Documentos

ISSN 1980-3958
Dezembro, 2011

219

Métodos estatísticos na seleção genômica ampla

Embrapa

ISSN 1980-3958
Dezembro, 2011

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

Documentos 219

Métodos estatísticos na seleção genômica ampla

Marcos Deon Vilela de Resende
Fabyano Fonseca e Silva
José Marcelo Soriano Viana
Luíz Alexandre Peternelli
Márcio Fernando Ribeiro Resende Jr.
Patricio Muñoz del Valle

Embrapa Florestas
Colombo, PR
2011

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
Capa: Mauro Marcelo Berté

1ª edição
Versão digital (2011)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Florestas

Métodos estatísticos na seleção genômica ampla [recurso eletrônico] / Marcos Deon Vilela de Resende ... [et al.]. Dados eletrônicos - Colombo : Embrapa Florestas, 2011.
(Documentos / Embrapa Florestas, ISSN 1980-3958 ; 219)

Sistema requerido: Adobe Acrobat Reader.
Modo de acesso: World Wide Web.
<http://www.cnpf.embrapa.br/publica/seriedoc/edicoes/doc219.pdf>
Título da página da web (acesso em 10 abr. 2012).

1. Métodos estatísticos. 2. Medição. 3. Melhoramento vegetal 4. Melhoramento animal. 5. Marcador genético. I. Resende, Marcos Deon Vilela de. II. Silva, Fabyano Fonseca e. III. Viana, José Marcelo Soriano. IV. Peternelli, Luíz Alexandre. V. Resende Júnior, Márcio Fernando Ribeiro. VI. Muñoz del Valle, Patricio. VII. Série.

CDD 631.52 (21. ed.)

Autores

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

Fabyano Fonseca e Silva
Zootecnista, Doutor,
Professor da UFV
fabyanofonseca@ufv.br

José Marcelo Soriano Viana
Engenheiro Agrônomo, Doutor,
Professor da UFV
jmsviana@ufv.br

Luíz Alexandre Peternelli
Engenheiro Agrônomo, Doutor,
Professor da UFV
peternelli@ufv.br

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

Patricio Muñoz Del Valle
Engenheiro Florestal, Mestre,
Estudante na Universidade da Flórida
pmunoz@ufl.edu

Apresentação

A seleção genômica ampla (GWS) aumenta a eficiência e rapidez do melhoramento genético. Essa tecnologia fundamenta-se na predição de fenótipos com base na leitura de marcadores genéticos e uso de métodos preditivos. Existem vários métodos para aplicação na GWS. O presente documento contempla mais de uma dezena desses métodos contemplando a teoria, a computação e a aplicação a dados simulados. Assim, seu conteúdo é abrangente e pode servir como um guia importante para os usuários da GWS.

Washigton L. E. Magalhães
Chefe de Pesquisa e Desenvolvimento

Sumário

Métodos estatísticos na seleção genômica ampla

Marcos Deon Vilela de Resende
Fabyano Fonseca e Silva
José Marcelo Soriano Viana
Luíz Alexandre Peternelli
Márcio Fernando Ribeiro Resende Jr.
Patricio Muñoz del Valle

Descrição genérica dos métodos de seleção genômica ampla (GWS)

Os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS) e seleção (ou estimação) genômica ampla (Genome Wide Selection - GWS) são importantes no melhoramento genético de animais e plantas e também na genética humana. No melhoramento genético, a GWS aumenta a eficiência e rapidez do processo seletivo. Em genética humana, as ferramentas da GWS propiciam a medicina personalizada ou medicina genômica (WRAY, 2005; WRAY et al., 2007; GODDARD et al., 2009; CAMPOS et al., 2010; MAKOWSKY et al., 2011), a qual fundamenta-se na predição de fenótipos com base na leitura de marcadores genéticos e uso de métodos preditivos. As predições geradas são usadas na diagose, prevenção e tratamento das doenças.

Um método ideal para GWS deve contemplar três atributos: (i) acomodar a arquitetura genética do caráter em termos de genes de pequenos e grandes efeitos e suas distribuições; (ii) realizar a regularização do processo de estimação em presença de multicolinearidade e grande número de marcadores, usando para isso estimadores do tipo *shrinkage*; (iii) realizar a seleção de covariáveis (marcadores) que afetam a característica em análise.

O problema principal da GWS é a estimação de um grande número de efeitos a partir de um limitado número de observações e também as colinearidades advindas do desequilíbrio de ligação entre os marcadores. Os estimadores do tipo *shrinkage* lidam adequadamente com isso, tratando os efeitos de marcadores como variáveis aleatórias e estimando-os simultaneamente (Resende et al., 2008).

Os principais métodos para a GWS podem ser divididos em três grandes classes: regressão explícita, regressão implícita e regressão com redução dimensional. Na primeira classe, destacam-se os métodos RR-BLUP, LASSO (*Least Absolute Shrinkage and Selection Operator*), Rede Elástica (*Elastic Net* – EN), BayesA e BayesB, dentre outros. Na classe de regressão implícita, citam-se os métodos RKHS (*Reproducing Kernel Hilbert Spaces*, que é um método semi-paramétrico, assim como o método de redes neurais) (GIANOLA; CAMPOS, 2009) e regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA et al., 2006). Dentre os métodos de regressão com redução dimensional, destacam-se o de quadrados mínimos parciais e de componentes principais. A Tabela 1 ilustra os métodos para GWS.

Tabela 1. Classificação dos Métodos para GWS.

Classe	Família	Método	Atributos
Regressão explícita	**Métodos de estimação penalizada (Regressão linear)**	RR-BLUP/GWS	Regularização, Arquitetura genética homogênea, Seleção indireta de covariáveis
		LASSO	Regularização, Arquitetura genética homogênea, Seleção direta de covariáveis
		EN	Regularização, Arquitetura genética homogênea, Seleção direta de covariáveis
		RR-BLUP-Het/GWS	Regularização, Arquitetura genética flexível, Seleção indireta de covariáveis
	Métodos de estimação bayesiana (Regressão não linear)	BayesA	Regularização, Arquitetura genética flexível, Seleção indireta de covariáveis
		BayesB	Regularização, Arquitetura genética flexível, Seleção direcionada de covariáveis
		Fast BayesB	Regularização, Arquitetura genética flexível, Seleção direcionada de covariáveis
		BayesCπ	Regularização, Arquitetura

			genética homogênea, Seleção direta de covariáveis
		BayesDπ	Regularização, Arquitetura genética flexível, Seleção direta de covariáveis
Regressão explícita	**Métodos de estimação bayesiana (Regressão não linear)**	BLASSO	Regularização, Arquitetura genética flexível, Seleção direta de covariáveis
		IBLASSO	Regularização, Arquitetura genética flexível, Seleção direta de covariáveis
Regressão implícita		Regressão Kernel RKHS Redes neurais	
Regressão com redução dimensional		Quadrados mínimos parciais Componentes principais	

Os métodos de regressão implícita são divididos em dois grupos: (i) métodos de estimação penalizada (RR-BLUP, LASSO, EN, RR-BLUP-Het); (ii) métodos de estimação bayesiana (BayesA, BayesB, Fast BayesB, BayesCπ, BayesDπ, BLASSO, IBLASSO e outros) (Tabela 1). Os estimadores penalizados são obtidos como solução para um problema de otimização, em que a função objetivo (função cujo valor é minimizado ou maximizado, dependendo do problema e objetivo) é definida pelo balanço entre precisão do ajuste (soma de quadrado dos resíduos) e complexidade

do modelo (componente de penalização). Os métodos de estimação penalizada diferem de acordo com as funções de penalização usadas, as quais produzem diferentes graus de *shrinkage*. Esse encurtamento previne a super-parametrização e pode conduzir à redução do erro quadrático médio de estimação.

Os métodos bayesianos estão associados a sistemas de equações não lineares e as predições não lineares podem ser melhores quando os efeitos de *Quantitative trait loci* (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 com mesma frequência alélica contribuem igualmente para a variação genética (ausência de genes de efeitos maiores). Na estimação bayesiana, o encurtamento das estimativas dos efeitos do modelo é controlado pela distribuição *a priori* assumida para esses efeitos. Diferentes prioris induzem a diferentes encurtamentos. Os métodos de estimação penalizada e os bayesianos podem ser com (BayesB, Fast BayesB,BayesCπ, BayesDπ, LASSO, BLASSO, IBLASSO) ou sem (RR-BLUP, EN, RR-BLUP-Het, BayesA) seleção direta de covariáveis. 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.

Comparações entre os métodos de predição de valores genéticos genômicos têm sido realizadas. Meuwissen et al. (2001) concluíram pela superioridade teórica do método BayesB, o qual mostrou-se ligeiramente superior ao RR-BLUP. Entretanto, o autor simulou os dados genotípicos segundo a mesma distribuição *a priori* empregada no processo de estimação. Isso conduziu a acurácias mais elevadas por esse método, as quais podem não ser

realísticas na prática, se a distribuição real associada aos efeitos genéticos diferir da distribuição *a priori* assumida na análise.

Comparando métodos bayesianos, Habier et al. (2011) relataram que o método BayesA mostrou-se superior na maioria das situações, mas nenhum dos métodos bayesianos são claramente superiores em todas as situações. Entretanto, BayesB, BayesCπ e BayesDπ apresentam a vantagem de propiciar informação sobre a arquitetura genética do caráter quantitativo e identificar as posições de QTL por modelagem da frequência de *Single nucleotide polymorphism*(SNP) não nulos. Também Mrode et al. (2010) concluíram pela superioridade do BayesA e Fast BayesB sobre o BayesB.

O método Fast BayesB foi desenvolvido por Meuwissen et al. (2009), visando diminuir o tempo de computação do método BayesB, originalmente implementado via simulação estocástica por meio de procedimento Monte Carlo Cadeia de Markov (MCMC). Esses autores derivaram um estimador não MCMC por meio de integração analítica. Esse método aproxima bem o método original e é muito mais rápido. Mrode et al. (2010) obtiveram, na prática, uma ligeira superioridade do Fast BayesB sobre o BayesB.

Os métodos BayesA e RR-BLUP em associação com um método de seleção de marcadores propiciam também informação sobre a arquitetura genética do caráter quantitativo. E essa seleção de covariáveis pode ser feita por meio da GWAS *a posteriori* (GWAS-PSE, conforme detalhado em tópico seguinte) e também pelo ordenamento do módulo dos efeitos estimados de marcadores.

Com distribuição exponencial e poucos 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 BayesA 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.

Gonzalez-Recio et al. (2008) compararam o método não paramétrico ou semi-paramétrico *Reproducing Kernel Hilbert Spaces* (RKHS) com a regressão bayesiana e RR-BLUP em termos de eficiência na seleção genômica. Concluíram que o método da regressão RKHS apresentou melhor capacidade preditiva do que os demais. Espaço de Hilbert (*Hilbert Spaces)* é um conceito muito usado em física estatística (física quântica) ou mecânica estatística (mecânica quântica) associado ao tema entropia, ou medida de desordem ou imprevisibilidade de um sistema (SALINAS, 2005). Também são emprestados da física estatística os conhecimentos da distribuição de Gibbs, usados na implementação da análise bayesiana.

Métodos de regressão com redução dimensional – regressão via quadrados mínimos parciais (PLSR) e regressão via componentes principais (PCR) – foram avaliados por Solberg et al. (2009). Concluíram que esses são mais simples e rápidos computacionalmente, porém menos acurados que o BayesB, com acurácias da ordem de 0,68 (PLSR e PCR) e 0,84 (BayesB).

Um procedimento BLASSO melhorado (IBLASSO ou *Improved Bayesian Lasso*) foi proposto por Legarra et al. (2011). O IBLASSO apresenta capacidade preditiva superior ao BLASSO e similar ao RR-BLUP-Het e BayesA com distribuições *a priori* não informativas para os efeitos aleatórios e componentes de variância.

Com base no exposto e nos resultados de literatura relatados, verifica-se que na classe dos métodos de regressão explícita, o BayesA, o LASSO bayesiano Melhorado (IBLASSO) e o RR-BLUP são os métodos favoritos quando o modelo poligênico infinitesimal se aplica. Na presença de genes de grande efeito, o método RR-BLUP necessita ser modificado de forma a permitir heterogeneidade de variância genética entre locos; isso gera o método RR-BLUP-Het. Adicionalmente, os métodos BayesA, RR-BLUP e RR-BLUP-Het necessitam ser complementados com a seleção de covariáveis por meio de alguma forma de GWAS. As variâncias genéticas de cada loco, necessárias no método RR-BLUP-Het, podem ser estimadas via os métodos BayesA (por meio de MCMC) ou IBLASSO.

O presente documento contempla os métodos BayesA, BayesB, Fast BayesB, BayesCπ, BLASSO, IBLASSO, RR-BLUP, RR-BLUP-Het, MCMC-BLUP, PLSR, e RKHS. Esses métodos propiciam, em determinadas situações, os três atributos desejáveis de acomodação da arquitetura genética do caráter, regularização da estimação e seleção de covariáveis.

Métodos de estimação penalizada

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 geralmente 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 no

caráter fenotípico y, via desequilíbrio de ligação com os genes que o controlam.

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

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

Isso implica que

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

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 $\beta \sim N(0, \sigma^2_\beta)$, $\hat{\beta}$ 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 $\sigma^2_{\beta_i}$ 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: $(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \lambda_{RR}(t)\sum_{i=1}^{n}\beta_i^2$, em que λ_{RR} é o parâmetro de penalização (associado ao *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 (medida da falta de ajuste do modelo) e o segundo termo é a penalização, a qual depende da magnitude dos coeficientes de regressão via $\sum_{i=1}^{n}\beta_i^2$. Por meio da função de penalização, um grande valor de λ cria um maior custo para β de grande valor, levando-o a encolher mais. Ocorre então a minimização da soma de quadrados dos resíduos, sujeita à restrição $\sum_{i=1}^{n}\beta_i^2 \leq t$. A solução para esse problema de otimização conduz a $\hat{\beta} = [X'X + \lambda_{RR}(t)I]^{-1}X'y$.

Outro método relacionado é o LASSO, que combina *shrinkage* (regularização) com seleção de variáveis e envolve o seguinte problema de otimização, via minimização de $(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \lambda_L\sum_{i=1}^{n}|\beta_i|$, 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. Ocorre então a minimização da soma de quadrados dos resíduos, sujeita a restrição $\sum_{i=1}^{n}|\beta_i| \leq t$. O componente $\lambda_L\sum_{i=1}^{n}|\beta_i|$

regulariza a regressão sem penalizar muito. O parâmetro de suavização λ_L controla a intensidade da regularização.

Para computação do Lasso, Tibshirani (1996) propôs o método de programação quadrática, o qual é muito complexo. A escolha do λ_L é de capital importância, pois o mesmo influencia o tamanho do grupo de marcadores selecionados. À medida que λ_L tende a zero a solução converge para método de regressão fixa via quadrados mínimos (FR-LS), ou seja, para $\hat{\beta} = (X' X)^{-1} X' y$. Nesse caso, não há seleção de covariáveis e a predição torna-se instável. Valores muito altos de λ_L reduzem muito os valores dos coeficientes de regressão. Para cômputo de λ_L de forma otimizada, Usai et al. (2009) propuseram o algoritmo da regressão de ângulo mínimo (LARS) associado a um passo de validação cruzada. O LASSO pode ser implementado também via abordagem bayesiana, em que λ_{BL} controla a precisão da distribuição *a priori* atribuída aos coeficientes de regressão.

Dois atributos importantes de um método estatístico de regressão ou modelo de predição são a **acurácia preditiva** e a **capacidade de interpretação**. O método de quadrados mínimos falha nos dois aspectos. É um método não viesado, mas pode apresentar estimativas com alta variância e, portanto, não apresenta mínimo erro quadrático médio e nem alta acurácia. O método RR apresenta pequeno viés e alta acurácia preditiva propiciada pelo *shrinkage*, o qual regulariza a estimação e melhora a estabilidade da solução. Ambos os métodos não produzem modelos interpretáveis, pois não selecionam covariáveis. Um terceiro método, denominado seleção de subconjunto de covariáveis (como o Garrote de Breiman) produz modelos interpretáveis, porém, com muita variabilidade nos resultados, pois se trata de um processo discreto. O

método Lasso foi proposto para conciliar esses dois atributos desejáveis (acurácia preditiva e capacidade de interpretação). Portanto, mantém a estabilidade da RR e produz modelos interpretáveis (pois produz alguns coeficientes que são exatamente zero) como o método de Breiman. Conforme Tibshirani (1996), os três métodos podem ser assim comparados:

a. Situação de pequeno número de grandes efeitos (controle genético por poucos genes de grandes efeitos): Garrote de Breiman é melhor, seguido por Lasso e RR.

b. Situação de moderado número de moderados efeitos: Lasso é melhor, seguido por RR e Garrote de Breiman.

c. Situação de grande número de pequenos efeitos (controle genético por muitos genes de pequenos efeitos): RR é melhor por pequena margem, seguido por Lasso e Garrote de Breiman.

Detalhes dos métodos de estimação penalizada

a. Regressão *Ridge* (RR-BLUP)

O método RR genômico foi proposto por Whittaker et al. (2000).

Função objetivo a ser minimizada:

$$\hat{\beta}_{RR} = \text{argmin} \left\{ \sum_{j}^{N} (y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 \left(+ \lambda_{RR} \sum_{i=1}^{n} \beta_i^2 \right) \right\}$$

Função de penalização, restrição ou regularização:

$$\lambda_{RR} \sum_{i=1}^{n} \beta_i^2$$

Solução para os coeficientes de regressão:

$$\hat{\beta} = [X' X + \lambda_{RR}(t)I]^{-1} X' y$$

Solução para os efeitos genéticos aditivos (a) dos indivíduos:

$$\hat{a} = X\hat{\beta} = X[X' X + \lambda_{RR}(t)I]^{-1} X' y$$

Características:

- Mantém todas as covariáveis, conduzindo a modelos complexos.

