UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA E BIOMETRIA

Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial

Marcos Deon Vilela de Resende
Fabyano Fonseca e Silva
Paulo Sávio Lopes
Camila Ferreira Azevedo

Disciplina EST792 - Métodos Estatísticos na Seleção Genômica Ampla

Citação: Resende, M.D.V.; Silva, F.F.; Lopes, P.S.; Azevedo, C.F. ***Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.*** **Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291 p. http://www.det.ufv.br/ppestbio/corpo_docente.php**

Viçosa – MG – 2012

ISBN: 978-85-89119-08-5

UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA E BIOMETRIA

Dados Internacionais de Catalogação na Publicação - CIP
Embrapa Florestas

Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência Bayesiana (MCMC), regressão aleatória multivariada (RRM) e estatística espacial [recurso eletrônico] / Marcos Deon Vilela de Resende ... [et al.].- Dados eletrônicos.- Viçosa, MG : Universidade Federal de Viçosa, 2012. 291 p.

Disciplina EST792 – Métodos Estatísticos na Seleção Genômica Ampla.

Sistema requerido: Adobe Acrobat Reader.
Modo de acesso: World Wide Web.
<http://www.det.ufv.br/ppestbio/corpo_docente.php.pdf>
Título da página da web (acesso em 12 nov. 2012).

ISBN 978-85-89119-08-5

1. Estatística biométrica. 2. Seleção genômica. 3. Genética quantitativa. 4. Matemática computacional. I. Resende, Marcos Deon Vilela de. II. Silva, Fabyano Fonseca e. III. Lopes, Paulo Sávio. IV. Azevedo, Camila Ferreira.

CDD 519.5 (21. ed.)

UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA E BIOMETRIA

Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial

Marcos Deon Vilela de Resende[1]
Fabyano Fonseca e Silva[2]
Paulo Sávio Lopes[3]
Camila Ferreira Azevedo[4]

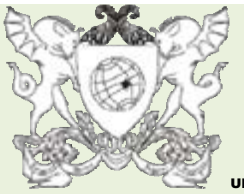
Apresentação

A Seleção Genômica veio unir a Genética de Populações à Genética Quantitativa. Estes dois ramos com forte orientação Biométrica tradicionalmente caminharam em separado, seja no Melhoramento Genético de Plantas e Animais ou na Genética Humana. Atualmente, a estimação de componentes da variação genética e de valores genéticos e a predição de fenótipos usa três conjuntos de dados ou informações: fenotípicos, genealógicos e genotípicos em locos marcadores moleculares em desequilíbrio de ligação com os genes de interesse. Ferramentas da Genética de Populações participam plenamente dos métodos de estimação atualmente empregados. Dessa forma, Genética de Populações, Genética Quantitativa, Genética Molecular e Estatística são demandados simultaneamente na análise genética dos caracteres de interesse. Esse texto aborda a nova Genética Quantitativa do terceiro milênio.

Viçosa – MG – 2012.

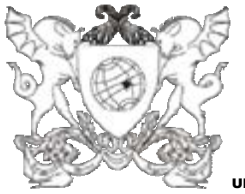
Os autores.

1Estatístico, Pós-Doutor em Estatística Biométrica e Estatística Genética (Inglaterra)
2Zootecnista, Pós-Doutor em Estatística Biométrica e Estatística Genética (USA)
3Zootecnista, Pós-Doutor em Genética Quantitativa e Melhoramento Animal (USA)
4Matemática, Mestre em Estatística Aplicada e Biometria (UFV)



UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA E BIOMETRIA

Sumário



UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA APLICADA E BIOMETRIA

1 Fundamentos Estatísticos da Seleção Genética

O melhoramento genético de animais e plantas fundamenta-se em duas ações: a identificação de indivíduos superiores; a criação de novas combinações genotípicas superiores por meio do cruzamento entre esses indivíduos elites. Em ambas as etapas a seleção tem papel fundamental e é realizada com base na avaliação genética dos indivíduos, a qual tem dois objetivos: (i) inferir sobre os valores genéticos dos indivíduos; (ii) ordenar os indivíduos com base em seus valores genéticos.

1.1 Propriedades dos Estimadores em Inferência Estatística

A escolha de um método ótimo de estimação/predição de valores genéticos deve basear-se no critério de uma inferência mais precisa e realista possível, a qual deve ser avaliada segundo parâmetros estatísticos adequados. Nesse contexto, os parâmetros mais importantes são a acurácia seletiva e o erro quadrático médio de estimação. A acurácia é conceituada como a correlação entre o valor genético verdadeiro e aquele estimado a partir das informações genotípicas (marcadores) e/ou fenotípica dos indivíduos. Um estimador acurado apresenta menor diferença quadrática entre valores verdadeiros e estimados, ou seja, apresenta mínimo erro quadrático médio (EQM) de estimação. A Tabela 1 ilustra essa questão.

Tabela 1. Ilustração de cálculo da acurácia e do erro de predição de valores genéticos a partir de dados simulados.

Individuo	Valor Genético Real (g)	Valor Genético Predito ($\hat{g}$)	Erro de Predição (%)($\hat{g} - g$)
1	65.929	71.716	0.09
2	76.213	74.242	0.03
3	55.333	62.620	0.13
4	54.678	60.012	0.10
5	74.766	76.409	0.02
6	92.742	75.515	0.19
7	81.241	76.785	0.05
8	62.385	72.929	0.17
9	83.280	81.906	0.02
10	66.279	67.104	0.01
11	59.107	63.747	0.08
12	63.325	64.381	0.02
13	60.807	68.552	0.13
14	66.864	65.872	0.01
15	78.432	67.242	0.14
16	54.042	56.527	0.05
17	75.274	77.499	0.03
18	86.995	76.232	0.12
19	72.250	78.856	0.09
20	80.547	70.806	0.12
			Erro Médio de Predição 0.08
			Correlação ou Acurácia 0.78

No exemplo apresentado, o erro médio de predição foi de 8 % e a correlação entre os valores verdadeiros e aqueles preditos foi de 78 %. Esse é o valor da acurácia seletiva ($r_{\hat{g}g}$) e seu quadrado ($r^2_{\hat{g}g}$) é denominado confiabilidade, confiança ou fidúcia seletiva. O valor genético estimado equivale ao verdadeiro mais o erro de predição, ou seja, $\hat{g} = g + (\hat{g} - g)$.

