcadores, de forma alternativa ou confirmativa do método REML. Assim, a h^2 e o parâmetro λ são obtidos por sintonia fina ou ajuste na própria validação cruzada via o modelo com maior capacidade preditiva ou maior determinação para predizer o fenótipo. Dessa forma, a h^2 otimizada para cada arquivo é que deve integrar o λ. Mas a h^2 usada para computar a acurácia a partir da capacidade preditiva, via $r_{\hat{g}g} = r_{gf} / h$, deve ser a h^2 total ajustada, estimada dos próprios dados fenotípicos. Essa tende a ser similar a h^2 estimada via marcadores, quando se usa o total de marcas em grande número.

Coeficientes de regressão

As duas últimas colunas da Tabela 11 apresentam os coeficientes de regressão envolvendo valores observados e preditos. Na primeira dessas colunas, a predição empregou a herdabilidade de 60% e na segunda dessas colunas, a predição empregou as herdabilidades estimadas por REML, conforme a Tabela 9. O coeficiente de regressão tem valor esperado igual a 1 e nessa situação indica que a predição foi não viesada. Observa-se que os valores da última coluna tenderam a 1, ao passo que aqueles da outra coluna foram menores que 1. Isso indica que as herdabilidades mais adequadas são aquelas estimadas por REML, ou seja, uma para cada conjunto de dados segundo o número de marcas.

Assim sendo, pode-se também usar o coeficiente de regressão para estimar a herdabilidade a ser empregada. Vários valores de herdabilidade são avaliados e aquele que fornecer uma regressão igual a 1 deve ser escolhido como melhor estimativa. Se a regressão der resultado menor que 1, o valor de herdabilidade avaliado foi de alta magnitude e deve ser diminuído até a convergência para 1. Se a regressão der resultado maior que 1, o

valor de herdabilidade avaliado foi de pequena magnitude e deve ser aumentado até a convergência para 1. Isso caracteriza o método R de estimação de componentes de variância, conforme descrito por Resende (2002).

Acurácia via Inversão da matriz dos coeficientes

O cálculo tradicional da acurácia pelo BLUP via inversão da matriz dos coeficientes das equações de modelo misto (e ponderação pelas frequências alélicas dos locos marcadores) não se adéqua bem à seleção genômica, uma vez que fornece acurácias sempre menores com o aumento do número de marcas (Tabela 13). Isto porque, estimada dessa forma, a acurácia reflete apenas o conteúdo de informação, basicamente dependente do tamanho amostral N e do número de parâmetros n (número de marcadores) a serem estimados. É um reflexo então da quantidade $\omega = N / n$, crescendo com o aumento dessa quantidade. Não considera, portanto, o quanto cada genótipo marcador dita o mérito genético. Assim, a acurácia via capacidade preditiva é uma medida mais adequada.

Tabela 13. Acurácia via Inversão da Matriz dos Coeficientes.

Númeroo Marcas	Número Locos	h^2	Acurácia
10	2,42	0,6	0,99
250	62	0,6	0,79
500	124	0,6	0,66
750	186	0,6	0,58
1.000	248	0,6	0,51

Softwares em R

Os seguintes softwares em linguagem R foram desenvolvidos: (i) GWS-R por Resende Júnior (2010), que emprega o método RR-BLUP; (ii) BLR-R (*Bayesian Linear Regression*) por Perez et al. (2010), que implementa vários métodos (RR, Bayesian

Lasso, *Bayesian Ridge Regression*). O pacote GWS-R realiza a estimação, a validação e também a análise de associação via regressão em marcas individuais e está disponível mediante solicitação aos autores desse trabalho.

A seguinte sequência pode ser adotada no uso do GWS-R.

No Ambiente R

As seguintes funções podem ser utilizadas no pacote:

myData: Carrega o arquivo de marcadores

myData(<"Diretório">, <"aquivo.extensão">, missingM = <"Valor do dado perdido">)

stanData: Padroniza os arquivos de dados como mencionado em tópicos anteriores de acordo com o tipo de marcador usado

stanData(<nome da variável em que foi salva a função myData>,<"tipo de marcador"(DARTs ou SNPs)>, "frequency

jacknife: Define o número de grupos utilizados na validação cruzada via Jacknife

Jacknife(<Variavel em que foi salvo a função stanData>, indsel = <>, nv = <>, <"diretório">, <"arquivo com referência para os fenótipos usados">, ngroups = <>, random = FALSE)

BLUP: Essa função usa os dados carregados nas funções anteriores para realizar a GS via Blup.

BLUP(<Variavel em que foi salva a função Jacknife)>, mark.val, siglev, acc.From.Data = TRUE, Ftest = FALSE, fdr = FALSE, k = 1)

Exemplo

```
> arquivo <- myData ("C:/Ano2010/GWS-R", "dartgM.txt",
missingM = "NA")

> stanD <- stanData(arquivo, "DARTs", "frequency")

> jacknife <- Jacknife(stanD, indsel = 0,nv = 1,"C:/Ano2010/
GWS-R", "caminhoMM.txt", ngroups = 460, random = FALSE)

> blup <- BLUP(jacknife, mark.val = c(2000), siglev = 0.05,
acc.From.Data = TRUE, Ftest = FALSE, fdr = FALSE, k = 1)

 > blup$correlation

 > blup$accuracy
```

As seguintes definições são necessárias:

1"C:/Ano2010/GWS-R": caminho do diretório;

2 dartgM.txt: arquivo com os dados dos marcadores moleculares (M marcadores x N indivíduos);

3 "caminhoMM.txt": arquivo que informa sobre o arquivo de dados fenotípicos e sobre a herdabilidade;

4 ngroups = 460: número de grupos para a validação Jacknife;

5 mark.val = c(2000): número de marcadores a ser incluído na análise;

6 O valor de n em lambda no programa pode ser alterado diretamente no conteúdo do arquivo R Pack.txt;

7 O conteúdo do arquivo "caminhoMM.txt" pode ser do tipo "CAP1" "C:/Ano2010/gws-r/CAP1.txt" 0.6 1.0, o qual informa que os dados fenotípicos estão no arquivo CAP1.txt e que a h^2 a ser usada é 0,60;

8 Nesse caso, para executar o programa são necessários os arquivos: dartgM.txt; caminhoMM.txt e CAP1.txt.

Referências

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

CAMPOS, G. de los; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers. **Genetics**, v. 182, p. 375-385, May. 2009.