- Produz bons resultados para o caso de muitos marcadores de pequenos efeitos.

- Previne problema de multicolinearidade (que conduziria a estimativas imprecisas) entre marcadores correlacionados.

- Regressa os coeficientes de preditores correlacionados igualmente na direção de zero e de cada um.

- $\sum_{i=1}^{n} \beta_i^2$ é a norma de penalização em β.

- Quanto maior o valor de lambda (parâmetro de sintonia ou complexidade, que regula a força da penalização ou *shrinkage*), maior o encurtamento.

- Se lambda é estimado por REML, a RR torna-se BLUP e tem-se o método RR-BLUP e

$$\lambda_{RR} = \sigma_e^2 / \sigma_{ai}^2 = \sigma_e^2 / \sigma_m^2 = \sigma_e^2 / (\sigma_a^2 / n_Q) = (1-h^2)/(h^2 / n_Q) = n_Q(1-h^2)/(h^2)$$

e $h^2 = n_Q / (n_Q + \lambda_{RR})$, em que $n_Q = 2\sum_i^n p_i(1-p_i)$ ou número de QTL, onde h^2 corresponde à herdabilidade do caráter, σ_a^2 é a variância genética aditiva do caráter e σ_e^2 é a variância residual.

- Se a matriz de parentesco A for computada via informação de marcadores e utilizada no método BLUP fenotípico tradicional, tem-se o método denominado GBLUP ou BLUP genômico, que é equivalente ao RR-BLUP em termos da predição dos efeitos aditivos a. Assim, tem-se para o GBLUP:

$\hat{a} = [Z'Z + A^{-1}(\sigma_e^2 / \sigma_a^2)]^{-1} y$, em que Z é a matriz de incidência dos indivíduos e y é vetor de fenótipos corrigidos para os efeitos fixos.

$A = (XX')/[2\sum_i^n p_i(1-p_i)]$, em que p_i é a frequência de um dos alelos do loco i e X* refere-se à matriz X corrigida para suas médias em cada loco ($2p_i$).

Tem-se então a equivalência
$\hat{a} = X\,\hat{\beta} = X[X'X + \lambda_{RR}(t)I]^{-1} X'y = [Z'Z + A^{-1}(\sigma_e^2 / \sigma_a^2)]^{-1} y$.

b. LASSO

Função objetivo a ser minimizada:

$$\hat{\beta}_L = \text{argmin} \left\{ \sum_{j}^{N} (y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 \left(+ \lambda_L \sum_{i=1}^{n} |\beta_i| \right) \right\}$$

Função de penalização:

$$\lambda_L \sum_{i=1}^{n} |\beta_i|$$

Características:

- Mantém as covariáveis mais significativas e remove as demais.

- $\sum_{i=1}^{n} |\beta_i|$ é a norma de penalização em β (com base em valores absolutos de β) e induz esparsidade na solução, conduzindo a seleção de covariáveis e *shrinkage*, simultaneamente.

- $\lambda_L \sum_{i=1}^{n} |\beta_i|$ regulariza o ajuste de quadrados mínimos e regressa alguns coeficientes a zero. Essa formulação do regularizador faz com que o Lasso regresse β de forma mais forte que o RR-BLUP, conduzindo alguns coeficientes a zero.

    - Instável com dados de alta dimensão, pois não pode selecionar mais covariáveis (n) do que do que o tamanho amostral (N) e, nesse caso, seleciona arbitrariamente um membro de um grupo de covariáveis altamente correlacionadas.

    - Não possui a propriedade oráculo ou de retidão, que se refere a coeficientes não zero assintoticamente

não viesados, normalidade assintótica e seleção consistente de covariáveis à medida que N e n tendem a infinito.

- O método Lasso adaptativo foi proposto visando atingir a propriedade oráculo, mas mantém a instabilidade com dados de alta dimensão.

c. Rede elástica (EN)

Função objetivo a ser minimizada:

$$\hat{\beta}_{EN} = \text{argmin}\left\{\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2\left(+ \lambda_{EN}\left(\alpha\sum_{i=1}^{n}\beta_i^2 + (1-\alpha)\sum_{i=1}^{n}|\beta_i|\right)\right)\right\}$$

Função de Penalização:

$$\lambda_{EN}\left(\alpha\sum_{i=1}^{n}\beta_i^2 + (1-\alpha)\sum_{i=1}^{n}|\beta_i|\right) \text{ ou}$$

$$\lambda_{EN}\left(\sum_{i=1}^{n}|\beta_i|^q\right)$$

Características:

- Se α = 0, EN = LASSO ou se q = 1, EN = LASSO.

- Se α = 1, EN = RR ou se q = 2, EN = RR.

- Se $1 \leq q \leq 2$ | tem-se EN.

- α varia entre 0 e 1 e λ é maior que 0.

- Usa duas penalizações: a norma de penalização do Lasso para a seleção de covariáveis e a norma de penalização da

RR para estabilizar a solução (quando as covariáveis são altamente correlacionadas) e melhorar a predição.

- O comportamento é semelhante ao Lasso, mas é robusta à extrema colinearidade entre as covariáveis.

- Permite selecionar um número de covariáveis maior que o tamanho da amostra (N).

- Não possui a propriedade oráculo.

- O método Rede elástica adaptativa foi proposto visando atingir a propriedade oráculo do Lasso adaptativo e a robustez do método EN à extrema colinearidade entre as covariáveis (ZOU; HASTIE, 2005).

d. Regressão *Ridge* com heterogeneidade de variâncias entre locos marcadores (RR-BLUP-Het)

Solução para os coeficientes de regressão:

$$\hat{\beta} = [X' X + \lambda_{RR_h}(t)I]^{-1} X' y$$

- É similar ao RR-BLUP, mas mesmo para marcas de mesma frequência, regressa os coeficientes de regressão diferentemente na direção de zero.

- Os fatores de penalização dos marcadores no sistema de equações de modelo misto são dados pelos elementos λ_{RR_i} do vetor λ_{RR_h} , em que i refere-se ao loco i.

- Os elementos λ_{RR_i} podem ser obtidos via os métodos bayesianos ou REML e usados para cômputo do método RR-BLUP-Het.

Métodos de Estimação bayesiana (BayesA, BayesB, Fast BayesB, BayesCπ, BayesDπ)

BayesA

O método BayesA proposto por Meuwissen et al. (2001) produz resultados similares ao método BLUP com variâncias heterogêneas, pois as variâncias dos segmentos cromossômicos diferem para cada segmento e são estimadas sob esse modelo, considerando a informação combinada dos dados (função de verossimilhança) e da distribuição *a priori* para estas variâncias. Neste caso, o modelo é ajustado por meio de uma abordagem bayesiana com estrutura hierárquica em dois níveis. Os efeitos dos marcadores são assumidos como amostras de uma distribuição normal com média zero e variância de cada marcador dada por uma distribuição qui-quadrado inversa e escalonada, conforme apresentado a seguir:

$$\beta_i \mid \sigma^2_{\beta i} \sim N(0, \sigma^2_{\beta i})$$

$$\sigma^2_{\beta i} \sim \chi^{-2}(v_\beta, S^2_\beta)$$

em que v_β é o número de graus de liberdades e S^2_β é o parâmetro da escala de distribuição. Tem-se que a distribuição marginal *a priori* dos efeitos genéticos dos marcadores, $\beta_i \mid v_\beta, S^2_\beta$, tem distribuição t de *Student* univariada, ou seja, $\beta_i \mid v_\beta, S^2_\beta \sim t(0, v_\beta, S^2_\beta)$. Assim, esta formulação resulta na modelagem dos efeitos dos marcadores como amostras de uma distribuição t de *Student*.

O valor de S_β^2 pode ser derivado com base no valor esperado de uma variável aleatória com distribuição qui-quadrado invertida escalonada. Essa esperança matemática é dada por $E(\sigma^2) = \frac{S^2 v}{v-2}$. Assim, o parâmetro de escala é dado por $S^2 = \frac{E(\sigma^2)(v-2)}{v}$. Então, para os efeitos genéticos dos marcadores tem-se $E(\sigma_{\beta i}^2) = \frac{S_\beta^2 v_\beta}{v_\beta - 2}$ e $S_\beta^2 = \frac{E(\sigma_{\beta i}^2)(v_\beta - 2)}{v_\beta}$. A esperança $E(\sigma_{\beta i}^2)$ equivale a $E(\sigma_{\beta i}^2) = \frac{\sigma_a^2}{\sum_{i=1}^{n} 2p_i(1-p_i)}$. Assim, $S_\beta^2 = \frac{\sigma_a^2}{\sum_{i=1}^{n} 2p_i(1-p_i)} \frac{(v_\beta - 2)}{v_\beta}$, em que v_β = 4,012 ou 4,2, conforme Meuwissen et al. (2001), σ_a^2 é a variância genética aditiva do caráter e p_i é a frequência alélica do marcador i. Meuwissen et al. (2001) consideraram $S_\beta^2 = 0{,}002 \; ou \; 0{,}0429$. Isto descreve uma distribuição moderadamente leptocúrtica. Qualquer valor maior que 4 pode ser usado para v_β. Valores menores ou iguais a 4 tornam -se *a priori* "flat" (não informativa).

Para os efeitos residuais tem-se $E(\sigma_e^2) = \frac{S_e^2 v_e}{v_e - 2}$ e $S_e^2 = \frac{E(\sigma_e^2)(v_e - 2)}{v_e}$. A esperança $E(\sigma_e^2)$ equivale $E(\sigma_e^2) = \tilde{\sigma}_e^2$. Assim, $S_e^2 = \tilde{\sigma}_e^2 \frac{(v_e - 2)}{v_e} = \tilde{\sigma}_e^2 \frac{(4.2 - 2)}{4.2}$, em que $\tilde{\sigma}_e^2$ é um valor *a priori* de σ_e^2.

Assumido $\beta_i \sim N(0, \sigma_{\beta_i}^2)$, em que $\sigma_{\beta_i}^2$ é tomado de uma distribuição qui-quadrado invertida, segundo o enfoque bayesiano, isso implica que grande número de marcadores apresenta efeitos pequenos e poucos marcadores apresentam efeitos grandes. O uso de uma mistura de distribuições normal e qui-quadrado invertida conduz a uma distribuição t para β e, portanto, com maior pico em zero e uma cauda mais longa que a distribuição normal. Este método pode ser implementado via amostragem de Gibbs, para obtenção dessa informação combinada ou da distribuição *a posteriori* das variâncias.

Os métodos associados a modelos hierárquicos bayesianos (BayesA e B) por meio de suas formulações em termos dos hiperparâmetros propiciam variâncias específicas para cada marcador. RR-BLUP são funções lineares dos dados e regressam as estimativas com o mesmo erro padrão (mesmas frequências alélicas e tamanho amostral) pela mesma quantidade. Prioris Gaussianas conduzem a *shrinkage* homogêneo através dos marcadores. Os métodos bayesianos são funções não lineares dos dados e regressam efeitos menores mais do que os maiores, ou seja, admitem maiores herdabilidades para os maiores efeitos.

O *shrinkage* homogêneo não é desejável, pois alguns marcadores estão ligados a QTLs e outros não estão. Mas assumindo uma distribuição *a priori* t escalonada ou dupla exponencial para os efeitos de marcadores tem-se os métodos BayesA e BLASSO, respectivamente, os quais produzem *shrinkage* específicos de acordo com o tamanho do efeito e da variância do marcador.

Além das distribuições consideradas para os efeitos aleatórios no modelo linear frequentista e para a verossimilhança do vetor de observações, a abordagem bayesiana requer atribuições para as distribuições *a priori* dos efeitos e componentes de variância. Essas distribuições podem ser informativas, conforme acima, ou não informativas. Distribuição *a priori* não informativa ou uniforme pode ser atribuída a esses componentes, refletindo conhecimento *a priori* vago. Para os componentes de variância, distribuições χ^2 invertidas podem ser consideradas como *priori* e, considerando $\upsilon_i = -2$ e $S_i^2 = 0$, a distribuição χ^2 se torna uniforme e, portanto, não informativa. A vantagem de usar distribuição qui-quadrado invertida como *priori* para os componentes de variância refere-se ao fato de que, com dados com distribuição normal, a distribuição *a posteriori* é também uma qui-quadrado invertida.

Considere o seguinte modelo:

$y = 1u + X\beta + e$, onde:

y : vetor de dados fenotípicos.

u : média geral.

β : vetor de efeitos genéticos aditivos (aleatórios) de marcadores.

e : vetor de erros.

1,X : matrizes de incidência que associam *u* e *β* aos dados fenotípicos (y).

Considera-se, inicialmente, que a distribuição condicional dos dados u, β e σ_e^2 é normal multivariada: $y|\mu,\ \beta,\ \sigma_e^2 \sim N\ (1\mu + X\beta,\ I\ \sigma_e^2)$, onde I é a matriz identidade e σ_e^2 a variância residual.

Os parâmetros de interesse para inferências são: $\mu, \beta, \sigma_{\beta i}^2 \, e \, \sigma_e^2$. Para conduzir a análise bayesiana, torna-se necessário especificar as distribuições *a priori* para $\beta, \sigma_{\beta i}^2 \, e \, \sigma_e^2$. Isto já foi realizado anteriormente. Definidas estas distribuições, pode-se agora escrever a distribuição conjunta *a posteriori* dos parâmetros do modelo.

$$p\ (\mu,\ \beta,\ \sigma_{\beta i}^2,\ \sigma_e^2|y)\ \alpha\ p\ (\mu,\ \beta,\ \sigma_{\beta i}^2,\ \sigma_e^2)\ p\ (y|\mu,\ \beta,\ \sigma_{\beta i}^2,\ \sigma_e^2)$$
$$= p\ (\mu)\ p\ (\beta_i|\sigma_{\beta i}^2)\ p\ (\sigma_{\beta i}^2)\ p\ (\sigma_e^2)\ p\ (y|\mu,\ \beta,\ \sigma_{\beta i}^2,\ \sigma_e^2)$$

Considerando a distribuição *a priori* dos componentes de variância como uma qui-quadrado escalonada invertida, tem-se que a distribuição conjunta *a posteriori* pode ser reescrita:

$$p\ (\mu,\ \beta,\ \sigma_{\beta i}^2,\ \sigma_e^2|y)\ \alpha\ \sigma_e^{2-\left(\frac{N+\upsilon_e}{2}+1\right)}\ \exp\left[-\frac{(y-1\mu-X\beta)'(y-1\mu-X\beta)+\upsilon_e\ S_e^2}{2\ \sigma_e^2}\right]$$
$$\sigma_{\beta i}^{2-\left(\frac{n+\upsilon_\beta}{2}+1\right)}\ \exp\left[-\frac{(\beta'\beta+\upsilon_\beta\ S_\beta^2}{2\ \sigma_{\beta i}^2}\right]$$

Para implementação do GS, deve-se derivar todas as distribuições condicionais *a posteriori* a partir da

distribuição conjunta *a posteriori*. A distribuição condicional *a posteriori* de $\sigma^2_{\beta i}$ é dada por uma qui-quadrado invertida escalonada por $S^2_\beta + \beta_i' \beta_i$ e com graus de liberdade υ_β, ou seja $P(\sigma^2_{\beta i} \mid \beta_i) = \chi^{-2}(\upsilon_\beta, S^2_\beta + \beta_i' \beta_i)$. Não se pode usar essa distribuição *a posteriori* diretamente para estimar $\sigma^2_{\beta i}$, pois ela é condicional aos efeitos β_i que são desconhecidos. Assim, a técnica de amostragem de Gibbs, baseada em distribuições *a posteriori* condicional a todos os outros efeitos, é usada para estimar os efeitos β_i e suas variâncias.

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

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

1. Fornecer os valores iniciais dos parâmetros de locação e dispersão do modelo. Estes valores iniciais podem ser calculados através de procedimentos padrões tais como a estimação de componentes de variância por REML ou quadrados mínimos. Considerando a média geral μ como único efeito fixo, pode-se calcular μ como a média aritmética das observações. O vetor dos efeitos de marcadores deve ser inicializado com um número positivo de pequena magnitude.

2. Atualizar $\sigma^2_{\beta i}$ para o i-ésimo marcador, amostrando-

o da distribuição condicional completa $P(\sigma^2_{\beta i} \mid \beta_i) = \chi^{-2}(\upsilon_\beta, S^2_\beta + \beta_i' \beta_i)$ com $\upsilon_g = 4{,}2$ e S^2_β calculado conforme a expressão acima.

3. Dados β_i e μ, calcular os valores de e via $e = (y - 1\mu - X\beta)$, em que X = [X1 X2 X3...] é a matriz de incidência para os efeitos de marcadores. Então, atualize a variância residual por meio da amostragem de $\chi^{-2}(N-2, e_i' e_i)$.

4. Amostrar, de uma distribuição normal com média $(1_n^{'} y - 1_n^{'} X\beta)$ e variância σ^2_e / N, a média geral, dada a atualizada variância residual.

5. Amostrar, de uma distribuição com média

$$\frac{X_{ij}^{'} y - X_{ij}^{'} X\beta_{ij=0} - X_{ij}^{'} 1_n u}{X_{ij}^{'} X_{ij} + \sigma^2_e / \sigma^2_{\beta i}}$$ e variância

$\sigma^2_e / (X_{ij}^{'} X_{ij} + \sigma^2_e / \sigma^2_{\beta i})$, todos os efeitos de marcadores β_{ij} dado a amostragem mais recente da média, σ^2_e e $\sigma^2_{\beta i}$, em que X_{ij} é o vetor coluna de X com efeitos β_{ij}. No caso, $\beta_{ij=0}$ equivale a β com efeito β_{ij} igualado a zero.