Um método ótimo de estimação/predição deve apresentar mínimo EQM, o qual é dado por EQM = Vício² + Precisão = Vício² + PEV. Assim, um estimador de mínimo EQM apresenta vício nulo ou baixo e alta precisão (baixa variância do erro de predição – PEV ou $Var(\hat{g} - g)$). Em ausência de vício, EQM = PEV. A Figura 1 ilustra os conceitos de vício, precisão e acurácia (Resende, 2008; Peternelli et al., 2011).

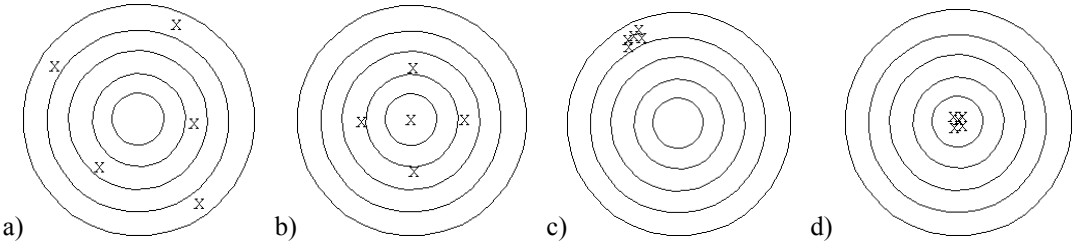


Figura 1: ilustração dos conceitos de acurácia, precisão e vício. (a): alto vicio, baixa precisão, baixa acurácia; (b): baixo vício, baixa precisão, baixa acurácia; (c): alto vício, alta precisão, baixa acurácia; (d): baixo vício, alta pecisão, alta acurácia.

Verifica-se pela Figura 1 que a alta acurácia (capacidade de acertar o alvo da predição nas várias tentativas) é uma combinação de alta precisão (baixa variação nas várias tentativas) e baixo vício (média das várias tentativas igual ao alvo da predição). Em outras palavras, pode-se dizer que a acurácia é a capacidade de acessar a verdade, e a precisão é a capacidade de acessar sempre a mesma estória mas não necessariamente a verdade. A acurácia e a precisão guardam entre si as seguintes relações:

- Acurácia ($r_{\hat{g}g}$)

$$r_{\hat{g}g} = [1 - PEV / \sigma^2_g]^{1/2}$$

- Precisão (PEV)

$$PEV = Var(\hat{g} - g) = (1 - r^2_{\hat{g}g})\ \sigma^2_g$$

A raiz quadrada da PEV equivale ao desvio padrão do erro de predição e pode ser usada para cômputo do intervalo de confiança do efeito genético (g) predito, por meio da expressão: $\hat{g}_i \pm t[Var(\hat{g} - g)]^{1/2}$ ou $\hat{g}_i \pm t[(1 - r^2_{\hat{g}g_i})\ \sigma^2_g]^{1/2}$, em que t é um valor tabelado (1,96) associado à distribuição t de Student a 95 % de confiança na inferência e σ^2_g é a variância genética aditiva da população.

A estimação da PEV com base na inversa da matriz dos coeficientes das equações de modelo misto é apresentada a seguir, com base em Resende (2002). A matriz dos coeficientes das equações do modelo misto $y = Xb + Zg + e$ equivale a

$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix}$ e a inversa generalizada de C é igual a $C^{-} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$, em que y, b e g são vetores de dados, efeitos fixos e genéticos aditivos, respectivamente, os últimos com matrizes de incidência X e Z. Tem-se também $Var(g) = A\sigma_g^2$, em que é uma matriz de correlação entre os elementos de g. O estimador da variância do erro de predição (PEV) dos efeitos genéticos é dado por $PEV = Var(g - \hat{g}) = C^{22}\sigma_e^2$.

Assim, a acurácia pode ser estimada por: $r_{\hat{g}g} = [1 - PEV / \sigma_g^2]^{1/2}$. Especificamente para um indivíduo *i*, tem-se:

$PEV_i = d_i\ \sigma_e^2 = (1 - r^2_{\hat{g}g_i})\ \sigma_g^2$

$r_{\hat{g}g_i} = (1 - d_i\ \sigma_e^2 / \sigma_g^2)^{1/2} = (1 - d_i\ \lambda)^{1/2}$, em que:

d_i : i-ésimo elemento da diagonal de C^{22}.

$\lambda = \dfrac{\sigma_e^2}{\sigma_g^2} = \dfrac{1 - h^2}{h^2}$.

O desvio padrão do erro de predição do valor genético do indivíduo *i* é dado por: $SEP_i = [d_i\ \sigma_e^2]^{1/2} = [(1 - r^2_{\hat{g}g_i})\ \sigma_g^2]^{1/2}$.

É importante relatar que para o caso em que $R \neq I\ \sigma_e^2$ e as equações de modelo misto não são simplificadas em relação ao termo R^{-1}, tem-se $PEV_i = C_i^{22}$.

Em inferência estatística, os estimadores devem apresentar as seguintes propriedades desejáveis:

a) **Não vício**, tal que a esperança matemática do estimador seja o próprio parâmetro.
b) **Consistência**, tal que, com o aumento do tamanho da amostra, a esperança do estimador convirja para o parâmetro e a variância do estimador, para zero (associado ao conceito de convergência em probabilidade): refere-se ao aumento da acurácia de uma estimativa com o aumento do tamanho da amostra.
c) **Eficiência**, tal que o estimador apresente variância mínima.
d) **Suficiência**, tal que o estimador condense o máximo possível a informação contida na amostra e não seja função (dependente) do parâmetro.
e) **Completitude** que está ligada à unicidade do estimador.
f) **Invariância à translação**, tal que a estimação dos componentes de variância não seja afetada por mudanças nos efeitos fixos.
g) **Admissibilidade**, que implica mínimo EQM global.
h) **Acurácia**, que congrega baixo vício e baixa PEV (alta precisão).
i) **Interpretabilidade:** complexidade mínima após seleção de covariáveis.
j) **Regularidade:** estimação sem consumir graus de liberdade.
k) **Estimabilidade:** possibilidade de estimação dados o método estatístico de estimação e o conjunto de dados (o conceito de estimabilidade envolve conceitos de estimação não tendenciosa e unicidade).
l) **Parcimônia:** eficácia com o menor número possível de parâmetros no modelo.

m) **Propriedade Shrinkage:** regressão ou penalização (ditada pelo tamanho da amostra e das variações dos efeitos aleatórios e residuais) e economia de graus de liberdade.
n) **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 (número de indivíduos) e n_m (número de covariáveis) tendem a infinito.
o) **Ajuste Válido**, produzindo estimativas no espaço paramétrico (variâncias positivas e herdabilidades entre zero e um).
p) **Identificabilidade**: solução única para os parâmetros do modelo.