DAETWYLER, H.; VILLANUEVA, B.; WOOLLIAMS, J. Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLoS ONE**, v. 3, n. 10, e3395, Oct. 2008.

DIAS, L. A. dos S.; RESENDE, M. D. V. de. Domesticação e melhoramento de cacau. In: BORÉM, A.; LOPES, M. T. G.; CLEMENT, C. R. (Org.). **Domesticação e melhoramento**: espécies amazônicas. Viçosa, MG: UFV, 2009. p. 251-274.

FALCONER, D. S. **Introduction to quantitative genetics**. 3rd ed. Harlow: Longman, 1989. 438 p.

FERNANDO, R. L.; HABIER, D.; STRICKER, C.; DEKKERS, J. C. M.; TOTTIR, L. R. Genomic selection. **Acta Agriculturae Scandinavica, Section A - Animal Science**, v. 57, n. 4, p. 192-195, 2007.

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

GIANOLA, D.; CAMPOS, G. de los. Inferring genetic values for quantitative traits non-parametrically. **Genetic Research**, v. 90, n. 6, p. 525-540, 2009.

GIANOLA, D.; FERNANDO, R. L; STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v. 173, p. 1761-1776, Jul. 2006.

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

GIANOLA, D.; PEREZ-ENCISO, M.; TORO, M. A. On marker-assisted prediction of genetic value: beyond the ridge. **Genetics**, v. 163, p. 347-365, Jan. 2003.

GIANOLA, D.; VAN KAAM, J. B. C. H. M. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, v. 178, n. 4, p. 2289–2303, Apr. 2008.

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

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

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

GONZALEZ-RECIO, O.; GIANOLA, D.; LONG, N.; WEIGEL, K. A.; ROSA, G. J. M.; AVENDANO, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, v. 178, n. 4, p. 2305–2313, Apr. 2008.

GONZALEZ-RECIO, O.; GIANOLA D.; ROSA, G. J. M.; WEIGEL, K. A.; KRANIS, A. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. **Genetics Selection Evolution**, v. 41, n. 3, 2009.

GRATTAPAGLIA, D.; RESENDE, M. D. V. Genomic selection in forest tree breeding. **Tree Genetics and Genomes**, Oct. 2010.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of genetic relationship on genome-assisted breeding values. **Genetics**, v. 117, p. 2389-2397, Dec. 2007.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the Bayesian alphabet for genomic selection. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 9, 2010, Leipzig, Germany. [**Proceedings...**]. [S.l.: s.n.], 2010. Disponível em: <http://www.kongressband.de/wcgalp2010/assets/pdf/0468.pdf>. Acesso em: 20 nov. 2010.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, v. 92, n. 2, p. 433-443, 2009.

HAYES, B. J.; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, v. 33, p. 209-229, 2001.

HAYES, B. J.; VISSCHER, P. M.; GODDARD, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics Research**, v. 91, p. 47-60, 2009.

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

LONG, N.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A; AVENDAÑO, S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, v.124, n. 6, p. 377–389, Dec. 2007.

MAHER, B. The case of the missing heritability. **Nature**, London, v. 456, p.18-21, 2008.

MEUWISSEN, T. H. E. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, v. 124, p. 321-322, 2007.

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

MEUWISSEN, T. H. E.; 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***,* v. 41, 2009.

MOSER, G.; MEHAR, S. K.; 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, n. 1, p. 37, 2010.

PARK, T.; CASELLA, G. The Bayesian LASSO. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, June. 2008.

PÉREZ, P.; CAMPOS, G. de los; CROSSA, J.; GIANOLA, D. **Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R.** 2010. Disponível em: <http://genomics.cimmyt.org/BLR_DRAFT.pdf>. Acesso em: 25 nov. 2010.

RESENDE, M. D. V. de; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. . Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, v. 56, p. 63-78, 2008.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 2008. 330 p.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica: Colombo, Embrapa Florestas, 2002. 975 p.

RESENDE, M. D. V. **Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético**. Colombo: Embrapa Florestas, 2007. 561 p.

RESENDE, M. D. V. de. **Selegen–Reml/Blup:** Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos. Colombo: Embrapa Florestas, 2007. 361 p.

RESENDE JÚNIOR, M. F. R. **Seleção genômica ampla no melhoramento vegetal**. 2010. 67 p. Dissertação (Mestrado em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa, MG.

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selections Evolution**, v. 41, n. 29, 2009. Disponível em: <http://www.gsejournal.org/content/41/1/29>. Acesso em: 30 out. 2010.

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

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistics Society Series B**, 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, n. 6, p. 427-36 Dec. 2009.

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

VAN RADEN, P. M.; VAN TASSELL, C. P.; WIGGANS, G. R., SONSTEGARD, T. S.; SCHNABEL, R. D.; SCHENKEL, F.; TAYLOR, J. F. Invited review: reliability of genomic predictions for North American dairy bulls. **Journal of Dairy Science**, v. 92, n. 1, p. 16-24, Jan. 2009.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker assisted selection using ridge regression. **Genetical Research**, v. 75, n. 2, p. 249-252, 2000.