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

De maneira genérica, na análise bayesiana os seguintes passos devem ser adotados: (i) especificação das distribuições *a priori* para os efeitos e componentes de

variância; (ii) especificação da função de verossimilhança para o vetor de observações (distribuição condicional dos dados): (iii) obtenção das distribuições conjuntas *a posteriori* para os efeitos e componentes de variância; (iv) obtenção das distribuições condicionais *a posteriori* para os efeitos e componentes de variância; (v) marginalização das distribuições condicionais *a posteriori* para os efeitos e componentes de variância. A marginalização analítica é praticamente impossível. Assim, têm sido usados métodos MCMC, como o amostrador de Gibbs, que atua por meio de amostragem e atualização de distribuições condicionais.

BayesB

O método BayesB apresenta as mesmas suposições que o BayesA para uma fração π dos SNPs e assume que (1 - π) dos SNPs apresenta efeitos nulos. Um problema desse método é a escolha da fração π. Com a seleção de covariáveis baseada no módulo de seus efeitos estimados, os dois métodos tendem a se equivaler. Na prática, o BayesA tem se mostrado superior ao BayesB com π igual a 0,66 (HABIER et al., 2011; MRODE et al., 2010).

Para os efeitos dos QTLs, o método BayesB usa uma distribuição *a priori* com alta densidade em $\sigma^2_\beta = 0$ e distribuição qui-quadrado invertida para $\sigma^2_\beta > 0$. Assim, considera que em muitos locos não existe variação genética, ou seja, não estão segregando. Assim, a distribuição *a priori* equivale a $\sigma^2_{\beta i} \sim \chi^{-2}(\upsilon, S^2)$ com probabilidade π e $\sigma^2_{\beta i} = 0$ com probabilidade (1 - π), em que π depende da taxa de mutação do gene. As quantidades $\upsilon = 4{,}234$ e $S^2 = 0{,}0429$ usadas por Meuwissen et al. (2001) produzem a média e variância de $\sigma^2_{\beta i}$, dado que

$\sigma^2_{\beta i} > 0$. Tais quantidades também dependem dos efeitos mutacionais e precisam ser estimadas na prática.

A distribuição *a priori* do método BayesA não tem um pico de densidade em $\sigma^2_{\beta i} = 0$. Uma vez que não é possível uma amostragem de $\sigma^2_{\beta i} = 0$, o método da amostragem de Gibbs não pode ser usado no método BayesB, pois não move sobre todo o espaço de amostragem. Assim, o algoritmo de Metropolis-Hastings deve ser usado. Esse método resolve esse problema por meio da amostragem simultânea de $\beta_i \; e \; \sigma^2_{\beta i}$. O amostrador de Metropolis-Hastings consiste em gerar amostras sequenciais como meio de aproximar uma distribuição da qual não há como amostrar diretamente. Tal amostrador pode amostrar diretamente de qualquer distribuição de probabilidade *f(x)*, desde que a densidade em *x* possa ser calculada. Detalhes da implementação desse algoritmo são apresentados por Sorensen e Gianola (2002) e Chib e Greenberg (1995).

A amostragem simultânea de $\beta_i \; e \; \sigma^2_{\beta i}$ é realizada da distribuição $P(\sigma^2_{\beta i}, \beta_i \mid y*) = P(\sigma^2_{\beta i} \mid y*).P(\beta_i \mid \sigma^2_{\beta i}, y*)$, em que *y** denota o vetor de dados corrigido para os efeitos fixos e para todos os efeitos genéticos, exceto β_i.

Essa expressão indica que se deve amostrar $\sigma^2_{\beta i}$ de $P(\sigma^2_{\beta i} \mid y*)$ sem condicionar em β_i (em contraste com o método BayesA) e em seguida amostrar β_i de $P(\beta_i \mid \sigma^2_{\beta i}, y*)$ condicional a $\sigma^2_{\beta i}$ e *y**, como no método BayesA. A distribuição $P(\sigma^2_{\beta i} \mid y*)$ não pode ser expressa na forma de uma distribuição conhecida e então deve-se usar o algoritmo MH para amostrar essa distribuição. A

distribuição *a priori* $p(\sigma^2_{\beta i})$ é usada como distribuição auxiliar para sugerir atualizações para a cadeia de MH.

Os métodos bayesianos teoricamente propiciam acurácias mais altas porque forçam muitos efeitos de segmentos cromossômicos a valores próximos a zero (BayesA) ou a zero (BayesB) e as estimativas dos efeitos dos demais segmentos cromossômicos são regressadas de acordo com uma quantidade ditada pelas distribuições *a priori* dos efeitos de QTL.

BayesCπ

Gianola et al. (2009) fazem uma análise crítica dos métodos associados a modelos hierárquicos bayesianos (BayesA 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 da *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). Segundo os autores, o método BayesB não é bem formulado no contexto bayesiano. Isto porque designar *a priori* que $\sigma^2_{\beta i} = 0$, não conduz necessariamente a $\beta_i = 0$, conforme intenção original de Meuwissen et al. (2001), em que β_i é o efeito genético do loco i. Sugere então que o estado zero seja especificado no âmbito dos efeitos e não no das variâncias. Assim, à probabilidade de mistura Π poderia ser atribuída uma distribuição *a priori* Beta. Surge então, o método

BayesC que é vantajoso e permite especificar uma distribuição *a priori* para Π, permitindo a modelagem da distribuição dupla exponencial.

Vários outros métodos bayesianos foram propostos (BayesCπ e BayesDπ, conforme Habier et al., 2011), todos eles com o propósito de permitir o aprendizado bayesiano. Habier et al. (2011) relataram que o método BayesA mostrou-se superior na maioria das situações, mas que nenhum dos métodos bayesianos são claramente superiores dentre eles; entretanto o BayesB, BayesDπ e especialmente o BayesCπ 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 não nulos.

No método BayesC uma variância comum é especificada para todos os locos. Adicionalmente, π é tratada como uma incógnita com distribuição *a priori* uniforme (0,1) caracterizando o método BayesCπ, que equivale então ao método RR-BLUP com seleção de covariáveis e implementado via MCMC. Também se π é igual a 1os métodos BayesCπ e RR-BLUP são iguais (se prioris vagas são usadas).

A modelagem de π é muito interessante para a análise de associação. A maioria das marcas não está em desequilíbrio de ligação com os genes. Assim, é necessária a seleção de um grupo de marcas que está em associação com o caráter. O método BayesB determina π subjetivamente. Usando a variável indicadora δ_i os métodos BayesCπ e BayesDπ modelam os efeitos genéticos aditivos como $a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$, em que $\delta_i = (0,1)$. A distribuição de $\delta = (\delta_1 ... \delta_n)$ é binomial com probabilidade π. Esse modelo

de mistura é mais parcimonioso do que o método BayesB. Seguindo a hierarquia do modelo, uma distribuição deve ser postulada para π e deve ser uma Beta (LEGARRA et al., 2011).

Se $\delta = 1$, não há seleção de marcas e o método torna-se o RR-BLUP implementado via MCMC (RR-BLUP bayesiano). Para o caso da distribuição Beta com parâmetros α e β, tem-se:

- Se α = 0 e β = 0: há problema na estimação, pois a distribuição Beta torna-se mal definida.

- Se α = 1 e β = 1: tem-se uma distribuição Uniforme em π.

- Se α = 1 e β = 10^{10}: tem-se π próximo de zero e a maioria das marcas terá efeito zero.

- Se α = 10^{8} e β = 10^{10}: tem-se π quase fixado em 0,01 e em torno de 1% das marcas terá efeito.

BayesDπ

O método BayesDπ mantém variâncias específicas para cada loco e modela π como uma variável aleatória. O método BayesD difere do BayesA e BayesB por considerar o parâmetro de escala das prioris qui-quadrado invertidas para as variâncias específicas para cada loco como uma incógnita com distribuição *a priori* Gama (1,1). Como o desconhecido parâmetro de escala é comum a todos os locos as informações de todos os locos contribuem para a su*a posteriori* e por meio desta para as posterioris das variâncias específicas de cada loco.

Adicionalmente, π é tratado como uma incógnita com distribuição *a priori* Uniforme (0,1) produzindo os métodos

BayesCπ e BayesDπ. Em contraste, π é igual a um no BayesA e pode ser da ordem de 0,01 no BayesB (HABIER et al., 2011).

Uma comparação entre os métodos bayesianos é apresentada na Tabela 2.

Tabela 2. Comparação entre os métodos bayesianos.

Método	Modelo para os efeitos genéticos	Parâmetros que estima	Método se π = 1
BayesDπ	$a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$	$\sigma^2_{\beta i}$, δ_i, σ^2_e, π	BayesD
BayesCπ	$a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$	σ^2_β, δ_i, σ^2_e, π	BayesC
BayesC	$a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$	σ^2_β, δ_i, σ^2_e	RR-BLUP bayesiano ($\delta_i = 1$)
BayesB	$a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i$	$\sigma^2_{\beta i}$, δ_i, σ^2_e	BayesA
BayesA	$a_j = \sum_{i=1}^{n} \beta_i x_{ij}$	$\sigma^2_{\beta i}$, σ^2_e	-
RR-BLUP	$a_j = \sum_{i=1}^{n} \beta_i x_{ij}$	σ^2_β, σ^2_e	-

Fast BayesB

O método Fast BayesB foi desenvolvido por Meuwissen et al. (2009) visando diminuir o tempo de computação do método BayesB. Esses autores derivaram um algoritmo de esperança condicional iterativa (ICE) para estimar β_i por meio de integração analítica. Os seguintes passos devem ser adotados.

a) Calcular as observações ajustadas, y_{-i}, que são corrigidas para os efeitos de todos os outros marcadores, usando a expressão $\hat{y}_{-i} = y - \sum_{j \neq i}^{n} x_j \hat{\beta}_j$.

   Estimar a estatística suficiente

   $\hat{Y}_i = (x_i' y - \sum_{j \neq i}^{n} (x_i' x_j) \hat{\beta}_j) / N$ e $\sigma^2 = \sigma_e^2 / N$.

b) Calcular $\hat{\beta}_i = E[\beta_i | Y_i]$, que é usado para atualizar a solução para o marcador i. A expressão para cômputo de $\hat{\beta}_i = E[\beta_i | Y_i]$ usa a função Delta Dirac e é apresentada por Meuwissen et al. (2009).

A natureza aproximada do algoritmo ICE é devida ao fato de y_{-i} e Y_i não serem conhecidos e sim serem estimados. Erros de estimação em $\hat{y}_{-i}$ e $\hat{Y}_i$ ocorrem devido a erros de estimação nos efeitos $\hat{\beta}_j$ dos outros marcadores.

Lasso bayesiano e Lasso bayesiano Melhorado (BLASSO e IBLASSO)

Os Lassos bayesianos são vantajosos em relação aos métodos bayesianos de Meuwissen et al. (2001) por serem assintoticamente livres de informação *a priori*. O parâmetro λ pode ser estimado dos próprios dados pelos métodos MCMC (esse algoritmo pode ser implementado usando informação *a priori* vaga) e MCEM (esse algoritmo EM não requer informação *a priori*). Os métodos BayesA e BayesB requerem a designação de distribuições *a priori* para a variância de cada marcador. Adicionalmente alguns métodos bayesianos requerem a estimação de π. Nos Lassos não existe π e uma distribuição controlada por λ é declarada para toda a coleção de variâncias dos locos marcadores.

No método Lasso original, uma moda conjunta é estimada e espera-se que a maioria dos marcadores tenham efeitos exatamente igual a zero (USAI et al., 2009). No Lasso bayesiano são estimadas médias *a posteriori*, produzindo valores muito pequenos, mas não zero. E médias *a posteriori* são o critério ótimo para seleção (LEGARRA et al., 2011). No Lasso original a solução admite até (N-1) coeficientes de regressão não nulos, em que N é o número de indivíduos. O Lasso bayesiano relaxa essa restrição, possivelmente produzindo um modelo mais acurado.

A formulação bayesiana do Lasso (BLASSO) inclui um termo de variância comum para modelar ambos os termos, os resíduos e os efeitos genéticos dos marcadores (PARK; CASELLA, 2008; CAMPOS et al., 2009b). Legarra et al. (2011) propuseram o método BLASSO melhorado (IBLASSO), o qual usa dois termos de variância, um para modelar os resíduos e outro para modelar os efeitos

genéticos dos marcadores. Esses termos se adequam aos conceitos de variação endógena e exógena no contexto dos modelos mistos, conforme Singer et al. (2011). Isso também é coerente com a teoria da genética quantitativa, que preconiza a decomposição da variação fenotípica em variação genética e residual.

Uma comparação entre os três métodos Lassos, o RR-BLUP e o RR-BLUP-Het é apresentada na Tabela 3.

Tabela 3. Características dos três métodos Lassos.

Método	Modelo	Variância de cada marcador	Variância genética aditiva	Parâmetro de forma
LASSO	$y = 1u + X\beta + e$ $e\|\sigma_e^2 \sim MVN(0, I\sigma_e^2)$ $p(\beta\|\sigma_e^2 = 1, \lambda) = (\lambda/2)\exp(-\lambda\|\beta\|)$ $\beta\|\lambda \sim \prod_i (\lambda/2)\exp[(-\lambda\|\beta_i\|)$	-	-	-
BLASSO	$y = 1u + X\beta + e$ $e\|\sigma^2 \sim MVN(0, I\sigma^2)$ $p(\beta\|\sigma^2, \lambda) = (\lambda/2\sigma)\exp[(-\lambda\|\beta\|)/\sigma]$ $p(\beta\|\tau) \sim N(0, D\sigma^2); diag(D) = \tau_1^2 ... \tau_n^2);$ $p(\tau\|\lambda) = \prod_i (\lambda^2/2)\exp(-\lambda^2\tau_1^2/2).$	$Var(\beta) = (2\sigma_e^2)/\lambda^2$ $Var(\beta_i) = \sigma_{\beta i}^2 = \tau_i^2\sigma^2$	$\sigma_a^2 = \sum_{i=1}^{m} 2p_i(1-p_i)(2\sigma_e^2)/\lambda^2$	$\lambda^2 = (2\sigma_e^2)/\sigma_\beta^2$
IBLASSO	$y = 1u + X\beta + e$ $e\|\sigma_e^2 \sim MVN(0, I\sigma_e^2)$ $\beta\|\lambda, \sigma_\beta^2 \sim \prod_i (\lambda/2\sigma_\beta)\exp[(-\lambda\|\beta_i\|)/\sigma_\beta]$ $p(\beta \mid \tau) \sim N(0, D); diag(D) = (\tau_1^2 ... \tau_n^2)$ $p(\tau \mid \lambda) = \prod_i (\lambda^2/2)\exp(-\lambda^2\tau_i^2/2)$	$Var(\beta) = 2/\lambda^2$ $Var(\beta_i) = \sigma_{\beta i}^2 = \tau_i^2$	$\sigma_a^2 = \sum_{i=1}^{m} 2p_i(1-p_i)\,2/\lambda^2$	$\lambda^2 = 2/\sigma_\beta^2$
RR-BLUP	$y = 1u + X\beta + e$ $e \mid \sigma_e^2 \sim MVN(0, I\sigma_e^2)$ $\beta \mid \sigma_\beta^2 \sim MVN(0, I\sigma_\beta^2)$	$Var(\beta) = \sigma_\beta^2$	$\sigma_a^2 = \sum_{i=1}^{m} 2p_i(1-p_i)\,\sigma_\beta^2$	$\lambda^2 = (\sigma_e^2/\sigma_\beta^2)^2$

Tabela 3. Continuacão.

Método	Modelo	Variância de cada marcador	Variância genética aditiva	Parâmetro de forma
RR-BLUP-Het	$y = 1\mathrm{u} + X\beta + e$ $e \mid \sigma_e^2 \sim MVN(0, I\sigma_e^2)$ $\beta \mid \lambda, \tau \sim MVN(0, D)$	$Var(\beta_i) = \sigma_{\beta i}^2 = \tau_i^2$	-	-

IBLASSO

A parametrização do IBLASSO é equivalente ao do LASSO original de Tibshirani (1996), porém, a implementação é bayesiana. Outra diferença refere-se ao fato de que a parametrização do LASSO original assume que a matriz de incidência X foi padronizada. O IBLASSO não assume isso. Essa diferença pode ser observada na descrição dos modelos apresentada na Tabela 3. A igualdade na parametrização advém da comparação entre os termos $(\lambda/2\sigma_\beta)$ e $(\lambda/2)$. Somente a proporção (λ/σ_β) é utilizada na prática e, portanto, λ e σ_β não podem ser estimados separadamente. Assim, o λ de Tibshirani equivale a (λ/σ_β) do IBLASSO e é, essencialmente, uma medida da variação genética dos marcadores na população. De forma equivalente, o modelo do IBLASSO poderia ser escrito em termos de σ^2_β, retirando λ.

A forma da distribuição dos efeitos das marcas é determinada pelo parâmetro de forma λ, que é relacionado à variação genética dos marcadores por meio da expressão $Var(\beta) = 2/\lambda^2$.Essa relação denota que λ^2 desempenha papel similar ao inverso da variância nos modelos sob normalidade. O parâmetro λ pode ser estimado por MCMC ou máxima verossimilhança marginal (MCEM ou REML). A estimação por MCEM evita o uso de super-priori para λ (PARK; CASELLA, 2008).

Partindo-se da relação $\sigma^2_a = \sum_{i=1}^{m} 2p_i(1-p_i)\sigma^2_\beta$ (GIANOLA et al., 2009), tem-se $\sigma^2_a = \sum_{i=1}^{m} 2p_i(1-p_i)\,2/\lambda^2$, em que σ^2_a é a variância genética aditiva. Uma vez que a variância

genética aditiva do caráter é geralmente conhecida *a priori* (de outros estudos), uma informação *a priori* para λ pode ser dada por $\lambda^2 = \sum_{i=1}^{m} 2p_i(1-p_i)\,2/\sigma_a^2$. Entretanto, nos modelos hierárquicos bayesianos propriamente ditos (caso dos Lassos bayesianos e não dos métodos bayesianos de Meuwissen) informação *a priori* é atribuída aos hiperparâmetros (λ e componentes de variância, por exemplo) de forma que a influência dessa informação desaparece assintoticamente.