Dessas propriedades, as mais importantes em conexão com a avaliação genética são apresentadas na Tabela 2. As demais são também importantes e serão invocadas em outras partes desse texto.

Tabela 2. Propriedades dos estimadores mais importantes em conexão com a avaliação genética.

Propriedades	Propriedades Componentes	Denominação das Propriedades
Admissibilidade	Viés^2 baixo + PEV mínima global	Não viés aproximado + eficiência global
Acurácia U	Viés 0 + PEV mínima na classe U	Não viés + eficiência local
Acurácia Global	Viés^2 baixo + PEV mínima global	Não viés aproximado + eficiência global
Interpretabilidade	Complexidade Mínima	Parcimônia
Regularidade	Estimabilidade + Ajuste válido	Shrinkage, economia de graus de liberdade

U: classe de estimadores não viesados.

O erro quadrático médio de predição equivale à distância Euclideana média entre os estimadores e os correspondentes parâmetros. Minimizar o erro quadrático médio significa maximizar a acurácia. Assim, o método ideal de estimação ou predição dos valores genotípicos é aquele que minimiza EQM. Verifica-se que tal método pode ser viciado em pequeno grau, pois o que importa é minimizar a soma $(Vício)^2 + PEV$. Na classe dos estimadores/preditores não viciados, a precisão é dada pelo parâmetro variância do erro de predição (PEV) e a estratégia de minimizar PEV conduz também à maximização da acurácia. Mas, de maneira geral (relaxando a necessidade de não vício), o que deve ser minimizado é o EQM, buscando a admissibilidade. Além da admissibilidade e acurácia, a interpretabilidade e a regularidade são relevantes, especialmente na seleção genômica.

1.2 Evolução dos métodos de avaliação genética

Em inferência estatística frequentista existem basicamente cinco classes de modelos de seleção. Fisher (1925) criou o método da ANOVA via quadrados mínimos ordinários (OLS) para a avaliação de variedades de cereais em delineamentos balanceados. O modelo genérico básico é dado por $y = Xb + e$, em que y é o vetor da variável resposta, b é o vetor de efeitos genéticos (fixos no caso) e e é o vetor de erros aleatórios com matriz de covariância $R = I\sigma_e^2$, caracterizando a Classe I de modelos de seleção. Nessa Classe I, os candidatos à seleção são de efeitos fixos, implicando na escolha entre tratamentos, representados por uma amostra aleatória de observações tomadas independentemente em cada tratamento.

A abordagem inicial da análise de dados desbalanceados é devida a Fisher e Yates, ambos trabalhando na *Rothamsted Experimental Station* na Inglaterra. Para este caso de representação desbalanceada, Yates (1934) apresentou as soluções de quadrados mínimos ponderados (WLS) para dois diferentes modelos de classificação cruzada. Nesse caso, matriz de covariância é diagonal dada por $R = I\sigma_{ei}^2$, em que σ_{ei}^2 é a variância do erro associada à observação i. Pela abordagem de Fisher e Yates os valores genéticos eram estimados como efeitos fixos.

Henderson et al. (1959) em um artigo influente apresentou estimadores de quadrados mínimos generalizados (GLS) de efeitos fixos contemplando a interferência de efeitos aleatórios (*g*) correlacionados na estimação daqueles efeitos. Nesse caso, o modelo é dado por $y = Xb + Zg + e$, em que X e Z são conhecidas matrizes de incidência. A matriz de covariância de y é dada por Var(y) = V = Var (g) + R = $Var(g) + I\sigma_e^2$ em que Var (g) pode ser não diagonal.

Na Classe II de modelos de seleção, a seleção envolve candidatos considerados como variáveis aleatórias não observáveis pertencentes a uma determinada população. Essa classe sempre foi considerada no melhoramento genético, associado aos índices de seleção envolvendo informações de parentes, desde o trabalho de Lush (1931). Sob esse modelo aleatório os preditores associados pertencem ao método BLP (melhor predição linear). O modelo (de médias) é dado por $y = Zg + e$, em que g é o vetor de valores genéticos, considerados como aleatórios. O BLP não especifica o que fazer com a média geral (u), o qual na prática tem sido estimada por OLS (Resende et al., 1993). Bueno Filho e Vencovsky (2009) relatam a utilidade do BLP no melhoramento vegetal.

O terceiro tipo de seleção foi negligenciado por estatísticos e melhoristas até o início da década de 1970. Essa Classe III de modelo de seleção, denominado Modelo Misto de Seleção (em analogia ao modelo misto de análise de variância), foi apresentada formalmente por Henderson (1973), contemplando o método BLUP (melhor predição linear não viesada). O modelo é dado por $y = Xb + Zg + e$, em que b é um vetor de efeitos fixos (efeitos ambientais identificáveis) e g é o vetor de efeitos genéticos, considerados como aleatórios. Neste caso, os candidatos à seleção são variáveis aleatórias não observáveis pertencentes a mais que uma população, e o mérito de cada candidato é a soma da média da população mais o valor predito da variável aleatória associada ao candidato. Neste caso, a seleção depende, também, de efeitos fixos desconhecidos. O modelo misto de seleção foi apresentado como BLUP por Henderson (1973), mas, foi concebido por volta de 1949 pelo próprio Henderson. Naquela época, Henderson derivou o método BLUP por meio da da maximização da função densidade de probabilidade conjunta de y (valores fenotípicos) e g (valores genéticos) (Henderson, 1973). A função maximizada não era uma função de verossimilhança e sim uma densidade conjunta.