O modelo genérico do Lasso é da forma:

$$y = 1u + X\beta + e$$
$$e \mid \sigma^2 \sim MVN(0, I\sigma^2)$$
$$p(\beta \mid \sigma^2, \lambda) = (\lambda / 2\sigma)\exp[(-\lambda|\beta|)/\sigma]$$

Essa distribuição exponencial do Lasso para β coaduna bem com a distribuição observada para os efeitos genéticos dos locos de um caráter quantitativo(GODDARD, 2009).

Com dois componentes de variância (σ_e^2 e σ_β^2) o modelo torna-se:

$$y = 1u + X\beta + e$$
$$e \mid \sigma_e^2 \sim MVN(0, I\sigma_e^2)$$
$$\beta \mid \lambda, \sigma_\beta^2 \sim \prod_i (\lambda / 2\sigma_\beta)\exp[(-\lambda|\beta_i|)/\sigma_\beta]$$

Notando-se a equivalência com o modelo de Tibshirani, tem-se:

$$\beta \mid \lambda \sim \prod_i (\lambda / 2) \exp[(-\lambda |\beta_i|)$$

Usando uma formulação em termos de um modelo hierárquico aumentado, incluindo um componente de variância extra τ_i^2 associado a cada loco marcador, tem-se:

$$p(\beta \mid \tau) \sim N(0, D); diag(D) = \tau_1^2 ... \tau_n^2)$$
$$p(\tau \mid \lambda) = \prod_i (\lambda^2 / 2) \exp(-\lambda^2 \tau_i^2 / 2)$$

Assim, tem-se: $Var(\beta_i) = \sigma_{\beta i}^2 = \tau_i^2$

A implementação prática desse modelo via amostrador de Gibbs é apresentada a seguir, conforme Legarra et al. (2011).

A distribuição *a priori* de σ_e^2 consiste de uma qui-quadrado invertida com 4 graus de liberdade. A distribuição *a priori* para λ pode ser deliberadamente vaga, como uma uniforme entre 0 e 1.000.000.

As distribuições condicionais *a posteriori* completas são apresentadas a seguir.

$$u \mid demais \propto N(1'(y - X\tilde{\beta})/1'1, 1/1'1\tilde{\sigma}_e^{-2})$$

$\beta_i \mid demais \propto N(x_i'(y - 1_i\tilde{\mu} - X\tilde{\beta}_{-i})\tilde{\sigma}_e^{-2} / LHS_I, 1/LHS_I)$, em que $LHS_i = x_i' x_i \tilde{\sigma}_e^{-2} + \tau_i^{-2}$ e x_i é a linha de X correspondente ao efeito i e $\tilde{\beta}_{-i}$ indica todas as variáveis $\tilde{\beta}$, exceto $\tilde{\beta}_i$.

$\tau_i^{-2} \mid demais \propto IG\left((\tilde{\lambda}^2 / \beta_i^2)^{1/2}, \lambda^2\right)$, em que IG refere-se a Gama Invertida.

$\lambda^2 \mid demais \propto G\left(m, 2/\sum \tilde{\tau}_i^2)\right)$, em que G refere-se a Gama com parâmetro de forma igual ao número m de marcas e parâmetro de escala igual a $2/\sum \tilde{\tau}_i^2)$.

$\sigma_e^2 \mid demais \propto \chi^{-2}\left(\tilde{e}'\tilde{e} + S_e^2, 4 + N\right)$, em que N é o número de indivíduos e S_e^2 é a escala da distribuição *a priori* da variância residual.

BLASSO

O modelo é da forma

$$y = 1u + X\beta + e$$
$$e \mid \sigma^2 \sim MVN(0, I\sigma^2)$$
$$\beta \mid \lambda, \sigma^2 \sim \prod_i (\lambda / 2\sigma)\exp[(-\lambda|\beta_i|)/\sigma]$$

Usando uma formulação em termos de um modelo hierárquico aumentado tem-se:

$$p(\beta|\tau) \sim N(0, D\sigma^2); diag(D) = \tau_1^2 ... \tau_n^2)$$
$$p(\tau|\lambda) = \prod_i (\lambda^2/2)\exp(-\lambda^2\tau_1^2/2)$$

Assim, tem-se que a variância genética em cada loco marcador é dada por $\sigma_{\beta i}^2 = \tau_i^2\sigma^2$.

As distribuições condicionais *a posteriori* completas são conforme descrito para o IBLASSO, porém com as seguintes modificações:

$$LHS_i = x_i' x_i \tilde{\sigma}_e^{-2} + \tau_i^{-2}\sigma^{-2}$$

$$\tau_i^{-2} \mid demais \propto IG\left((\tilde{\lambda}^2\sigma^2 / \beta_i^2)^{1/2}, \lambda^2\right)$$

$$\sigma^2 \mid demais \propto \chi^{-2}\left(\tilde{\beta}' \tilde{D}^{-1}\sigma^2\tilde{\beta} + \tilde{e}'\tilde{e} + S_e^2, 4 + m + N\right)$$

Essa última distribuição condicional mostra que os efeitos de marcadores são na prática considerados como pseudo resíduos no BLASSO.

GBLUP com heterogeneidade de variâncias

O método GBLUP ou BLUP genômico pode também ser implementado considerando a heterogeneidade de variância entre marcadores. Nesse caso, a matriz A é dada por $A = (X^* DX^{*\prime})/[2\sum_i^n p_i(1-p_i)]$, em que p_i é a frequência de um dos alelos do loco i e X* refere-se à matriz X corrigida para suas médias em cada loco ($2p_i$). A matriz D é dada por $diag(D) = (\tau_1^2 ... \tau_n^2)$ e os elementos τ_i^2 podem ser obtidos pelos métodos IBLASSO, BLASSO, BayesA, BayesB, etc. Essa abordagem apresenta também os seguintes pontos favoráveis: (i) permite a análise simultânea de indivíduos genotipados e não genotipados; (ii) permite o cômputo direto da acurácia seletiva via inversão da matriz dos coeficientes das equações de modelo misto; (iii) a matriz D pode ser estimada em apenas uma amostra da população e ser usada em toda a população de seleção e em várias gerações.

Comparação entre distribuições assumidas para os efeitos genéticos nos diferentes métodos

Na Tabela 4 são apresentadas as distribuições assumidas para os efeitos genéticos de marcadores nos diferentes métodos de GWS.

Tabela 4. Distribuições assumidas para os efeitos genéticos de marcadores nos diferentes métodos de GWS.

Método	**Distribuição *a priori* dos efeitos**	**Distribuição *a priori* das variâncias**	**Distribuição a *posteriori***
RR-BLUP (bayesiano)	Normal com variância comum	qui-quadrado invertida não informativa	qui-quadrado invertida
BayesA	Normal com heterogeneidade de variâncias entre marcas (t dado priori qui-quadrado para as variâncias)	qui-quadrado invertida (equivale ao BayesB com $\pi = 1$)	qui-quadrado invertida
BayesB	Normal com heterogeneidade de variâncias entre marcas, média zero e variância finita (t dado priori qui-quadrado para as variâncias)	Mistura de distribuições 0 com probabilidade $(1-\pi)$ e qui-quadrado invertida com probabilidade π	qui-quadrado invertida
BayesCπ	Mistura de distribuições 0 e normal com variância comum (t dado priori qui-quadrado para as variâncias)	qui-quadrado invertida, π com distribuição Uniforme entre 0 e 1	
Lassos	Exponencial Dupla	Exponencial Dupla	Gama Invertida

A Figura 1 ilustra as formas das distribuições normal (RR-BLUP) e exponencial (LASSO).

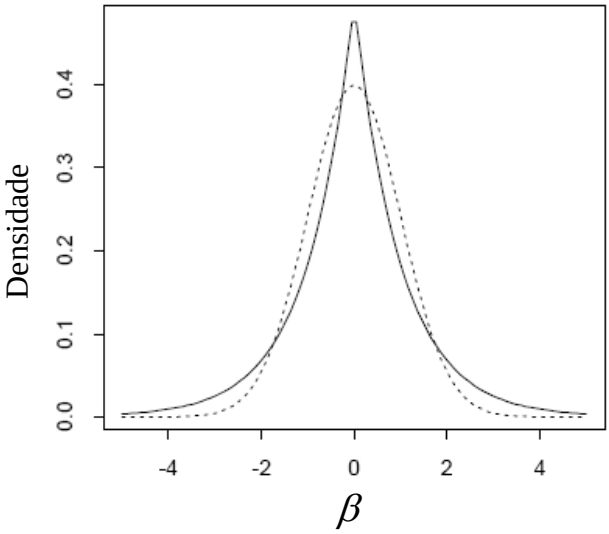

Figura 1. Densidades das distribuições normal (curva pontilhada) e exponencial dupla (curva cheia), ambas com médias iguais a zero e variâncias iguais à unidade.

Observa-se que a densidade *a priori* utilizada no LASSO Bayesiano apresenta maior massa de densidade no valor zero e caudas mais robustas, exercendo maior encurtamento sobre coeficientes de regressão próximos de 0 e menor encurtamento sobre coeficientes de regressão distantes de zero.

Regressão Kernel Hilbert Spaces (RKHS)

Os métodos regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA et al., 2006), regressão semi-paramétrica RKHS (*Reproducing Kernel Hilbert Spaces*) (GIANOLA; KAAM, 2008) e de redes neurais pertencem à classe de regressão implícita e são métodos não paramétricos ou semi-paramétricos. Esses métodos são uma alternativa para o ajuste de modelos com muitas interações epistáticas e de dominância.

Gonzalez-Recio et al. (2008) compararam métodos não paramétricos (RKHS), regressão bayesiana e RR-BLUP 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*. O método semi-paramétrico RKHS parece ter maior capacidade preditiva quando aplicado a dados reais (GIANOLA et al., 2009), sem fazer fortes suposições *a priori*.

Regressões não paramétricas são representações funcionais entre um grande número de covariáveis e uma variável dependente, gerando uma estrutura menos parametrizada, com menos suposições e com facilidade para acomodar efeitos de interações.

As funções de *kernel* podem ser usadas em métodos não paramétricos para estimar densidades a partir de uma amostra (BISHOP, 2006). A regressão de Naradaya-Watson (NWR) aplicando o *kernel* binomial para estimação da função do valor alélico tem sido usada para implementação do modelo não paramétrico usando a teoria do modelo

aditivo (HASTIE; TIBSHIRANI, 1986; GIANOLA et al., 2006). Este método apresenta resultado similar ao do RR-BLUP, sendo que o NWR depende do fator de alisamento e o RR-BLUP depende do fator de *shrinkage*.

RKHS

Modelo

O modelo genérico para o fenótipo é dado por $y_j = u + g(x_j) + e_j$, em que: y_j é o fenótipo do indivíduo j; u é a média do caráter em estudo; e_j é o erro aleatório e $g(x_j)$ é uma função desconhecida que relaciona os genótipos marcadores (covariáveis) com os fenótipos (variável dependente).

A função g(x) é definida por $g(x) = E(y|x) = \frac{\int_{-\infty}^{\infty} y\, p(y,x)\, dy}{p(x)}$.

Função objetivo a ser minimizada:

$$\hat{\beta}_{RKHS} = \arg\min \left\{ \sum_{j}^{N} [(y_j - u - g(x_i)]^2 + h\left\|g(x)\right\|_H^2 \right\}.$$

Função de penalização

$h\left\|g(x)\right\|_H^2$, em que h é o parâmetro de suavização e $\left\|g(x)\right\|_H^2$ é a norma de g(x) em um espaço de Hilbert, a qual induz regularização, cuja força é ditada por h.

Características

No espaço infinito de Hilbert, procura-se a função g(x) que minimize a soma de quadrados penalizada

$SS[g(x)] = \left\{ \sum_{j}^{N} [(y_j - u - g(x_i)]^2 + h \left\| g(x) \right\|_H^2 \right\}$. A solução para essa minimização é dada por:

$g(x) = \alpha_0 + \sum_{j=1}^{N} \alpha_j k(x - x_i)$, em que α_j são coeficientes desconhecidos (com total equivalente ao número N de indivíduos genotipados) e $k(x-x_j)$ é o *kernel* de reprodução, cuja escolha define o espaço de Hilbert em que se dará a minimização da soma de quadrados. A regularização realizada produz nos modelos de regressão RKHS um menor número de parâmetros do que em outros métodos.

Na RKHS uma coleção de funções reais é implicitamente definida pela escolha de um *kernel* de reprodução, $k(x_i,x_j)$. Esta função mapeia pares de genótipos em números reais. Sob uma perspectiva bayesiana o *kernel* de reprodução define correlações *a priori* entre as avaliações da função (valores genéticos) em pares de genótipos $(Cor[g(x_i),g(x_j)]$. A escolha do *kernel* é fundamental na especificação do modelo e a RR pode ser representada como regressões RKHS. De maneira geral, os *kernels* são escolhidos por algoritmos de forma a maximizar a performance do modelo, maximizando a capacidade preditiva. Uma grande variedade de *kernels* é avaliada e é selecionado aquele que é ótimo segundo o critério de seleção do modelo (aquele que maximiza a capacidade preditiva) (CAMPOS et al., 2009a). A capacidade preditiva na população de validação é a capacidade de prever futuras observações. Na população de estimação é uma medida da qualidade do ajustamento entre os dados de treinamento e o modelo.

Na regressão RKHS a estrutura de covariância é proporcional a uma matriz de *kernel* K, dada por $Cov(g_i,g_j) \propto K_{RKHS}(x_i,x_j)$, em que x_i,x_j são vetores de genótipos

marcadores para os indivíduos i e j, e K(.,.) é uma função positiva definida avaliada nos genótipos marcadores. Uma grande vantagem da RKHS é que o modelo é representado em termos de N incógnitas, fato que é uma grande vantagem computacional quando n é muito maior que N.

Nos modelos de regressão explicita e na RKHS, as funções base (funções das covariáveis usadas para construir a regressão, por exemplo, polinômios) para regressar fenótipos em marcadores são definidas *a priori* e isto impõe restrições nos padrões que podem ser capturados pelos métodos. No método de redes neurais as funções base usadas são inferidas dos próprios dados e isso confere grande flexibilidade a esse método. Porém, há o risco de superparametrização e a interpretação dos parâmetros não é trivial. A superparametrização significa que a capacidade preditiva na população de estimação apresenta boa performance mas não a apresenta na população de validação (em dados que não foram usados para ajustar o modelo) (CAMPOS et al., 2009a; 2009b).

O modelo pode então ser expandido da seguinte forma:

$$y_j = u + g(x_j) + e_j$$

$$y_j = u + \sum_{j=1}^{N} \alpha_j k(x - x_i) + e_j \text{ , em que } \alpha_0 \text{ faz parte de u.}$$

Em termos vetoriais, tem-se:

$$y = 1\mathrm{u} + T(h)\alpha + e \text{, em que:}$$

$$T(h) = \begin{bmatrix} t_1(h) \\ t_2(h) \\ . \\ . \\ . \\ t_n(h) \end{bmatrix}, \; t_i(h) =_1 [k_h(x_i - x_1)\; k_h(x_i - x_2)...k_h(x_i - x_n)]_n$$

e $\alpha' =_1 [\alpha_1\; \alpha_2\; ...\alpha_n]_n$

Assumindo $\alpha_j \sim N(0, \sigma_\alpha^2)$ e que os componentes de variância e h são conhecidos, têm-se as equações de modelo misto para obtenção das soluções de u e α_j :

$$\begin{bmatrix} 1'1 & T(h)'1 \\ T(h)1' & T(h)'T(h) + I\frac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 1'y \\ T(h)'y \end{bmatrix}$$

Após a escolha do parâmetro de suavização h, pode-se obter estimativas REML para os componentes de variância σ_α^2 e σ_e^2 . O parâmetro de suavização h pode ser determinado via validação cruzada ou via abordagem bayesiana, atribuindo-se distribuições *a priori* próprias para todos os parâmetros do modelo (GIANOLA; CAMPOS, 2009).

O modelo KRHS pode ser também assim especificado: $y = 1\mathrm{u} + K_h\alpha + e$, em que u é uma constante, K_h é a matriz positiva definida de *kernels*, dependente do parâmetro de suavização h; α é um vetor contendo coeficientes não paramétricos que são assumidos com distribuição normal $\alpha_j \sim N(0, K_h^{-1}\sigma_\alpha^2)$, com σ_α^2 representando a recíproca do

parâmetro de alisamento $(\sigma_\alpha^2 = \lambda^{-1})$. Os resíduos têm distribuição normal com matriz de covariância R=I σ_e^2. A solução para α é dada por $[\sigma_e^{-2}K_h + \sigma_\alpha^{-2}I]\hat{\alpha} = \sigma_e^{-2}y$.

Os fenótipos são preditos por $\hat{y} = \hat{u}1 + K_h^*\hat{\alpha}$, onde uma linha de K_h^* tem a forma $K_t^* = [K_h^*(x_i - x_j)]$, com $K_h^*(x_i - x_j)$ sendo o *kernel* entre o genótipo do indivíduo i no grupo de validação e o genótipo do indivíduo j no grupo de estimação.

RKHS com efeito poligênico

Nesse caso, o efeito genético de um indivíduo j é dado pelo modelo $g_j = p_j + \alpha_j$, em que p_j é a regressão sobre o pedigree e α_j é a regressão semi-paramétrica sobre os marcadores. Na RKHS, a suposição é de que $\alpha = (\alpha_1, \alpha_2, ...\alpha_n)'$ é um processo gaussiano com média nula e função de covariância proporcional a um *kernel* de reprodução, $K_{RKHS}(x_i, x_j)$, avaliada nos genótipos marcadores, em que x_i e x_j são vetores de genótipos marcadores para os indivíduos i e j.