Em termos mais rigorosos, a seleção é um problema puramente estatístico, visto que na prática seleciona-se uma fração de indivíduos segundo seus valores genéticos os quais seguem uma distribuição de probabilidade. Pearson (1903) derivou as médias e variâncias condicionais para a distribuição normal multivariada. Os

resultados de Pearson foram apresentados em notação matricial por Aitken (1934) e empregados por Henderson no contexto dos preditores BLUP, os quais podem ser vistos como valores genéticos condicionais a um conjunto de (N-r) funções lineares dos dados, linearmente independentes e invariantes `a translação, em que N é o número de observações e r é o posto de X, a matriz de incidência para os efeitos fixos. Os índices de seleção podem ser vistos como computações das médias condicionais dos valores genéticos dadas as observações. Lush (1931) foi o primeiro cientista a utilizar preditores de valores genéticos baseados em médias condicionais e Cochran (1951) estendeu as propriedades ótimas dos índices de seleção para quaisquer distribuições.

A média fenotípica, média aritmética ou média estimada pelo método de quadrados mínimos não é um estimador de mínimo EQM quando se tem mais que dois tratamentos ou materiais genéticos em avaliação. O trabalho de Stein (1955), que constituiu um verdadeiro paradoxo na Estatística, demonstrou que a média aritmética é estimador não admissível, isto é, que existem estimadores que propiciam menor erro quadrático médio ou menor risco que a média aritmética, quando mais que duas médias necessitam ser estimadas. Neste contexto, James e Stein (1961) apresentaram um estimador melhorado para a média populacional, que é dado por $M^* = k\ (\overline{Y}_{i..} - \overline{Y}_{...}) + \overline{Y}_{...}$, em que *k* é um fator regressor (ou de *shrinkage*) da média amostral de determinado tratamento ($\overline{Y}_{i.}$) sobre a média geral ($\overline{Y}_{...}$), em que $k = 1 - [(T-3)/(T-1)]/F$ e T é o número de genótipos em avaliação.

Os métodos (viciados ou não) que minimizam o EQM conduzem a estimadores/preditores do tipo *shrinkage*. Genericamente, um estimador do tipo *shrinkage* tem a forma de um escalar (variando entre zero e um) multiplicado por um vetor de médias estimadas por quadrados mínimos ou por máxima verossimilhança. Ou seja, para o caso balanceado, esse tipo de estimador multiplica as médias fenotípicas por um fator que varia entre zero e um, dependendo da confiabilidade (herdabilidade) que se tem nas médias fenotípicas estimadas.

Estimadores do tipo *shrinkage* começaram a ser usados por Lush (1931) no contexto do melhoramento animal associado ao método da melhor predição linear (BLP) e, posteriormente, foram também usados no método da melhor predição linear não viciada (BLUP) conforme Henderson (1973; 1975) e Thompson (1976; 1979). Esses métodos assumem os efeitos de materiais genéticos como aleatórios e o BLUP é, adicionalmente, um preditor não viciado. Entretanto, conforme Stein (1955), para mais que dois tratamentos, estimadores do tipo *shrinkage* são necessários, independentemente se os efeitos forem tomados como fixos ou aleatórios. O estimador melhorado de James e Stein (1961) não necessita de qualquer suposição referente a efeitos fixos ou aleatórios, ou sobre as distribuições das médias a serem estimadas (Efron e Morris 1977) e pertencem à Classe IV de modelos de seleção. Requer apenas o relaxamento da suposição de não vício. Este estimador é viesado, mas tem menor erro quadrático médio que o estimador de quadrados mínimos, em determinada região do espaço paramétrico.

No contexto da avaliação genética, é importante relatar que o vício propiciado pelo estimador de James-Stein é pequeno e só pode existir quando o

número de tratamentos é baixo (inferior a dez). À medida que o número de tratamentos aumenta, o estimador viesado torna-se não viesado e, por isso, o estimador de James-Stein é denominado como "aproximadamente não viesado". Conforme Schaeffer (1999), a princípio, somente estimadores não viesados eram usados pelos estatísticos. Os desenvolvimentos teóricos, porém, evidenciaram que tais estimadores podem gerar estimativas fora do espaço paramétrico admissível. Assim, atualmente, procedimentos aproximadamente não viesados, desde que admissíveis (de mínimo erro quadrático médio), têm sido considerados como os ideais.

Os estimadores de James e Stein (1961) propiciam, com o aumento do número de tratamentos em avaliação, uma transição natural de um modelo de efeitos fixos para um modelo de efeitos aleatórios. E isso só depende do tamanho da população (número de tratamentos). Com grande número de tratamentos, os estimadores de James-Stein e o método BLUP (cujo regressor é $k = 1 - 1/F$) se equivalem (Tabela 3). Nesse caso, a metodologia BLUP é a melhor escolha pela facilidade de implementação e por poder ser estendida para o caso não balanceado. Quando o número de tratamentos é superior a cinco, o modelo se aproxima mais de aleatório (devendo-se usar o método BLUP) e, quando menor que cinco o modelo se aproxima mais de fixo (devendo-se usar o método de quadrados mínimos, cujo fator de regressão é igual a 1). Logicamente o estimador de James-Stein é o mais eficaz em qualquer das situações (Resende e Duarte, 2007).

Tabela 3. Valores dos regressores (de James-Stein) dos desvios das médias fenotípicas em relação à média geral, em experimentos balanceados, para obtenção de estimativas precisas de valores genéticos para diferentes números de tratamentos ou genitores na população.

Número de tratamentos	Regressor[1]	Número de tratamentos	Regressor
3	1 - 0,33/*F**	14	1 - 0,85/*F*
4	1 - 0,33/*F*	15	1 - 0,86/*F*
5	1 - 0,50/*F*	16	1 - 0,87/*F*
6	1 - 0,60/*F*	17	1 - 0,88/*F*
7	1 - 0,67/*F*	18	1 - 0,88/*F*
8	1 - 0,71/F	19	1 - 0,89/*F*
9	**1 - 0,75/*F***	20	1 - 0,89/*F*
10	1 - 0,78/*F*	**21**	**1 - 0,90/*F***
11	**1 - 0,80/*F***	38	1 – 0,95/*F*
12	1 - 0,82/*F*	135	1 – 0,99/*F*
13	1 - 0,83/*F*	400	1 – 1/*F*

[1]- *F**: *F* de Snedecor centrado em zero, sendo que esse regressor deve multiplicar diretamente a média fenotípica e não o desvio; *F*: *F* de Snedecor centrado na média geral

O procedimento de estimação bayesiana pertence à Classe V de modelos de seleção e foi recomendado para avalição genética por Gianola e Fernando (1986). O teorema de Bayes foi derivado em 1763 e, portanto, é bem mais antigo do que o método de Stein, e também minimiza o erro quadrático esperado. Por isso, o estimador de James-Stein é muito similar ao estimador de Bayes, tornando-se inclusive idênticos para grande número de tratamentos (Efron e Morris 1977). Por isso, são também denominados como estimadores de Bayes-Stein, Bayes empírico ou regra empírica de Bayes. Em inferência bayesiana não existe qualquer distinção entre efeitos fixos ou aleatórios, e os parâmetros

a serem estimados são considerados variáveis aleatórias que devem ser estimadas considerando as incertezas a elas associadas.