A distribuição a priori conjunta de p, α e componentes de variância associados σ_p^2, σ_α^2 e σ_e^2 é dada por:

$$p\ (u,\ \alpha,\ p,\ \sigma_\alpha^2,\ \sigma_p^2,\ \sigma_e^2 | df_e, S_e, df_\alpha, S_\alpha, df_p, Sf_p)\ \alpha\ N(\alpha | 0, K_{RKHS}\sigma_\alpha^2) N(p | 0, A\sigma_p^2)$$
$$x\ \ \chi^{-2}(\sigma_e^2 | df_e, S_e)\chi^{-2}(\sigma_\alpha^2 | df_\alpha, S_\alpha)\chi^{-2}(\sigma_p^2 | df_p, S_p)$$

Qualquer função positiva definida satisfazendo $\sum_i \sum_j \alpha_i \alpha_j K_{RKHS}(x_i, x_j)$ para todas as sequências não nulas $\{a_i\}$ é uma escolha válida de *kernel*. Pode-se escolher $K_{RKHS}(x_i, x_j)$ como um *kernel* Gaussiano

$K_{RKHS}(x_i, x_j) = \exp\{-2(d_{ij} / q_{0,5})\}$, em que $d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ é o quadrado da distância euclidiana, e $q_{0,5}$ é a mediana amostral da matriz de quadrados das distâncias euclidianas amostrais $\{d_{ij}\}$.

Combinando a distribuição *a priori* conjunta com a função de verossimilhança, a distribuição condicional completa do modelo torna-se (CROSSA et al., 2010):

$$p\ (u,\ \alpha,\ p,\ \sigma_\alpha^2,\ \sigma_p^2,\ \sigma_e^2 | \bar{y}, H\ \alpha \left\{ \prod_{i=1}^{n} N\left(\bar{y}_i \middle| u + \alpha_j + p_j, \sigma_e^2 / n_j\right) \right\} N(\alpha | 0, K_{RKHS}\sigma_\alpha^2) N(p | 0, A\sigma_p^2)$$
$$x\ \ \chi^{-2}(\sigma_e^2 | df_e, S_e) \chi^{-2}(\sigma_\alpha^2 | df_\alpha, S_\alpha) \chi^{-2}(\sigma_p^2 | df_p, S_p)$$

Amostras são retiradas dessa distribuição.

Um modelo sem o efeito poligênico pode ser ajustado removendo p_j das equações acima. Assim, as distribuições a seguir são dadas por:

a priori:

$$p\ (u, \alpha,\ \sigma_\alpha^2,\ \sigma_p^2,\ \sigma_e^2 | df_e, S_e, df_\alpha,\ S_\alpha, df_p,\ Sf_p)\ \alpha\ N(\alpha | 0, K_{RKHS}\sigma_\alpha^2) \chi^{-2}(\sigma_e^2 | df_e, S_e)$$
$$x\ \ \chi^{-2}(\sigma_\alpha^2 | df_\alpha, S_\alpha) \chi^{-2}(\sigma_p^2 | df_p, S_p)$$

e *a posteriori:*

$$p\ (u,\ \alpha,\ \sigma_\alpha^2,\ \sigma_p^2,\ \sigma_e^2|\bar{y}, H\ \alpha \left\{\prod_{i=1}^{n} N\left(\bar{y}_j \middle| u+\alpha_j, \sigma_e^2/n_i\right)\right\} N(\alpha|0, K_{RKHS}\sigma_\alpha^2)$$
$$x\ \ \chi^{-2}(\sigma_e^2|df_e, S_e)\chi^{-2}(\sigma_\alpha^2|df_\alpha, S_\alpha)\chi^{-2}(\sigma_p^2|df_p, S_p)$$

O modelo animal univariado tradicional pode também ser expresso em termos de $y = g + e$ em que $g|0, K_{RKHS}\sigma_\alpha^2 \sim N(0, K_{RKHS}\sigma_\alpha^2)$, conduzindo ao estimador $[\sigma_e^{-2} I + \sigma_\alpha^{-2} K_{RKHS}^{-1}]\ \hat{g} = \sigma_e^{-2} y$ (CAMPOS et al., 2009a).

Regressão via quadrados mínimos parciais (PLSR)

A regressão via quadrados mínimos parciais (PLSR) é um método de redução dimensional que pode ser aplicado à seleção de marcadores com efeitos significativos em um caráter. É um método muito usado em quimiometria na situação em que se tem um grande número de variáveis com relações desconhecidas e o objetivo é a construção de um bom modelo preditivo para a variável resposta (WOLD et al., 2001). No PLS variáveis latentes são extraídas como combinações lineares das variáveis originais e são usadas para a predição da variável resposta, conforme descrito a seguir.

$y_j = f(x_j) + e_j$: valor fenotípico do indivíduo j.

$f(x_j)$: função que relaciona genótipos marcadores aos fenótipos.

e_j : termo residual.

Pelo PLS, a função $f(x_j)$ é definida como $f(x_j) = \sum_{l=1}^{h} t_{jl}\beta_l$, em que t_{jl} é o componente latente l (l = 1, 2..., h) no individuo j e geralmente h é menor que o número de variáveis. β_l é o efeito genético associado ao componente latente l. O efeito genético (regressão) associado ao marcador i é dado por $\beta_i = \sum_{l=1}^{h} \beta_l x_{li}$.

As variáveis latentes são componentes ortogonais, o que elimina o problema de multicolinearidade e a PLSR é similar à regressão via componentes principais (PCR). Ambos os métodos constroem a matriz T de componentes latentes, como transformação linear da matriz X das variáveis originais por meio de T = XW, em que W é uma matriz de pesos. A diferença é que a PCR extrai componentes que explicam a variância de X e a PLSR extrai componentes que têm maior covariância com y. Na PLSR as colunas de pesos na matriz W são definidas de forma que o quadrado da matriz de covariância amostral entre y e os componentes latentes é maximizado sob a restrição de que os componentes latentes sejam não correlacionados.

Existem diferentes técnicas para extração dos componentes latentes. A complexidade ótima do modelo, ou seja, o número de componentes latentes, pode ser determinada por validação cruzada.

Relação entre RR-BLUP, BLASSO e IBLASSO

Resultados práticos têm revelado que a capacidade preditiva não varia muito com o valor de λ_{RR} e λ_L associados à herdabilidades entre 5% e 95%, quando o número de locos é grande (SILVA et al., 2011).

Em presença de genes maiores, o RR-BLUP difere consideravelmente do BLASSO e IBLASSO. Nesse caso, o IBLASSO e o RR-BLUP-Het são melhores. O IBLASSO é similar ao BayesA mas com maior *shrinkage* nas marcas de menor efeito, conforme discutido em tópicos anteriores.

Em termos de ordenamento dos candidatos à seleção, têm-se as seguintes tendências. Com seleção indireta de covariáveis nos métodos que não o fazem diretamente: (i) BayesA é igual a BayesB; (ii) RR-BLUP é igual ao Lasso em *ranking*, desde que a arquitetura genética seja homogênea; (iii) RR-BLUP é igual ao BayesA e BayesB, desde que a arquitetura genética seja homogênea e as *prioris* utilizadas nos métodos bayesianos sejam não informativas; (iv) Com arquitetura genética heterogênea, RR-BLUP-Het é similar ao IBLASSO em *ranking*; (v) RR-BLUP é igual ao BayesCπ desde que as prioris utilizadas no método bayesiano sejam não informativas; (vi) RR-BLUP é igual ao BayesDπ, desde que a arquitetura genética seja homogênea e as *prioris* utilizadas no método bayesiano sejam não informativas. Se $\pi = 1$, RR-BLUP é igual ao BayesCπ.

RR-BLUP e Lasso podem ser implementadas sob o enfoque frequentista e bayesiano. Se *prioris* não informativas forem utilizadas, tem-se que RR-BLUP frequentista é semelhante ao RR-BLUP bayesiano e Lasso frequentista é semelhante ao Lasso bayesiano.

A seleção indireta de covariáveis no RR-BLUP usando os maiores módulos dos efeitos estimados dos marcadores produz o método RR-BLUP_B (RESENDE et al., 2010; RESENDE JUNIOR et al., 2012), o qual pode apresentar acurácia superior. Mas esse método e também o RR-BLUP tradicional dividem toda a variação genética aditiva do caráter por uma função do número de marcadores ajustados. E os marcadores usados não capturam toda essa variação genética. No RR-BLUP_B maior variação genética é atribuída a cada marcador do que de fato deveria. Assim, o RR-BLUP_B deve usar somente a variação genética capturada pelos marcadores ajustados em cada análise. Portanto, deve-se usar o REML para estimar essa variação ou outro método bayesiano, como o BLASSO ou IBLASSO, produzindo o método REML/RR-BLUP_B ou BLASSO/RR-BLUP_B ou IBLASSO/RR-BLUP_B. Também, a escolha do melhor modelo REML/RR-BLUP_B deve basear-se na validação cruzada.

Relação entre RR-BLUP e BLASSO

Considerando todos os locos que controlam o caráter:

$$\lambda_{BLUP} = \sigma_e^2 / \sigma_a^2$$

Considerando cada loco i:

$$\lambda_{RR} = \sigma_e^2 / \sigma_{ai}^2 = \sigma_e^2 / \sigma_\beta^2$$

Pelo BLASSO e com homogeneidade de variâncias genéticas entre locos (LEGARRA et al., 2011):

$$\lambda_{BL} = [\,2\sigma_e^2 / \sigma_\beta^2\,]^{1/2}$$

Como função do penalizador no RR-BLUP:

$$\lambda_{BL} = [2\sigma_e^2 / \sigma_\beta^2]^{1/2} = [2\lambda_{RR}]^{1/2} = 1.414\,[\lambda_{RR}]^{1/2}$$

No BLASSO tem-se (CAMPOS et al., 2009b):

$$\sigma_{\beta i}^2 = \tau_i^2 \sigma_e^2$$

Com homogeneidade de variancias genéticas entre locos:

$$\sigma_\beta^2 = \tau^2 \sigma_e^2 \text{ e}$$

$\tau^2 = \sigma_\beta^2 / \sigma_e^2 = 1/\lambda_{RR}$, em que τ^2 é a média dos valores de τ_i^2.

Assim, $\lambda_{BL} = [2\sigma_e^2 / (\tau^2 \sigma_e^2)]^{1/2} = [2/\tau^2]^{1/2}$ e

$\tau^2 = 2/\lambda_{BL}^2$. Substituindo em $\sigma_\beta^2 = \tau^2 \sigma_e^2$, tem-se $\sigma_\beta^2 = 2\sigma_e^2 / \lambda_{BL}^2$.

Para cômputo da herdabilidade, tem-se (RESENDE et al., 2010):

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2}$$

Fazendo-se as substituições tem-se:

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2} = \frac{2\sum_i^n p_i(1-p_i)\tau^2\sigma_e^2}{2\sum_i^n p_i(1-p_i)\tau^2\sigma_e^2 + \sigma_e^2} = \frac{1}{1+1/\{[2\sum_i^n p_i(1-p_i)]\tau^2\}}$$

De forma alternativa e usando $\sigma_\beta^2 = 2\sigma_e^2 / \lambda_{BL}^2$, tem-se:

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2} = \frac{2\sum_i^n p_i(1-p_i)2\sigma_e^2/\lambda_{BL}^2}{2\sum_i^n p_i(1-p_i)2\sigma_e^2/\lambda_{BL}^2 + \sigma_e^2} = \frac{1}{1+\lambda_{BL}^2/[4\sum_i^n p_i(1-p_i)]} = \frac{1}{1+\lambda_{BL}^2/(2n_Q)}$$

, pois $n_Q = 2\sum_i^n p_i(1-p_i)$. Assim, com arquitetura genética homogênea, a h^2 pode ser obtida a partir do parâmetro de penalização do BLASSO e das frequências alélicas nos locos marcadores.

Sendo $\lambda_{BL} = [2\lambda_{RR}]^{1/2}$, tem-se:

$$h^2 = \frac{1}{1+\lambda_{BL}^2/(2n_Q)} = \frac{1}{1+2\lambda_{RR}/(2n_Q)} = \frac{1}{1+\lambda_{RR}/n_Q} = \frac{n_Q}{n_Q+\lambda_{RR}}$$

Pelo método RR-BLUP, a h^2 é dada por $h^2 = n_Q/(n_Q + \lambda_{RR})$, fato que confirma a equivalência dos métodos na situação de arquitetura genética homogênea.

Como λ_{RR} é assumido como conhecido no RR-BLUP, o estimador para a h^2 capturada por todos os marcadores em conjunto tem que ser especificado em função do parâmetro de penalização λ_{BL} do BLASSO, sendo dado por $\hat{h}^2 = \frac{1}{1+\hat{\lambda}_{BL}^2/(2n_Q)} = \frac{2n_Q}{2n_Q+\hat{\lambda}_{BL}^2}$. Utilizando no RR-BLUP

essa h^2 estimada, o coeficiente de regressão envolvendo valores observados e preditos pela GWS serão próximos de 1, desde que o caráter seja de arquitetura genética homogênea. Isso indica que as avaliações são não viesadas e são efetivas em predizer as reais magnitudes das diferenças entre os indivíduos em avaliação. Se a estimativa de tal coeficiente de regressão (em análise usando a h^2 estimada dessa maneira) se afastar muito de 1, há indícios de presença de genes de efeitos maiores e, nesse caso, o método RR-BLUP não é adequado, devendo-se preferir o BLASSO, o IBLASSO ou o RR-BLUP-Het.

Relação entre RR-BLUP, BLASSO e IBLASSO

Considerando todos os locos que controlam o caráter:

$$\lambda_{BLUP} = \sigma_e^2 / \sigma_a^2$$

Considerando cada loco i:

$$\lambda_{RR} = \sigma_e^2 / \sigma_{ai}^2 = \sigma_e^2 / \sigma_\beta^2$$

Pelo IBLASSO e com homogeneidade de variâncias genéticas entre locos (LEGARRA et al., 2011):

$$\lambda_{IBL} = [\,2 / \sigma_\beta^2\,]^{1/2}$$

Como função do penalizador no RR-BLUP:

$$\lambda_{IBL} = [\,2\lambda_{RR} / \sigma_e^2\,]^{1/2} = 1.414\,[\lambda_{RR} / \sigma_e^2\,]^{1/2}$$

Como função do penalizador no BLASSO, dado por $\lambda_{BL} = [\,2\sigma_e^2 / \sigma_\beta^2\,]^{1/2}$, tem-se:

$$\lambda_{IBL} = [\lambda_{BL} / \sigma_e^2]^{1/2}$$

No IBLASSO tem-se (LEGARRA et al., 2011): $\sigma_{\beta i}^2 = \tau_i^2$.

Com homogeneidade de variâncias genéticas entre locos:

$\sigma_\beta^2 = \tau^2$ e $\tau^2 = \sigma_\beta^2 = 2/\lambda_{IBL}^2$, em que τ^2 é a média dos valores de τ_i^2 .

Assim, $\lambda_{IBL} = [2/\tau^2]^{1/2}$ e como $\lambda_{BL} = [2/\tau^2]^{1/2}$ tem-se também a equivalência entre BLASSO e IBLASSO quando existe homogeneidade de variância entre locos.
Do mesmo modo, $\tau^2 = 2/\lambda_{BL}^2$ e, substituindo em $\sigma_\beta^2 = \tau^2$, tem-se $\sigma_\beta^2 = 2/\lambda_{BL}^2 = 2/\lambda_{IBL}^2$.

Para cômputo da herdabilidade, (RESENDE et al., 2010):

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2}$$

Para o IBLASSO, fazendo-se as substituições, tem-se:

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2} = \frac{2\sum_i^n p_i(1-p_i)\tau^2}{2\sum_i^n p_i(1-p_i)\tau^2 + \sigma_e^2}.$$

De forma alternativa e usando $\sigma_\beta^2 = 2/\lambda_{IBL}^2$, tem-se:

$$h^2 = \frac{2\sum_i^n p_i(1-p_i)\sigma_\beta^2}{2\sum_i^n p_i(1-p_i)\sigma_\beta^2 + \sigma_e^2} = \frac{2\sum_i^n p_i(1-p_i)2/\lambda_{IBL}^2}{2\sum_i^n p_i(1-p_i)2/\lambda_{IBL}^2 + \sigma_e^2} = \frac{1}{1+\sigma_e^2\lambda_{IBL}^2/[4\sum_i^n p_i(1-p_i)]} = \frac{1}{1+\sigma_e^2\lambda_{IBL}^2/(2n_Q)}$$

pois $n_Q = 2\sum_i^n p_i(1-p_i)$. Assim, com arquitetura genética homogênea, a h^2 pode ser obtida a partir do parâmetro de penalização do IBLASSO, das frequências alélicas nos locos marcadores e da variância residual.

Sendo $\lambda_{IBL} = [\,2\lambda_{RR}/\sigma_e^2\,]^{1/2}$, tem-se:

$$h^2 = \frac{1}{1+\sigma_e^2\lambda_{IBL}^2/(2n_Q)} = \frac{1}{1+2\lambda_{RR}/(2n_Q)} = \frac{1}{1+\lambda_{RR}/n_Q} = \frac{n_Q}{n_Q+\lambda_{RR}}$$

Pelo método RR-BLUP, a h^2 é dada por $h^2 = n_Q/(n_Q + \lambda_{RR})$, fato que confirma a equivalência dos três métodos na situação de arquitetura genética homogênea.
Como λ_{RR} é assumido como conhecido no RR-BLUP e a h^2 via λ_{IBL} depende também de σ_e^2, o estimador para a h^2 capturada por todos os marcadores em conjunto tem que ser especificado em função do parâmetro de penalização λ_{BL} do BLASSO (o qual é estimado dos dados), sendo dado por $\hat{h}^2 = \frac{1}{1+\hat{\lambda}_{BL}^2/(2n_Q)} = \frac{2n_Q}{2n_Q+\hat{\lambda}_{BL}^2}$. Utilizando no RR-BLUP, essa h^2 estimada, o coeficiente de regressão envolvendo valores observados e preditos pela GWS serão próximos de 1, desde que o caráter seja de arquitetura genética homogênea.

Análise simultânea de indivíduos genotipados e não genotipados via GBLUP

A avaliação genética em um programa de melhoramento genético envolve simultaneamente indivíduos fenotipados e genotipados, apenas fenotipados e apenas genotipados. Essas três classes de indivíduos necessitam ter seus valores genéticos preditos para que sejam ordenados e comparados. Uma opção é realizar três predições isoladas e fazer o ordenamento global. Outra opção para o grupo de indivíduos apenas genotipados é estabelecer um índice combinando a predição genômica com a predição baseada nos valores genéticos preditos de seus genitores.

No entanto, a alternativa mais eficiente é realizar toda a predição em um único passo, conforme relatado por Misztal et al. (2009) e Aguilar et al. (2010) e apresentado a seguir.

Para o grupo de indivíduos genotipados e fenotipados, o seguinte modelo linear misto geral é ajustado para estimar os efeitos genéticos aditivos usando informações fenotípicas e dos marcadores (RESENDE, 2008; RESENDE et al., 2010): $y = Wb + Za + e$, em que y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, a é o vetor dos efeitos genéticos aditivos (aleatórios) e e refere-se ao vetor de resíduos aleatórios. W e Z são as matrizes de incidência para b e a.

Esse modelo é equivalente a: $y = Wb + ZXm + e$, em que m é o vetor dos efeitos aleatórios de marcadores, X é a matriz de incidência para m e $a = Xm$.

A matriz de incidência X contém os valores 0, 1 e 2 para o número de alelos do marcador (ou do suposto QTL) em um

indivíduo diploide. Outra forma equivalente de codificar é usar os valores -1, 0 e 1.

As equações de modelo misto para a predição de *a* via o método G-BLUP equivalem a:

$$\begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z + G^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} W'y \\ Z'y \end{bmatrix}$$, em que

$G = (XX')/k = (XX')/[2\sum_i^n p_i(1-p_i)]$ e

$k = 2\sum_i^n p_i(1-p_i)$. Com padronização prévia dos elementos de X (dividindo-os por $2\sum_i^n p_i(1-p_i)]^{1/2}$) e centrando a média em zero tem-se $G = XX'$.

O parâmetro de escala $k = 2\sum_i^n p_i(1-p_i)$ assume independência entre efeitos de SNPS. Visando contornar essa suposição, Gianola et al. (2009) determinaram o seguinte parâmetro de escala:

$$k = \left((p_0 - q_0)^2 + 2\left([\sum_i^n p_i(1-p_i)]/n \right) \left((\alpha + \beta + 2)/(\alpha + \beta) \right) \right) n$$

em que $p_0 = \alpha/(\alpha + \beta)$ é a frequência alélica esperada, $q_0 = (1 - p_0)$ e α e β são parâmetros da distribuição beta ajustando a frequência alélica básica e n é o número de SNP.

O estimador de a pode ser resumido em:

$$[\hat{a}] = \left[Z'Z + G^{-1} \frac{\sigma_e^2}{\sigma_a^2} \right]^{-1} [Z\acute{y}].$$

Para a avaliação global das três classes de indivíduos em um único passo, o mesmo modelo *y = Wb + Za + e* pode ser usado, porém com uma alteração (substituição da matriz G pela matriz H) nas equações de modelo misto, conforme Misztal et al.(2009):

$$\begin{bmatrix} W'W & W'Z \\ Z'W & Z'Z + H^{-1} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} W'y \\ Z'y \end{bmatrix}$$

A matriz H inclui ambas as relações, baseadas em pedigree (A) e diferenças (A_δ) entre essas e as relações genômicas, de forma que H = A + A_δ . Assim, H é dada por

$H = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & G \end{vmatrix} = A + \begin{vmatrix} 0 & 0 \\ 0 & G - A_{22} \end{vmatrix}$, em que os subscritos 1 e 2 representam indivíduos não genotipados e genotipados, respectivamente.

A inversa de H, que permite computações mais simples, é dada por:

$H^{-1} = A^{-1} + \begin{vmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{vmatrix} = \begin{vmatrix} A^{11} & A^{12} \\ A^{21} & G^{-1} + A^{22} - A_{22}^{-1} \end{vmatrix}$, em que A_{22}^{-1} é a inversa da matriz de parentesco baseada em pedigree para os indivíduos somente genotipados.

O valor genético genômico global do indivíduo *j* é dado por $\hat{a}_j = \sum_i X_{ij}\hat{\beta}_i$. Esse, quando estimado quando o individuo j não participa da estimação de β, pode ser correlacionado com o fenótipo observado de j, visando fazer a validação.

A partir da estimação dos valores genéticos ($\hat{a}$) pelo GBLUP, os efeitos estimados dos marcadores ($\hat{\beta}$) podem ser obtidos, conforme desenvolvido a seguir:

$$\hat{a} = X\hat{\beta}$$
$$X'\hat{a} = X'X\hat{\beta}$$
$$\hat{\beta} = (X'X)^{-1}X'\hat{a}$$

Modelos com efeitos de dominância (d) podem ser ajustados. Esses são da forma $y = Wb + X\beta + Td + e$. Nesse caso, os elementos de X são codificados como $(2)^{1/2}$, 0 e $-(2)^{1/2}$ para os genótipos MM, Mm e mm, respectivamente. E os elementos de T são codificados como −1, 1 e −1 para os genótipos AA, Aa e aa, respectivamente. Valores de X e T codificados dessa forma são independentes e apresentam média zero e variância 1. Se os elementos de X são codificados com os valores -1, 0 e 1, os modelos com efeitos de dominância apresentam os elementos de T dados por 0, 1 e 0, para os genótipos MM, Mm e mm, respectivamente.

A análise pelo GBLUP é favorável computacionalmente, pois resulta em um menor número de equações a serem resolvidas. Outro uso importante dessa análise refere-se à estimação da herdabilidade total explicada por todos os marcadores simultaneamente. Com matriz de parentesco

dada por $G = (XX')/k = (XX')/[2\sum_{i}^{n} p_i(1-p_i)]$, essa h^2 pode ser estimada por REML fazendo uso das equações de modelo misto para a estimação dos componentes de variância σ_a^2 e σ_e^2. Os elementos da matriz G representam o parentesco realizado médio multi-locos e são dados por $G_{jk} = (1/n)\sum_{i=1}^{n} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$. Outro ponto favorável do GBLUP refere-se à possibilidade de estimação direta (via PEV) da acurácia da GWS. Para indivíduos com fenótipos, essa acurácia será aquela sem validação cruzada, válida para a população de estimação. No G-BLUP, a população de validação tem seus fenótipos substituídos por dados perdidos e, portanto, os indivíduos dessa população tem uma estimativa validada da acurácia.

Na população de estimação recomenda-se ajustar o vetor de fenótipos para os efeitos dos genitores antes de se fazer a análise genômica (GARRICK et al., 2009; RESENDE et al., 2010). Outra forma de realizar esse ajuste é por meio do ajuste dos efeitos de genitores como efeitos fixos (VAZQUEZ et al., 2010). Este ajuste suga dos valores genéticos individuais os efeitos dos genitores, deixando somente os efeitos da segregação mendeliana, os quais devem ser desregressados.

Modelos em nível de indivíduos contemplando as interações genótipos ambientes (ae) podem também ser ajustados, desde que existam indivíduos aparentados no mesmo ambiente e também entre ambientes. Neste caso, o modelo equivale a $y = Wb + Za + Zae + e$, em que *ae* é o vetor dos efeitos da interação entre os efeitos genéticos aditivos

e de ambientes (aleatórios) e *Z* é a matriz de incidência para *a* e *ae*. As equações de modelo misto para a predição de *a* e *ae* via o método BLUP equivalem a:

$$\begin{bmatrix} W'W & W'Z & W'Z \\ Z'W & Z'Z + G^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'W & Z'Z & Z'Z + G_{ae}^{-1}\frac{\sigma_e^2}{\sigma_{ae}^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{ae} \end{bmatrix} = \begin{bmatrix} W'y \\ Z'y \\ Z\acute{y} \end{bmatrix},$$

em que:

G_{ae} = G para pares de indivíduos no mesmo ambiente e G_{ae} = 0 para pares de indivíduos em diferentes ambientes. A variância da interação entre os efeitos genéticos aditivos e de ambientes é denotada por σ_{ae}^2 .

Análise de associação genômica ampla (GWAS)

A análise de associação genômica ampla pode ser realizada pelos seguintes métodos.

(A) Análise de associação genômica ampla (GWAS – Modelo fixo sobre fenótipos observados y; Fator de penalização $\lambda = 0$)

O modelo para o valor fenotípico em analise é dado por

$y = 1u + Xm_i + e,$

A estrutura de médias e variâncias é definida como:

$$E(y) = 1u + Xm_i$$
$$e \sim N(0,\ \ R = I\sigma_e^2) \qquad\qquad Var(y) = V = R$$

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

$$\begin{bmatrix} 1'1 & 1'X \\ X'1 & X'X \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}_i \end{bmatrix} = \begin{bmatrix} 1'y \\ X'y \end{bmatrix}$$ em que y é o vetor de fenótipos.

Resolvendo-se esse sistema, obtém-se o vetor solução $\begin{bmatrix} \hat{u} \\ \hat{m}_i \end{bmatrix}$.

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 associados à distribuição F tabelada. A hipótese alternativa é de que o marcador afeta o caráter, ou seja, o marcador e QTL encontram-se em desequilíbrio de ligação.

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

$$F_i = \frac{QM\ \mathrm{Re}\,gressão}{\hat{\sigma}_e^2} = \frac{\hat{m}_i\ X'y + \hat{u}\ 1'y - (1/n)\ (1'y)^2}{(y'y - \hat{m}_i\ X'y - \hat{u}1'y)/(n-2)}.$$

As características da GWAS tradicional são:

a. Regressão fixa em marcas únicas;

b. Fenótipos observados;

c. *Shrinkage*: 0;

d. Imprecisão devido à correlação entre efeitos dos marcadores (não considerada na análise);

e. Imprecisão devido a *shrinkage* nulo e diferentes frequências alélicas dos marcadores (não consideradas simultaneamente na análise);

f. Superestimação: cada marcador suga seu efeito e mais de outros.

(B) Análise de associação genômica ampla pós estimação simultânea (GWAS-PSE- Modelo aleatório sobre fenótipos estimados $\hat{y}$; Fator de penalização $\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$)

O modelo da GWS para o valor fenotípico em analise é dado por:

$y = 1\mu + Xm^r + e$, em que m^r é o vetor de efeitos aleatórios simultâneos de todas as marcas.

As equações de modelo misto genômicas para a predição de m^r via o método RR-BLUP-Het equivalem a:

$$\begin{bmatrix} 1'1 & 1'X \\ X'1 & X'X + I\dfrac{\sigma_e^2}{\sigma_{gi}^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}^r \end{bmatrix} = \begin{bmatrix} 1'y \\ X'y \end{bmatrix}$$

As características da GWS são:

a. Regressão aleatória em marcas simultâneas;

b. Fenótipos observados;

c. *Shrinkage* diferenciado: f($\hat{\sigma}_{gi}^2$);

d. Precisão: devida a *shrinkage* diferenciado (heterogeneidade de $\hat{\sigma}^2_{gi}$) e consideração das diferentes frequencias alélicas simultaneamente;

e. Ausência de superestimação: análise simultânea de efeitos correlacionados.

Para a GWAS-PSE deve-se inicialmente obter $\hat{y} = 1\hat{u} + X\hat{m}^r$ e realizar nova análise sob o modelo:

$$\hat{y} = 1u + Xm_i^* + e$$

$$E(\hat{y}) = 1u$$
$$e \sim N(0,\ \ R = I\sigma_e^2)$$
$$m^* \sim N(0,\ \ I\hat{\sigma}^2_{gi})$$

As equações de modelo misto para marcas individuais são:

$$\begin{bmatrix} 1'1 & 1'X \\ X'1 & X'X + \dfrac{\sigma_e^2}{\hat{\sigma}^2_{gi}} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}_i^* \end{bmatrix} = \begin{bmatrix} X'\hat{y} \\ Z'\hat{y} \end{bmatrix}$$

$$F_i^* = \frac{QM\ \mathrm{Re}\,gressão}{\hat{\sigma}_e^2} = \frac{\hat{m}_i^*\ X'\hat{y} + \hat{u}\ 1'\hat{y} - (1/n)\ (1'\hat{y})^2}{(\hat{y}'\hat{y} - \hat{m}_i^*\ X'\hat{y} - \hat{u}1'\hat{y})/(n-2)}$$

As características do método GWAS-PSE são:

a. Regressão aleatória em marcas únicas;

b. Fenótipos estimados;

c. *Shrinkage* diferenciado: f($\hat{\sigma}^2_{gi}$);

d. Precisão: devida a *shrinkage* diferenciado (heterogeneidade de $\hat{\sigma}^2_{gi}$);

e. Ausência de superestimação: análise simultânea de efeitos correlacionados;

f. Ausência de superestimação: cada marcador suga apenas seu efeito, ditado por seu $\hat{\sigma}^2_{gi}$ estimado via análise simultânea.

Para obtenção de $\hat{\sigma}^2_{gi}$ pode-se usar os métodos IBLASSO, BLASSO, BayesA, BayesB, BayesCπ.

(C) Análise de associação genômica ampla pós estimação simultânea (GWAS-PSE- Modelo fixo sobre fenótipos estimados $\hat{y}$; Fator de penalização $\lambda = 0$)

Equivale ao modelo descrito em (A), porém aplicado sobre fenótipos estimados.

(D) Análise de associação genômica ampla (GWAS-PSE- Modelo aleatório sobre fenótipos observados y; Fator de penalização $\lambda = \sigma^2_e / \hat{\sigma}^2_{gi}$)

Equivale ao modelo descrito em (B), porém aplicado sobre fenótipos observados.

Essas quatro abordagens foram aplicadas a dados reais (nível de significância 5% pelo teste F), gerando os resultados mostrados na Tabela 5.

Tabela 5. Comparação entre os modelos de análise de associação (GWAS).

Método	Modelo para efeitos de marcas	Fenótipos	Penalização	N marcas significativas
A	Fixo	y	$\lambda = 0$	687
C	Fixo	$\hat{y}$	$\lambda = 0$	652
B	Aleatório	$\hat{y}$	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	394
D	Aleatório	y	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	63

Verifica-se que tratar os efeitos de marcas como fixos conduz à seleção de um maior número de marcas. Isso ocorre porque essa abordagem superestima os efeitos de marcas individuais. Verifica-se também que o método D conduz à seleção do menor número de marcas e o método B é o mais equilibrado.

As vantagens dos métodos GWAS-PSE são: estimação simultânea dos efeitos de marcas, consideração dos efeitos de marcas como aleatórios, consideração da heterogeneidade de variâncias entre marcas e estimação após validação cruzada.

Associação genômica ampla (GWAS) em humanos

Os primeiros estudos em genética quantitativa humana visando ao entendimento do controle genético dos caracteres basearam-se na estimação da herdabilidade (h^2) via análise de pares de gêmeos, usando o conceito de semelhança entre parentes baseada em pedigree (alelos idênticos por descendência, IBD). Essa abordagem

considera todos os locos, variantes comuns e raros (genes de baixa frequência), ou seja, todos genes que controlam o caráter ou h^2 total.

O papel de genes individuais no controle genético dos caracteres passou a ser estudado pela metodologia de Fulker e Cardon (1994), por meio da estimação da h^2 de um loco marcado no contexto do mapeamento de QTL, conforme descrito por Resende (2008) e Cruz et al. (2009). A aplicação do método fundamenta-se na análise de ligação dentro de família de irmãos completos, usando marcas moleculares duas a duas.

Visscher et al. (2006; 2008) apresentaram uma abordagem para a estimação da h^2 usando simultaneamente todos os locos marcados e também usando análise de segregação dentro de família de irmãos completos. Essa abordagem genômica ampla baseia-se também em IBD e capitaliza o parentesco exato ou realizado. A h^2 estimada foi de 0,80 para altura em humanos. O método considera variantes comuns e raros (genes de baixa frequência), ou seja, todos os genes ou h^2 total, pois usa também o pedigree via genotipagem dos genitores, estimando alelos IBD em todos os locos.

Outro método de estudo do controle dos caracteres em nível populacional e não apenas dentro de famílias é a GWAS. Essa baseia-se em análise de desequilíbrio de ligação em nível populacional, porém usando apenas um loco marcador de cada vez, via análise de regressão fixa sobre indivíduos não aparentados. A h^2 capturada pelos marcadores significativos foi de apenas 0,10 para altura em humanos.

A GWAS entre membros de uma família (de irmãos completos) pode ser descrita como uma análise de ligação.

Em tal análise, marcadores a alguma distância de um QTL exibirá uma associação com o caráter porque houve apenas uma geração de recombinação entre os genitores e os filhos irmãos completos. Consequentemente, um alelo marcador e um alelo do QTL no mesmo cromossomo tenderão a ser herdados juntos.

Um procedimento mais eficaz para capturar a maioria da herdabilidade de um caráter é a análise de desequilíbrio de ligação em nível populacional usando todos os locos marcadores simultaneamente de maneira similar ao método da GWS. É baseado em regressão aleatória para a predição de efeitos latentes. Utiliza indivíduos não aparentados, embora todos os indivíduos de uma espécie sejam aparentados em algum grau porque compartilham ancestrais comuns e, portanto, compartilham alelos idênticos em estado (IBS).

Os marcadores SNPS captam esses parentescos ancestrais e, portanto, estimam relações genéticas entre indivíduos baseadas em IBS (POWELL et al., 2010; VISSCHER et al., 2010). O uso simultâneo da genética de populações (análise de ligação, desequilíbrio de ligação e mapeamento genético) e da genética quantitativa (estimação da herdabilidade), tradicionalmente foram usados separadamente na genética humana. A GWS combinando essa duas áreas permitiu capturar uma h^2 de 0,45 para altura em humanos. O restante (0,80 – 0,45 = 0,35) não capturado é devido a muitos variantes de baixa frequência (incluindo locos de grande efeito).