Na Tabela 4 é apresentada a evolução dos métodos de avaliação genética. Em cada linha da tabela o primeiro autor citado refere-se ao trabalho mais influente e os demais referem-se a trabalhos básicos e/ou teóricos que já haviam abordado o tema.

Tabela 4. Evolução dos métodos de estimação de componentes de médias (valores genéticos).

Observações em y são Variáveis Aleatórias				
Método	**Autores**	**Modelo**	**Estimador**	**Estrutura de Variâncias**
OLS	Fisher (1925)	Fixo $y = Xb + e$	$\hat{b} = (X'X)^{-1}X'y$	$e \sim N(0, I\sigma_e^2)$
WLS	Yates (1934)	Fixo $y = Xb + e$	$\hat{b} = (X'R^{-1}X)^{-1}X'R^{-1}y$	$e \sim N(0, R = I\sigma_{ei}^2)$
GLS	Henderson et al. (1959)	Fixo $y = Xb + Zg + e$	$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$	$y \sim N(Xb, V)$ $V = Var(g) + I\sigma_e^2$
BLP	Lush (1931; 1945); Pearson (1903); Aitken (1934)	Aleatório $y = Zg + e$	$\hat{g} = [Z'R^{-1}Z + (A\sigma_g^2)^{-1}]^{-1}Z'R^{-1}y$	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$
BLUP (A-BLUP)	Henderson (1973); Thompson (1976); Henderson (1949)	Misto $y = Xb + Zg + e$	$\hat{g} = [Z'R^{-1}Z + (A\sigma_g^2)^{-1}]^{-1}$ $Z'R^{-1}(Y - X\hat{b})$	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$
James-Stein	Efron e Morris (1977); James e Stein (1962); Stein (1955)	$y = Xb + Zg + e$	$\hat{g} = k\ (\overline{Y}_{i..} - \overline{Y}_{...}) + \overline{Y}_{...}$ $k = (1 - 1/F)$	$e \sim N(0, R = I\sigma_{ei}^2)$
MAP (Bayes)	Gianola e Fernando (1986); Fernando e Gianola (1986); Robertson (1955); Dempfle (1971); Bayes (1763)	Aleatório $y = Xb + Zg + e$	$P(g\|y) = \frac{P(y\|g)\ P(g)}{P(y)}$ $\hat{g} = [Z'R^{-1}Z + (A\sigma_g^2)^{-1}]^{-1}$ $Z'R^{-1}(Y - X\hat{b})$	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$ $b \sim N(0, I\sigma_b^2)$ $\sigma_b^2 \to \infty$
MAS (LE e LD) via OLS e BLUP	Lande e Thompson (1990, OLS);	Fixo $y = u + Zg + \sum_{i=1}^{s} Q_i q_i + e$ ou $y = u + Zg + \sum_{i=1}^{s} W_i m_i + e$	$\hat{g} = Zg + \sum_{i=1}^{s} W_i m_i$ s é o número de marcas significativas	$e \sim N(0, I\sigma_e^2)$
	Fernando e Grossman (1989); Goddard (1991)	Misto $y = Xb + Zg + \sum_i Q_i q_i + e$ ou $y = Xb + Zg + \sum_i W_i m_i + e$	$\hat{g} = Zg + \sum_i W_i m_i$	$e \sim N(0, I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$
GWS (RR-BLUP); GBLUP; Bayes; RR-BLUP_B)	Meuwissen et al. (2001); Whittaker et al. (2000); Van Raden (2008); Nejati-Javaremi et al. (1997); Resende et al. (2010); Resende Jr. et al. (2012)	Misto $y = Xb + \sum_i Q_i q_i + e$ ou $y = Xb + \sum_i W_i m_i + e$ ou $y = Xb + Z\sum_i W_i m_i + e$ ou $y = Xb + ZWm + e$	$\hat{g} = W\hat{m} = W(W'R^{-1}W + I\lambda)^{-1}W'R^{-1}(y - X\hat{b})$ em_que $\hat{m} = (W'R^{-1}W + I\lambda)^{-1}W'R^{-1}(y - X\hat{b})$ ou $\hat{m} = (W'W + I\lambda)^{-1}W'(y - X\hat{b})$ $\hat{g} = [R^{-1} + G^{-1}(\sigma_e^2/\sigma_g^2)]^{-1}R^{-1}(y - X\hat{b})$ em_que $A = G = (WW')/[2\sum_i^n p_i(1-p_i)]$	$e \sim N(0, I\sigma_e^2)$ $g \sim N(0, G\sigma_g^2)$ $G = (WW')/[2\sum_i^n p_i(1-p_i)]$
Observações em y são Variáveis Mistas (Aleatórias + Determinísticas)				
Método	**Autores**	**Modelo**	**Estimador**	**Estrutura de Variâncias**
Modelos Espaciais: Krigagem e Autoregressivos	Matheron (1971); Robinson (1991); Gilmour et al. (1995)	Misto $y = Xb + Zg + e$	$\hat{g} = [Z'\Sigma^{-1}Z + (A\sigma_g^2)^{-1}]^{-1}$ $Z'\Sigma^{-1}(Y - X\hat{b})$	$e \sim N(0, \Sigma)$ $g \sim N(0, A\sigma_g^2)$ $\Sigma = \sigma_\varepsilon^2[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)]$