A variação genética no loco i é dada por $\sigma_{ai}^2 = 2p_i(1-p_i)a_i^2$, ignorando a dominância. Assim, um alelo raro não pode explicar grande parte da variação genética, mesmo se for de grande efeito. Para que esses locos sejam capturados pelos marcadores e detectados é necessário um grande

tamanho amostral. Pelo método GWS a variação genética aditiva total é estimada por $\sigma_a^2 = \sum_i 2p_i(1-p_i)a_i^2$.

Outra forma muito usada para a estimação da h^2 é via análise de ligação usando toda a genealogia (ALMASY; BLANGERO, 1998; HAMZA; PAYAMI, 2010). O software Solar (*Sequential Oligogenic Linkage Analysis Routines*) tem sido usado para estimação.

Aulchenko et al. (2007) propuseram o método GRAMMAR para a GWAS em múltiplos estágios, conforme descrito a seguir. Após o ajuste do modelo $y = Xb + Zg + e$ obtém-se $\hat{e} = y - X\hat{b} - Z\hat{g}$, em que g é um vetor de efeitos poligênicos. Ajusta-se então o modelo $\hat{e} = 1u + Wm_i + e$, identificando-se os marcadores significativos. Apenas com os SNPs significativos, ajusta-se o modelo $y = Xb + Wm_i + Zg + e$. Isso reduz o tempo de computação. Os efeitos m são ajustados como efeitos fixos (pois assim os SNPs não modelam estrutura familiar em g, isto é, não explicam correlação entre indivíduos aparentados, com alelos IBD). Fundamenta-se no fato de que os efeitos de genes maiores integram o vetor de resíduos condicionais, após o ajuste para g sob modelo poligênico infinitesimal (ajuste ou eliminação dos efeitos de família ou variação entre pedigrees ou estrutura ou do parentesco). Na análise final, volta-se com o modelo completo. Nesse caso, o efeito poligênico é incluído visando corrigir os dados para a estrutura de famílias por meio da matriz de parentesco, visto que $g \sim N(0, A\sigma_g^2)$.

A comparação de modelos hierárquicos, mas com mesma estrutura de efeitos fixos, é realizada pelo LRT ou análise

de *deviance*. A comparação de modelos não hierárquicos, mas com mesma estrutura de efeitos fixos, deve ser feita por meio dos procedimentos AIC e BIC. O AIC está relacionado aos conceitos de informação de kullback-Leibler e máxima verossimilhança (ANDERSON et al., 2000). Informação de kullback-Leibler é um conceito da física para medir a diferença entre o modelo (aproximação da realidade) e a realidade. Akaike (1974) percebeu que o log da verossimilhança de um modelo é um estimador da informação de kullback-Leibler, porém viesado. E esse viés é igual ao número de parâmetros do modelo. Então,definiu o AIC como a *deviance* mais duas vezes o número de parâmetros do modelo. Como o objetivo é minimizar a perda de informação, o modelo com o menor AIC tem o maior suporte nos dados.

Captura da h^2 em humanos, imperfeito LD entre SNPs e variantes causais

Visscher et al. (2010) abordam os resultados da GWAS referente ao caráter altura em humanos. A h^2 capturada pela GWAS nos estudos tradicionais foi da ordem de 0,10. Esse baixo valor ocorreu devido ao fato de variantes de baixa frequência (MAF < 0.10) não estarem em perfeito LD com marcadores comuns (MAF > 0.10), ou seja, o r^2 é baixo e também variantes de pequenos efeitos não são detectados significativamente pela GWAS tradicional, mesmo se em LD com marcadores comuns. No estudo de Yang et al. (2010), a h^2 capturada foi de 0,45. Isso ocorreu porque variantes de pequenos efeitos não são detectados significativamente, mas em LD com marcadores comuns, são capturados pela GWS a qual não faz uso de significância para efeitos de marcas.

O valor máximo que r^2 pode atingir é fortemente determinado pelas frequências alélicas nos dois locos

(WRAY, 2005). Quanto mais diferentes as frequências alélicas, menor o valor de r^2. Assim, como a maioria dos SNP genotipados são comuns, se os variantes são raros r^2 será baixo e, então a variação σ^2_{mi} associada aos SNP é substancialmente menor que a variação σ^2_{ai} no QTL (VISSCHER et al., 2010). As expressões abaixo ilustram essa questão.

$$r^2 = \sigma^2_{mi} / \sigma^2_{ai}$$

$$\sigma^2_{mi} = r^2 \sigma^2_{ai}$$

Na prática, pode-se estimar o LD apenas entre os SNP. Essa estimativa pode ser útil apenas quando SNP e gene apresentam frequências alélicas similares. Um gene pode estar em LD com múltiplos SNPs, então esses coletivamente podem capturar o variante causal mesmo que nenhum SNP esteja em perfeito LD com ele (VISSCHER et al., 2010). Assim, um SNP pode não ser detectado como significativo, mas, em conjunto com outros, ser importante para explicar a variação genética e maximizar a acurácia seletiva. Dessa forma, recomenda-se não aplicar teste de significância antes da GWS.

Mesmo com o uso de dezenas de milhares de marcadores, se os variantes são raros, e sendo comuns os marcadores, ainda assim, os marcadores não capturarão toda a variação genética. Assim, a eficiência da GWS depende da arquitetura genética do caráter na população. Se o mesmo for governado por um grande número de variantes raros que explicam grande parte da variação genética, a GWS terá menor sucesso. Nesse caso, é recomendável ajustar no modelo, o efeito poligênico residual, como forma de capturar esses variantes raros.

Em resumo, as causas da herdabilidade perdida são: (i) variantes de baixa frequência (MAF < 0,10) não estão em perfeito LD com marcadores comuns (MAF > 0,10), causando baixo r^2; (ii) pequeno número de marcas, causando baixo r^2; (iii) uso apenas dos SNPs significativos na GWAS.

A estimação simultânea é necessária porque os SNPs estão em LD, ou seja, são dependentes e correlacionados. A regressão simultânea é equivalente a regressar o fenótipo em todos os componentes principais derivados dos marcadores, sendo que o grau de *shrinkage* experimentado por cada efeito estimado é proporcional ao seu associado valor singular quadrático (CAMPOS et al., 2010). Isso dá suporte ao método GWAS-PSE e, mais ainda, à própria GWAS com estimação simultânea (GWAS-SE), conforme Yang et al. (2011). Baseados nesse princípio há também os métodos regressão via quadrados mínimos parciais (PLSR) e regressão via componentes principais (PCR) (SOLBERG et al., 2009).

Ilustra-se a seguir a dependência de r^2 em relação às frequências alélicas nos dois locos considerados. O r^2 é um coeficiente de determinação e equivale ao quadrado do coeficiente de correlação entre duas variáveis ou locos a e b, dado por:

$$r = \frac{Cov(a,b)}{[Var(a)Var(b)]^{1/2}} = \frac{\sum ab - \sum a \sum b}{[Var(a)]^{1/2}[Var(b)]^{1/2}} = \frac{\Pr ob(ab) - \Pr ob(a) \Pr ob(b)}{[pq]^{1/2}[rs]^{1/2}} = \frac{D}{[pq\ rs]^{1/2}}$$

O quadrado dessa quantidade equivale a $r^2 = \frac{D^2}{[pq\ rs]}$, que é a medida padrão de desequilíbrio de ligação. Usando as matrizes de incidência X dos marcadores o valor de r pode ser dado por $r_{(a,b)} = \frac{Cov(X_{ia}, X_{ib})}{[Var(X_{ia})]^{1/2}[Var(X_{ib})]^{1/2}}$.

Definem-se as quantidades $D = \text{Prob}(ab) - \text{Prob}(a)\text{Prob}(b)$, em que Prob (a) é a frequência do alelo a e Prob (ab) é a frequência do genótipo ab. Genericamente, p é a frequência do alelo A, q é a frequência do alelo a, r é a frequência do alelo B e s é a frequência do alelo b. A igualdade $Var(a) = pq$ assume distribuição Bernoulli para a presença do alelo.

Comparação entre 12 métodos de seleção genômica ampla

Para a comparação entre vários métodos estatísticos na GWS foram simulados dois conjuntos de dados usando o aplicativo RealBreeding (VIANA, 2011), (Tabela 6).

Tabela 6. Parâmetros usados na simulação.

Caráter	Va	Ve	h2	Soma 2pq	N genes menores	N genes maiores	N indivíduos	N SNP
Sem gen maior	4,826202	11,26114	0,300	233,47	100	0	300	500
Com gen maior	114,5132	267,1974	0,300	231,80	98	2*	300	500

* os dois explicando 30% da variação genética e os 98 explicando 70%.

Foram empregados os seguintes softwares e métodos na GWS (Tabela 7).

Tabela 7. Softwares e métodos usados na GWS.

Método	Software	Referência
1 FR-LS	Selegen Genômica	Resende (2007)
2 RR-BLUP	Selegen Genômica	Resende (2007)
3 RR-BLUP-Het	Selegen Genômica	Resende (2007)
4 RR-BLUP Padronizado	*Genome Wide Prediction*	Meuwissen et al (2009)
5 Fast BayesA	*Genome Wide Prediction*	Meuwissen et al (2009)
6 Fast BayesB	*Genome Wide Prediction*	Meuwissen et al (2009)
7 IBLASSO	GS3	Legarra et al (2011)
8 BayesCPi	GS3	Legarra et al (2011)
9 MCMC-BLUP	GS3	Legarra et al (2011)
10 BLASSO	BLR	Perez et al. (2010)
11 RKRS	R	Campos et al. (2009a)
12 PLSR	R	Os autores

Os resultados referentes à GWS são apresentados na Tabela 8.

Tabela 8. Resultados de acurácia referentes à GWS.

Método	Acurácia – Caráter 1	Acurácia – Caráter 2
1 FR-LS	0,59	0,44
2 RR-BLUP	0,71	0,78
3 RR-BLUP-Het (IBLASSO)	0,71	**0,80**
4 RR-BLUP Padronizado	0,71	0,78
5 Fast BayesA	0,71	0,79
6 Fast BayesB	0,71	0,79
7 IBLASSO	0,71	**0,80**
8 BayesCPi	0,59	0,70
9 MCMC-BLUP	0,71	**0,80**
10 BLASSO	0,68	0,63
11 RKRS	0,99	0,99
12 PLSR	0,99	0,99

Verifica-se que, para o caráter 1, com arquitetura genética homogênea, a maioria dos métodos forneceram acurácia idêntica de 0,71. Apenas os métodos FR-LS, BLASSO e BayesCPi foram inferiores. Os métodos RKRS e PLSR não usam herdabilidade e, portanto, os resultados (0,99) obtidos na população de estimação referem-se a coeficientes de determinação fenotípica e não a acurácias. Para a comparação desses métodos com os demais torna-

se necessária a realização de validação cruzada em todos os métodos.

Para o caráter 2, com arquitetura genética heterogênea, os métodos diferiram mais, destacando-se como superiores os métodos IBLASSO, RR-BLUP-Het (com componentes de variância estimados pelo IBLASSO) e MCMC-BLUP, concordando com Legarra et al. (2011). Os métodos FR-LS e BLASSO foram inadequados para os dois caracteres. Os métodos RR-BLUP e RR-BLUP padronizado, se aplicados corretamente, são idênticos.

Foram também comparados cinco métodos na GWAS, conforme a Tabela 9.

Tabela 9. Softwares e métodos usados na GWAS.

Método	**Software**	**Referência**
1 GWAS-FR-OBS	Selegen Genômica	Resende (2007)
2 GWAS-PSE-FR-EST	Selegen Genômica	Resende (2007)
3 GWAS-PSE-RR-OBS	Selegen Genômica	Resende (2007)
4 GWAS-PSE-RR-EST	Selegen Genômica	Resende (2007)
5 IBLASSO	GS3	Legarra et al (2011)

*FR: regressão fixa; RR: regressão aleatória

Os resultados referentes à GWAS para o caráter 2 são apresentados na Tabela 10. São apresentados o número de marcas retidas em cada método, a acurácia da GWS com o emprego das referidas marcas retidas e as 16 marcas de maiores efeitos em cada método de GWAS.

Tabela 10. Resultados referentes à GWAS.

Método1	Método 2	Método 3	Método 4	Método 5
Número marcas = 95	**Número marcas = 139**	**Número marcas = 6**	**Número marcas = 97**	**Número marcas = 169**
Acurácia = 0,80	Acurácia = 0,80	Acurácia = 0,56	Acurácia = 0,81	Acurácia =0,79
38	31	2	16	2
16	43	16	31	394
2	11	38	43	38
31	16	7	38	16
7	38	31	11	218
43	19	43	2	7
49	5	-	19	84
33	49	-	5	17
11	7	-	49	330
17	36	-	6	35
6	10	-	10	190
42	6	-	33	33
36	29	-	36	303
29	39	-	7	252
4	33	-	29	49

Verifica-se que os métodos retiveram número de marcas ligeiramente diferentes mas conduziram a acurácias similares, exceto pelo método 3. Essas acurácias foram também praticamente idênticas àquelas obtidas com o uso de todas as 500 marcas. Assim, é possível a seleção de um subconjunto de marcas. O método 3 é útil em fornecer um ponto de corte para a seleção de um número muito restrito de marcas mais associadas com o caráter. As marcas com maior associação com o caráter foram aproximadamente coincidentes nos diferentes métodos. O nível de significância de 5% na GWAS parece adequado para a GWS. Isso difere dos níveis bem rigorosos (menos que 1/1000) adotados na GWAS propriamente dita.

Pesos das marcas nos diferentes métodos e frequências alélicas

O conhecimento dos pesos dados às diferentes fontes de informação nos procedimentos de estimação é relevante no estudo das propriedades dos diferentes métodos de estimação. Mrode et al. (2010) abordaram essa questão.

A equação de estimação dos efeitos de marcadores pelo método RR-BLUP é dada por $\hat{\beta} = (X' X + \lambda_{RR} I)^{-1} X' y$. O estimador do efeito de uma marca i equivale a $\hat{\beta}_i = (x_i' x_i + \lambda_{RR} I)^{-1} x_i' x_i yd_i = w_i yd_i$, em que yd_i é o desvio fenotípico associado à marca i corrigido para todos os demais efeitos ambientais e genéticos de outras marcas, sendo dado por $yd_i = x_i'(y - \mu - x_j \hat{\beta}_j)$, i ≠ j. O valor genético aditivo do indivíduo k é dado por $\hat{a}_k = \sum_i x_i w_i yd_i$.

Pelos métodos bayesianos BayesA e BayesB existe um componente adicional resultante da amostragem da

distribuição condicional *a posteriori* de β tal que $\hat{a}_k = \sum_i x_i w_i y d_i + N(\hat{\beta}_i, (x_i' x_i + \lambda_i)^{-1} \sigma_e^2)$. O segundo termo dessa equação tende a zero quando se faz as médias de todas as amostras de Gibbs salvas após o período de *burn in*.

Diferenças nos pesos dos marcadores, ou seja, diferentes *shrinkages* podem surgir mesmo quando se usa o método RR-BLUP, como resultado da variação nas frequências alélicas. Mrode et al. (2010) relatam os seguintes pesos associados a cada categoria (alta, média e baixa) de frequência alélica: 0,19, 0,12 e 0,04, respectivamente. Para os métodos BayesA e BayesB, os pesos não variaram entre as categorias de frequência alélica, equivalendo a 0,52 e 0,88, respectivamente. O peso maior associado ao BayesB deve-se ao fato desse método efetivamente ajustar um menor (66% no caso) número de marcadores.

Verifica-se então que os pesos diferem entre métodos. Isso afeta as alterações nas frequências alélicas como resultado da seleção. E o método RR-BLUP enfatiza pouco os alelos de baixa frequência, podendo ser desfavorável para o melhoramento a longo prazo. Para contornar isso, um índice de seleção enfatizando mais os alelos de baixa frequência poderia ser estabelecido.

As correlações entre pesos e frequências alélicas foram 0,99; 0,40 e -0,05 para o RR-BLUP, BayesA e BayesB, respectivamente. No método RR-BLUP, a quantidade e magnitude de informação depende essencialmente das frequências alélicas. No BayesA e BayesB, dependem também da variação genética diferencial entre locos. Conforme Mrode et al. (2010), a correlação entre os efeitos dos marcadores pelos métodos BayesA e RR-BLUP usando

componentes de variância obtidos pelo método BayesA foi de 0,99.

Formas de parametrização da matriz de incidência genotípica

Parametrização 1

A matriz de incidência X contém os valores 0, 1 e 2 para o número de alelos do marcador (ou do suposto QTL) em um indivíduo diploide e, 2p para os indivíduos com dados perdidos de marcas. Esses valores devem ser centrados em zero para que os efeitos das marcas codominantes sejam efeitos de substituição alélica com média zero na população, e, nesse caso, assumindo equilíbrio de Hardy-Weinberg, a variação genética aditiva do caráter na população equivale a $\sigma_a^2 = 2\sum_i^m p_i(1-p_i)\sigma_\beta^2$. Dessa forma, os valores de X_i devem ser subtraídos pela média de X (via 0 - 2p, 1 - 2p e 2 - 2p, respectivamente) obtendo-se uma variável com média zero. Assim, com centralização, no método RR-BLUP deve-se usar $n_Q = 2\sum_i^m p_i(1-p_i)$ e os efeitos genéticos aditivos dos indivíduos são dados por $\hat{a} = X\hat{\beta}$.

Adicionalmente, pode-se padronizar os dados dos marcadores na matriz X, da seguinte forma para cada elemento X_i da matriz, referente ao loco i:

$X_i = (0 - 2p_i) / (Var(X_i))^{1/2}$ se o indivíduo é homozigoto para o primeiro alelo (mm);

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

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

$X_i = 0$ se o indivíduo apresenta dado perdido de marca.