				$\sum_c(\Phi_c)=\begin{bmatrix}1 & \rho^1 & \rho^2 & \rho^3\\ \rho^1 & 1 & \rho^1 & \rho^2\\ \rho^2 & \rho^1 & 1 & \rho^1\\ \rho^3 & \rho^2 & \rho^1 & 1\end{bmatrix}$
Modelos de Regressão Aleatória Multivariada: Polinômios de Legendre, Splines cúbicas e B	Schaeffer e Dekkers (1994); White et al. (1999); Meyer (2005)	Misto $y = Xb + Zg + e$	$\hat{g} = [Z'\Sigma^{-1}Z + (A \otimes K_g)^{-1}]^{-1}$ $Z'\Sigma^{-1}(Y - X\hat{b})$	$e \sim N(0, \Sigma)$ $g \sim N(0, A \otimes K_g)$ $\Sigma = \sigma_e^2[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)]$
Modelos de Competição: Efeitos Associativos ou Indiretos	Resende et al. (2005) Van Vleck e Cassady (2005); Arango et al. (2005);	Misto $y = Xb + Zg + e$ $y = Xb + Z\tau + NZ\phi + e$ $y = Xb + Z\tau + NZ\phi + \xi + \eta$	$\hat{g} = [Z'\Psi^{-1}Z + G^{*-1}]^{-1}$ $Z'\Psi^{-1}(Y - X\hat{b})$	$\Psi = \sigma_\xi^2[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)] + I\sigma_\eta^2$ $G^* = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$

A: matriz de correlação genética aditiva construída via pedigree; G: matriz de correlação genética aditiva construída via marcadores. Notação: Vetores y, b, g, m, q: referentes aos dados fenotípicos, efeitos fixos, genéticos aditivos poligênicos aleatórios, genéticos aditivos aleatórios de marcadores, genéticos aditivos aleatórios de QTL, respectivamente, com variâncias 0, σ_g^2, σ_m^2 e σ_q^2. Matrizes X, Z, W, Q: incidência para b, g, m, q, respectivamente.

Na Tabela 5 é apresentada a evolução na forma de consideração do modelo genético associado aos caracteres quantitativos nos métodos de avaliação genética.

Tabela 5. Evolução na forma de consideração do modelo genético associado aos caracteres quantitativos nos métodos de avaliação genética.

Modelo	Efeitos	Método de Seleção	Autores
Poligênico Infinitesimal	Pequenos - Infinitos	BLUP	Fisher (1918)
Misto de Herança: genes maiores + poligênico residual	Grandes + Pequenos Infinitos	LE - MAS	Fernando e Grossman (1989)
Misto de Herança: genes maiores + poligênico residual	Grandes + Pequenos Finitos (segregando dentro de famílias)	LE - MAS	Fernando et al. (1994)
Misto de Herança: genes maiores + poligênico residual	Grandes + Pequenos Finitos (segregando na população: entre famílias)	LD - MAS e GWS	Meuwissen et al. (2001)

O modelo linear misto convencional contempla os efeitos fixos (b), genéticos aleatórios (g) e ambientais aleatórios (e) por meio de $y = Xb + Zg + e$ (Modelo Individual). Incluindo os efeitos (q) dos QTLs de grandes efeitos para os locos *i*, o modelo torna-se $y = Xb + Zg^* + \sum_i Q_i q_i + e$ (Modelo de QTL), quando se conhecem os genes ou $y = Xb + Zg^* + \sum_i W_i m_i + e$ quando se conhecem apenas os marcadores, em que Q_i é uma matriz de incidência que relaciona os indivíduos com os alelos do loco *i*, e q_i e m_i contém os efeitos alélicos para cada loco gênico e marcador, respectivamente. As matrizes de incidência Q não são conhecidas e nem as suas dimensões, dadas pelo número de alelos em cada loco. Também não é conhecido o número de locos que afeta o caráter. Isto contrasta com o primeiro modelo, em que as matrizes de incidência para *b* e *g* (X e Z, respectivamente) são conhecidas. Se Q fosse conhecida as equações de modelo misto poderiam ser usadas sem qualquer alteração. Um outro modelo melhor seria $y = Xb + \sum_i Q_i q_i + e$ ou $y = Xb + Z\sum_i W_i m_i + e$ (Modelo GWS), no qual todos os locos seriam individualizados e não haveria necessidade de inclusão do resíduo genético poligênico ou infinitesimal (g^*).

O que torna a análise genômica diferenciada é o fato da matriz Q ser desconhecida. No entanto, ela pode ser estimada com base nas informações dos marcadores (matriz W). Segundo Perez-Enciso e Misztal (2004), a forma como os marcadores são usados para estimar Q e a forma de definição de *q* resulta em distintos modelos que contemplam os vários delineamentos para a análise de QTLs e formas de seleção genômica.

Whittaker et al. (2000) e Meuwissen et al. (2001) foram pioneiros em propor a predição simultânea dos efeitos dos marcadores, sem o uso de testes de significância para marcas individuais. Isto contrasta com o método da MAS proposto por Lande e Thompson (1990). Uma comparação entre as três proposições pode ser vista na Tabela 6.

Tabela 6. Comparação entre as três proposições de seleção auxiliada por marcadores.

Autores	Método	População	Número de Marcadores (n_m)	Teste de Significância	Extensao para o Enfoque Bayesiano
Lande e Thompson (1990)	MAS – Indice de Seleção Reg. Mult.	Dentro de família ou cruzamento	Muito menor que tamanho do cruzamento (N): $n_m << N$	Sim	Não
Whittaker et al. (2000)	MAS – Ridge Regression	Dentro de família ou cruzamento	Maior ou igual ao tamanho do cruzamento (N): $n_m >= N$	Não	Não
Meuwissen et al. (2001)	GWS – RR-BLUP	Toda a População	Muito maior que tamanho da população de estimação (N): $n_m >> N$	Não	Sim

Verifica-se pela Tabela 6, que a inovação de Meuwissen et al. (2001) não foi em termos de metodologia estatística mas, em termos conceituais enfatizando o uso do conceito de desequilíbrio de ligação em nível populacional e não apenas dentro de família e o não uso de testes de significância para marcas. E o maior mérito foi a demonstração, via simulação, do fato de que a GWS pode realmente funcionar na prática. Por outro lado, a versão G-BLUP da GWS, enfatizando a troca da matriz A pela G no BLUP tradicional (Van Raden, 2008) já havia sido proposta por Nejati-Javaremi et al. (1997) e Fernando (1998).

O não uso de significância estatística para a seleção de marcas pela GWS a distingue da GWAS (Genome Wide Association Studies), a qual procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses visando detectar efeitos com significância estatística. A GWAS sofre com a alta taxa de falsos negativos devido ao uso de pontos de corte muito rigorosos visando evitar a ocorrência de falsos positivos. A GWS equivale à GWAS aplicada sobre todos os locos simultaneamente e baseando-se em estimação e predição em vez de teste de hipótese. Dessa forma consegue explicar parte muito maior da variabilidade genética e evitar a chamada herdabilidade faltante ou perdida (missing heritability), típica dos estudos de análise de ligação e de associação.

1.3 Modelos Estatísticos Lineares

Os modelos estatísticos lineares tem a forma geral $y = u + b + g + e$, em que u é uma constante ou média geral, b é um fator de blocagem cujos níveis são efeitos fixos ou aleatórios, g é um fator de tratamentos cujos níveis são efeitos fixos ou aleatórios e e é um erro aleatório. Esses modelos podem ser classificados em:

- Modelo Fixo: todos os fatores possuem níveis com efeitos fixos, exceto o erro aleatório (e).
- Modelo Aleatório: todos os fatores possuem níveis com efeitos aleatórios, exceto a média geral (u).
- Modelo Misto: possui efeitos fixos, além da média geral, e efeitos aleatórios além do erro experimental.

A natureza dos efeitos estatísticos pode ser definida:

- Fator de efeitos fixos: os níveis são constantes; são escolhidos; a inferência é válida para os níveis em estudo; a informação entre níveis não afeta a estimação de cada nível.
- Fator de efeitos aleatórios: os níveis são variáveis aleatórias amostradas segundo uma distribuição de probabilidade; os níveis são amostras aleatórias de uma população; a inferência é válida para toda a população; a informação entre níveis afeta a estimação de cada nível.

No contexto dos modelos mistos, as seguintes regras práticas podem ser adotadas para a definição de efeitos fixos ou aleatórios, a qual depende de: (i) número de níveis do fator (com 38 níveis o modelo aproxima 95% ao modelo aleatório, conforme a Tabela 3); (ii) tamanho de cada nível do fator (com 5 indivíduos de cada genitor em cada nível, 15% da variação genética fica retida entre níveis ou grupos e para utilizá-la deve-se tomar o fator grupo como de efeitos aleatórios); (iii) magnitude da variação entre níveis do fator em relação à variação residual (à medida que o coeficiente de determinação c^2 do fator tende a 1, o modelo tende de aleatório para fixo); (iv) presença de tratamento preferencial aos melhores indivíduos, caso em que os grupos de indivíduos devem ser tratados como de efeitos fixos, explorando a propriedade do Blup de invariância à translação nos efeitos fixos.

Força relativa dos efeitos fixos e efeitos aleatórios com matrizes de correlação A e I

Os efeitos fixos dominam efeitos aleatórios com matriz de correlação A e I. Efeitos aleatórios com matriz de correlação A dominam efeitos aleatórios com matriz de correlação I. Isto é ilustrado a seguir.

(A) - Modelo de reprodutor: ajustes não concorrentes

Efeitos fixos – Pop (p)	Touro (t)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Touro
1	1	11	200.10	$y = 1u + Tt + e$	Aleatório em t
1	2	12	160.50		Aleatório em t
1	2	13	302.45	$t \sim N(0, I\sigma_t^2)$	Aleatório em t
1	3	14	112.67		Aleatório em t
1	3	15	145.89		Aleatório em t

(B) - Modelo de reprodutor: ajustes concorrentes: efeitos fixos dominam efeitos aleatórios com matriz de correlação I: $0^{-1} > I^{-1}$

Efeitos fixos – Pop (p)	Touro (t)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Touro
1	1	11	200.10	$y = Xp + Tt + e$	Fixo em p e zero em t
2	2	12	160.50		Aleatório em t
2	2	13	302.45	$t \sim N(0, I\sigma_t^2)$	Aleatório em t
2	3	14	112.67		Aleatório em t
2	3	15	145.89		Aleatório em t

(C) - Modelo individual ou animal: ajustes concorrentes: efeitos fixos dominam efeitos aleatórios com matriz de correlação A: $0^{-1} > A^{-1}$**. O indivíduo 11 terá seu efeito genético predito em g mas o valor refere-se somente à parte dentro de família.**

Efeitos fixos – Pop (p)	Touro (t)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Touro
1	1	11	200.10	$y = Xp + Zg + e$	Fixo em p e zero em g
2	2	12	160.50		Aleatório em g
2	2	13	302.45	$g \sim N(0, A\sigma_g^2)$	Aleatório em g
2	3	14	112.67		Aleatório em g
2	3	15	145.89		Aleatório em g

(D) - Modelo individual ou animal: ajustes concorrentes: efeitos aleatórios com matriz de correlação A dominam efeitos aleatórios com matriz de correlação I: $A^{-1} > I^{-1}$**. Nesse caso, o vetor estimado t conterá apenas valores zero.**

Efeitos fixos – Pop (p)	Touro (t)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Touro
1	1	11	200.10	$y = 1u + Tt + Zg + e$	Aleatório em g
1	2	12	160.50		Aleatório em g
1	2	13	302.45	$g \sim N(0, A\sigma_g^2)$	Aleatório em g
1	3	14	112.67		Aleatório em g
1	3	15	145.89	$t \sim N(0, I\sigma_t^2)$	Aleatório em g

(E) - Modelo individual ou animal: ajustes concorrentes: efeitos fixos dominam efeitos aleatórios com matrizes de correlação A e I simultaneamente: $0^{-1} > A^{-1}\ e\ I^{-1}$

Efeitos fixos – Pop (p)	Touro (t)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Touro
1	1	11	200.10	$y = Xp + Tt + Zg + e$	Fixo em p e zero em g e t
2	2	12	160.50		Aleatório em g
2	2	13	302.45	$g \sim N(0, A\sigma_g^2)$	Aleatório em g
2	3	14	112.67		Aleatório em g
2	3	15	145.89	$t \sim N(0, I\sigma_t^2)$	Aleatório em g