A quantidade p_i é a frequência do segundo alelo do marcador. Dessa forma, a variância de X com X_i ajustado é igual a 1, obtendo-se uma variável com média zero e variância unitária.

Sendo β o efeito do marcador na população, a variância devida ao marcador é dada por $Var(X_i\beta) = Var(X_i)\ Var(\beta)$. Com a transformação acima, $Var(X_i) = 1$ e portanto, $Var(X_i\beta) = Var(\beta)$. Em outras palavras, modelando a variância do efeito do marcador, modela-se diretamente a variância do marcador, independentemente de sua frequência. Assim, com centralização e padronização $\sigma_a^2 = m\sigma_\beta^2$. Dessa forma, no método RR-BLUP deve-se usar $n_Q = m$ e os efeitos genéticos aditivos dos indivíduos são dados por $\hat{a} = X\hat{\beta}$.

Parametrização 2

Em outra parametrização, a matriz de incidência X contém os valores -1, 0 e 1 para o número de alelos do marcador (ou do suposto QTL) em um indivíduo diploide, ou seja, para os genótipos mm, Mm e MM, respectivamente. Essa parametrização é ligeiramente inferior à anterior (LEGARRA et al., 2011). Para essa parametrização deve-se usar, nométodo RR-BLUP, $n_q = 2\sum_i^m p_i(1-p_i)$ e o efeito genético aditivo do indivíduo j é dado por

$$\hat{a}_j = \sum_i^m [I(x_{ij} = 1)(2p_i\hat{\beta}_i) + I(x_{ij} = 0)(p_i\hat{\beta}_i - q_i\hat{\beta}_i) + I(x_{ij} = -1)(-2q_i\hat{\beta}_i)$$

Imputação de genótipos marcadores

Dados perdidos associados aos genótipos marcadores podem ser imputados cientificamente usando a informação de parentesco entre os indivíduos genotipados e não genotipados. Assim, para funcionar, esse método demanda que haja algum parentesco entre os indivíduos da população.

O conteúdo alélico c para os indivíduos genotipados (Y) é dado por 0, 1 ou 2 para os genótipos aa, Aa e AA, respectivamente, para marcadores bialélicos e codominantes. O conteúdo alélico para os indivíduos não genotipados (X) é dado por (GENGLER et al., 2007):

$c_X = \begin{pmatrix} 1 & A_{XY}A_{YY}^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix}$, em que A refere-se à matriz de parentesco (correlação) genético aditivo entre indivíduos genotipados (A_{YY}) e entre indivíduos genotipados e não genotipados (A_{XY}); c_Y é o vetor de conteúdo alélico dos indivíduos genotipados; μ é a média geral, calculada diretamente dos dados genotípicos: 1 é um vetor de uns.

A média geral pode também ser calculada simultaneamente ao vetor c_X por meio das equações de modelo misto:

$\begin{pmatrix} 1'1 & 1'M \\ M'1 & M'M + A^{-1}\alpha \end{pmatrix}\begin{pmatrix} \hat{\mu} \\ \hat{c}_Y \\ \hat{c}_X \end{pmatrix} = \begin{pmatrix} 1'c_Y \\ M'c_y \end{pmatrix}$, em que M é uma matriz de incidência que associa c_Y a $\begin{pmatrix} c_Y \\ c_X \end{pmatrix}$. M pode ser reescrita como $M = \begin{pmatrix} I_Y & 0_X \end{pmatrix}$, em que I é uma matriz identidade. A matriz de parentesco é dada por $A = \begin{pmatrix} A_{YY} & A_{YX} \\ A_{XY} & A_{XX} \end{pmatrix}$. O fator α é necessário para que o sistema tenha solução e é dado por $\alpha = \sigma_e^2 / \sigma_c^2$, em que σ_e^2 é a variância do erro de genotipagem e σ_c^2 é variância do conteúdo alélico c. O componente σ_e^2 deve ser mantido próximo de zero, ou seja, da ordem de 0,001. Isso está associado a um coeficiente de determinação de c equivalente a 0,999. Dessa forma, $\alpha = \sigma_e^2 / \sigma_c^2 = 0{,}001 / 0{,}999 = 0{,}001001$. O modelo associado ao sistema de equações equivale a $c_Y = \mu + Mc_Y^* + e$, em que $c_Y^* = [c_Y \quad c_X]$.

Considere o seguinte exemplo, com quatro indivíduos genotipados (não aparentados e com contagem de alelos marcadores 1, 0, 2 e 2, respectivamente) e 1 não genotipado e irmão completo do indivíduo número 4. Tem-se as seguintes matrizes e resolução pelas equações de modelo misto:

1´ = [1 1 1 1]

c_Y´ = [1 0 2 2]

M = [1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0]

A = [1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0,5
0 0 0 0,5 1]

Sendo $\alpha = 0{,}001$, tem-se

Matriz dos Coeficientes = MC

$$MC = \begin{pmatrix} 1'1 & 1'M \\ M'1 & M'M + A^{-1}\alpha \end{pmatrix}$$

MC = [4,0000 1,0000 1,0000 1,0000 1,0000 0
1,0000 1,0010 0 0 0 0
1,0000 0 1,0010 0 0 0
1,0000 0 0 1,0010 0 0
1,0000 0 0 0 1,0013 -0,0007
0 0 0 0 -0,0007 0,0013]

Lado Direito das Equações = LD

$$LD = \begin{pmatrix} 1'c_Y \\ M'c_y \end{pmatrix}$$

LD´ = [5 1 0 2 2 0].

Solução

$$\begin{pmatrix} \hat{\mu} \\ \hat{c}_Y \\ \hat{c}_X \end{pmatrix} = (MC)^{-1}LD = \begin{pmatrix} 1{,}2500 \\ -0{,}2498 \\ -1{,}2488 \\ 0{,}7493 \\ 0{,}7493 \\ 0{,}3746 \end{pmatrix}.$$

Assim, o genótipo imputado para o indivíduo 5 foi 0,3746.

Resolvendo-se via fórmula tem-se:

$$c_X = \begin{pmatrix} 1 & A_{XY}A_{YY}^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix}$$

$$= \begin{pmatrix} 1 & A_{XY}I_{(4)} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix} = \begin{pmatrix} 1 & A_{XY} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 0.5 \end{pmatrix} \begin{pmatrix} 1.25 \\ -0.25 \\ -1.25 \\ 0.75 \\ 0.75 \end{pmatrix} = 1.625$$

O valor 1,625 menos a média geral 1,25, fornece o valor 0,375.

Aumento na eficiência seletiva do melhoramento de plantas e animais

O aumento da eficiência seletiva com o uso da GWS pode ocorrer pela alteração dos quatro componentes da expressão do progresso genético, dada por $G_S = (k\, r_{g\hat{g}}\sigma_g)/L$, em que k é o diferencial de seleção padronizado (dependente da intensidade de seleção), $r_{g\hat{g}}$ é a acurácia seletiva, σ_g é o desvio padrão genético (variabilidade genética) do caráter na população e L é o tempo necessário para completar um ciclo seletivo.

Espécies vegetais perenes (florestais, fruteiras, forrageiras, cana-de-açúcar) e animais

Nessas espécies, o benefício da GWS se dá devido ao aumento de $r_{g\hat{g}}$ e redução em L. O aumento em $r_{g\hat{g}}$ se dá devido ao uso da matriz de parentesco real e própria de cada caráter (RESENDE, 2007). E esse aumento depende do tamanho da população de estimação e da densidade de marcadores. O fator L é enormemente reduzido com a GWS, pois a predição genômica e a seleção podem ser feitas no estágio de plântulas. Assim, mesmo que $r_{g\hat{g}}$ seja de mesma magnitude que aquela obtida com a seleção fenotípica, a GWS será ainda superior à seleção baseada em fenótipos, devido à redução em L.

Espécies vegetais alógamas anuais (milho, girassol)

Nessas espécies o benefício da GWS se dá devido a três fatores: aumento de $r_{g\hat{g}}$, aumento de k e redução em L.

Há também um aumento da variação genética explorada pelo método da seleção recorrente.

Nesse caso, o aumento de $r_{g\hat{g}}$ se dá devido ao uso da matriz de parentesco real e também devido ao fato de se explorar toda a variação genética da população e não somente aquela entre famílias. Uma vez que a seleção pela GWS é praticada precocemente e antes do florescimento, torna-se possível a seleção em nível de indivíduo e nos dois sexos (como se faz no melhoramento de plantas perenes), sem a necessidade de duas estações de plantio: uma para a avaliação de famílias e outra para o estabelecimento do lote de recombinação. Consequentemente, o tempo L também é reduzido. Essa coincidência entre unidade de seleção e unidade de recombinação maximiza também a herdabilidade do método de seleção (explora adicionalmente 0,50 ou 0,75 da variação genética aditiva que estava dentro de progênies). A seleção em nível de indivíduo propicia também o aumento da intensidade de seleção k.

Espécies vegetais autógamas anuais (soja, feijão, arroz, trigo)

Nessas espécies, usando a duplicação de haplóides para a obtenção direta de linhagens, o benefício da GWS se dá devido aos quatro fatores: aumento de $r_{g\hat{g}}$, aumento de k, aumento de σ_g (por meio da exploração de duas vezes a variação genética aditiva) e redução em L.

Seguindo o método normal ou genealógico de melhoramento, tem-se que a seleção via GWS não pode ser realizada na geração F_2, pois deve-se caminhar até a homozigose para a seleção final. Assim, não se reduz L. Mas pode-se identificar os bons alelos com a GWS na geração F_2 e direcionar o cruzamento entre as melhores

plantas, fazendo-se a seleção recorrente intrapopulacional em autógamas. Isso permite aumentar $r_{g\hat{g}}$ e σ_g e, consequentemente, aumenta-se o ganho genético. Adicionalmente aumenta-se k, pois é possível avaliar um número muito maior de plantas F_2 do que de famílias $F_{2:3}$.

Para o avanço de plantas S_0 até linhagens homozigotas pode-se praticar a seleção precoce via GWS em cada geração (sem a necessidade de experimentar progênie), maximizando-se então a acurácia seletiva. A estimação dos efeitos de marcas é baseada em plantas S_0 da geração F_2.

Referências

AGUILAR I.; MISZTAL, I.; JOHNSON, D. L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T. J. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, Champaign, v. 93, n. 2, p. 743-52, 2010.

ALMASY, L.; BLANGERO, J. Multipoint quantitative-trait linkage analysis in general pedigrees. **The American Journal of Human Genetics**, Chicago, v. 62, n. 5, p. 1198-1211, 1998.

ANDERSON, D. R.; BURNHAM, K. P.; THOMPSON, W. L. Null hypothesis testing: problems, prevalence, and an alternative. **Journal of Wildlife Management**, Bethesda, v. 64, p. 912-923, 2000.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transaction on Automatic Control**, v. 19, p. 716-723, 1974.

AULCHENKO, Y. S.; KONNING, D.; HALEY, C. Grammar: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. **Genetics**, Austin, v. 177, p. 577-585, 2007.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American Statistician**, Washington, DC, v. 49, n. 4, p. 327-335, 1995.

CRUZ, C. D.; GOD, P. I. V. ; BHERING, L. L. Mapeamento de QTLs em populações exogâmicas. In: BORÉM, A.; CAIXETA, E. T. (Org.). **Marcadores Moleculares**. 2. ed. Viçosa, MG: Folha de Viçosa, 2009. v. 1. p. 443-481.

CAMPOS, G. de los; GIANOLA, D.; ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, Champaign, v. 87, p.1883-1887, 2009a.

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**, Austin, v. 182, p. 375-385, 2009b.

CAMPOS, G. de los; GIANOLA, D.; ALLISON, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. **Nature Reviews Genetics**, London, v. 11, p. 880-886 Dec. 2010.

FULKER, D. F.; CARDON, L. R. A sib-pair approach to interval mapping of quantitative trait loci. **American Journal of Human Genetics**, Chicago, v. 54, p. 1092-1103, 1994.

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

GENGLER, N.; MAYERES, P.; SZYDLOWSKI, M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. **Animal**, Cambrige, v. 1, n. 1, p. 21-28, 2007. DOI: 10.1017/S1751731107392628

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

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

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

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

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

GODDARD, M. E.; WRAY, N. R.; VERBYLA, K.; VISSCHER, P .M. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, Hayward, v. 24, p. 517-529, 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**, Austin, v. 178, n. 4, p. 2305–2313, 2008.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, London, v. 12, p. 186, 2011.

HAMZA, T. H.; PAYAMI, H. The heritability of risk and age at onset of Parkinson's disease after accounting for known genetic risk factors. **Journal of Human Genetics**, v. 55, p. 241–243, 2010.

HASTIE, T.; TIBSHIRANI, R. Generalized Additive Models (with discussion). **Statistical Science**, v. 1, n. 3, p. 297-318, 1986.

LEGARRA, A.; ROBERT-GRANIÉ, C.; CROISEAU, P.; GUILLAUME, F.; FRITZ, S. Improved Lasso for genomic selection. **Genetics Research**, Cambridge, v. 93, n. 1, p. 77-87, 2011.

MAKOWSKY, R.; PAJEWSKI, N. M.; KLIMENTIDIS, Y. C.; VAZQUEZ, A. I.; DUARTE, C. W.; ALLISON, D. B.; CAMPOS, G. de los. Beyond missing heritability: prediction of complex traits. **Plos Genetics**, San Francisco, CA, v. 7, n. 4, 2011.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, 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**, London, v. 41, p. 2, 2009. DOI:10.1186/1297-9686-41-2.

MISZTAL, I.; LEGARRA, A.; AGUILAR I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, Champaign, v. 92, n. 9, p. 4648-55, 2009.

MRODE, R.; COFFEY, M.; BERRY, D.P. Understanding genomic evaluations from various evaluation methods and GMACE. **Interbull Bulletin**, v. 42, p. 52-55, 2010.

PARK, T.; CASELLA, G. The Bayesian LASSO. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, 2008. DOI: 10.1198/016214508000000337

PEREZ, P.; CAMPOS, G; CROSSA, J.; GIANOLA, D. Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. **Plant Genome**, v. 3, n. 2, p. 106–116, 2010.

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

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. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. v. 1. 561 p.

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

RESENDE M. D. V.; RESENDE JUNIOR, M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGGIA A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 2010. 79 p.

RESENDE JR., M. F. R. ; VALLE, P. R. M. ; RESENDE, M. D. V. ; GARRICK, D. J. ; FERNANDO, R. L. ; DAVIS, J. M. ; JOKELA, E. J. ; MARTIN, T. A. ; PETER, G. F. ; KIRST, M. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, Austin, v. 190, 2012. DOI: 10.1534/genetics.111.137026

SALINAS, S. R. A. **Introdução à física estatística**. 2. ed. São Paulo: EDUSP, 2005. 462 p.

SILVA, F. F. E.; VARONA, L.; RESENDE, M. D. V.; BUENO FILHO, J. S. S.; ROSA, G. J. M.; VIANA, J. M. S. A note on accuracy of Bayesian LASSO regression in GWS. **Livestock Science**, New York, v. 141, n. 1-3, p. 310-314, Dec. 2011. DOI:10.1016/j.livsci.2011.09.010.

SINGER, J. M.; STANEK, E. J.; LENCINA, V. B.; GONZÁLEZD, L. M.; LIE, W.; MARTIN, S. S. Prediction with measurement errors in finite populations. **Statistics and Probability Letters**, Amsterdam, v. 82, n. 2, Feb. 2011. DOI: 10.1016/j.spl.2011.10.013.

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, London, v. 41, n. 29, 2009. DOI:10.1186/1297-9686-41-29.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag, 2002. 740 p.

TIBSHIRANI, R. Regression *shrinkage* and selection via the Lasso. **Journal of the Royal Statistics Society Series B**, Oxford, v. 58, p.267-288, 1996.

USAI, M. G; GODDARD, M. E.; HAYES, B. J. LASSO with cross-validation for genomic selection. **Genetics Research**, Cambridge, v. 91, n. 6, p. 427-36, Dec. 2009 .

VIANA, J. M. S. **RealBreeding**. Viçosa: UFV, 2011.

VAZQUEZ, A. I.; ROSA, G. J.; WEIGEL, K. A.; CAMPOS, G. de los; GIANOLA, D.; ALLISON, D. B. Predictive ability of subsets of SNP with and without parent average for several traits in US Holsteins. **Journal of Dairy Science**, Champaign, v. 93, n. 1, p. 5942-5949, 2010. DOI: 10.3168/jds.2010-3335.

VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era: concepts and misconceptions. **Nature Reviews Genetics**, London, v. 9, p. 255-266, 2008.

VISSCHER, P. M.; MEDLAND, S. E.; FERREIRA, M. A. R.; MORLEY, K. I.; ZHU G.; CORNES, B. K.; MONTGOMERY, G. W.; MARTIN, N. G. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. **PLoS Genetics**, San Francisco, CA, v. 2, n. 3, e41, 2006. DOI: 10.1371/journal.pgen.0020041.

VISSCHER, P. M.; YANG, J.; GODDARD, M. E. A commentary on "Common SNPs explain a large proportion of the heritability for human height" by Yang *et al*. (2010). **Twin Research and Human Genetics**, v. 13, n. 6, p. 517–524, 2010.

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

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics, **Chemometrics and Intelligent Laboratory Systems**, Amsterdam, v. 58, 109–130, 2001.

WRAY, N. R. Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. **Twin Research and Human Genetics**, v. 8, p. 87-94, 2005.

WRAY, N. R.; GODDARD, M. E.; VISSCHER, P. M. Prediction of individual risk to disease from genome-wide association studies. **Genome Research**, New York, v. 17**,** p. 1520–1528, 2007.

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

YANG, J.; LEE, S. H.; GODDARD, M. E.; VISSCHER, P. M. GCTA: a tool for genome-wide complex trait analysis. **The American Journal of Human Genetics**, Chicago, v. 88, p. 76-82, 2011.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society B**, Oxford, v. 67, p. 301-320, 2005 .

Embrapa
Florestas

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 9678