(F) - Modelo de famílias de irmãos completos: ajustes não concorrentes: o vetor f estima os efeitos de família contemplando $\sigma_f^2 = (1/2)\sigma_g^2 + (1/4)\sigma_{do\,\mathrm{min}ancia}^2$**.**

Efeitos fixos – Pop (p)	Famíla (f)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para Família
1	1	11	200.10	$y = 1u + Ff + e$	Aleatório em f
1	2	12	160.50		Aleatório em f
1	2	13	302.45	$f \sim N(0, I\sigma_f^2)$	Aleatório em f
1	3	14	112.67		Aleatório em f
1	3	15	145.89		Aleatório em f

(G) - Modelo individual ou animal com famílias de irmãos completos: ajustes concorrentes: efeitos aleatórios com matriz de correlação A dominam efeitos aleatórios com matriz de correlação I: $A^{-1} > I^{-1}$**. O vetor f estima os efeitos da capacidade específica de combinação (CEC) associados a cada família, contemplando** $\sigma_f^2 = (1/4)\sigma_{do\,\mathrm{min}ancia}^2$**.**

Efeitos fixos – Pop (p)	Famíla (f)	Indivíduo (g)	Peso	Modelo Ajustado	Ajuste para CEC de Família
1	1	11	200.10	$y = 1u + Ff + Zg + e$	Aleatório em f
1	2	12	160.50		Aleatório em f
1	2	13	302.45	$g \sim N(0, A\sigma_g^2)$	Aleatório em f
1	3	14	112.67		Aleatório em f
1	3	15	145.89	$f \sim N(0, I\sigma_f^2)$	Aleatório em f

Assim, os efeitos associados à matriz de incidência X são mais fortes do que aqueles associados à matriz de incidência Z abrangendo os seguintes casos: $\begin{bmatrix} X'X+0^{-1}(\sigma_e^2/\sigma_b^2) & X'Z \\ Z'X & Z'Z+I^{-1}(\sigma_e^2/\sigma_g^2) \end{bmatrix}$; $\begin{bmatrix} X'X+0^{-1}(\sigma_e^2/\sigma_b^2) & X'Z \\ Z'X & Z'Z+A^{-1}(\sigma_e^2/\sigma_g^2) \end{bmatrix}$ e $\begin{bmatrix} X'X+A^{-1}(\sigma_e^2/\sigma_g^2) & X'Z \\ Z'X & Z'Z+I^{-1}(\sigma_e^2/\sigma_g^2) \end{bmatrix}$, em que A é uma matriz não diagonal de correlação entre valores genéticos aditivos, com elementos dados por a_{XY}, o numerador do coeficiente de parentesco de Wright entre os indivíduos X e Y dado pela correlação $r_{a\,XY} = \frac{a_{XY}}{(a_{XX}a_{YY})^{1/2}}$, em que $a_{XX} = 1+F$ é o parentesco do indivíduo com ele mesmo e F é o coeficiente de endogamia. Se F = 0, $r_{a\,XY} = a_{XY}$.

1.4 Modelos Estatísticos de Seleção

Os modelos estatísticos de seleção tem a forma geral $\hat{g} = f(y)$, em que $\hat{t}$ é um estimador dos efeitos de tratamentos genéticos e $y = u + b + g + e$. Os modelos estatísticos de seleção podem ser classificados em (Resende, 2008):

A) Estimadores não Viesados

(i) Modelo I (Fixo): tem como alvo a escolha entre tratamentos independentes e de efeitos fixos; assume implicitamente que g^2 = Var(g) / Var(y) = 1, ou seja, que o coeficiente de determinação dos efeitos de tratamento equivale a 100%; utiliza na seleção os procedimentos de comparação de médias fenotípicas estimadas por quadrados mínimos (OLS).

(ii) Modelo II (Aleatório): tem como alvo a seleção entre variáveis aleatórias não observáveis pertencentes a uma mesma população estatística (ambiente); assume g^2 = Var(g) / Var(y) = h^2, em que h^2 é a herdabilidade de cada nível do fator de tratamentos; utiliza na seleção o procedimento da melhor predição linear (BLP) ou índice de seleção (SI).

(iii) Modelo III (Misto): tem como alvo a seleção entre variáveis aleatórias não observáveis pertencentes a várias populações estatísticas (ambientes ou raças, de efeitos fixos); assume g^2 = Var(g) / Var(y) = h^2, em que h^2 é a herdabilidade de cada nível do fator de tratamentos; estima as médias das várias populações por quadrados mínimos generalizados (GLS), produzindo melhores estimativas lineares não viciadas (BLUE) dessas médias; utiliza na seleção o procedimento da melhor predição linear não viciada (BLUP).

O procedimento BLUP pode ser assim caracterizado:

- B: minimiza a variância do erro de predição (PEV), ou seja, maximiza a precisão.
- L: é uma função linear das observações.
- U: é não viciado, propriedade essa que, em conjunção com a minimização da PEV, maximiza a acurácia na classe dos preditores não viesados.
- P: preditor de uma variável aleatória.

As propriedades B e U, simultaneamente, caracterizam um procedimento acurado, na classe dos preditores não viesados. Assim, o BLUP poderia também ser traduzido como preditor linear acurado (ALP).

B) Estimadores Aproximadamente não Viesados

(iv) Modelo IV: tem como alvo a escolha entre tratamentos com coeficientes de determinação dados por $g^2 = 1-[(T-3)/(T-1)]/F$, em que T é o numero de níveis dos efeitos aleatórios g e F é a estatística F de Snedecor, função da proporção entre variância entre tratamentos e variância residual. Utiliza na seleção médias fenotípicas estimadas por quadrados mínimos (OLS) ponderadas pelo fator de shrinkage g^2 (Estimadores de James-Stein).

(v) Modelo V: tem como alvo a escolha entre variáveis aleatórias obtidas como médias a posteriori (MAP) de uma distribuição condicional dos valores genéticos dados o vetor de dados e os valores atualizados dos componentes de variância e efeitos fixos (Estimadores de Bayes ou MAP).

1.5 Métodos Estatísticos de Estimação

Os métodos estatísticos de estimação de componentes de média e de variância, associados aos cinco tipos de modelos estatísticos de seleção, são apresentados na Tabela 7.

Tabela 7. Métodos estatísticos de estimação de componentes de média e de variância e testes de hipóteses .

Modelo Estatístico Linear e de Seleção	Método de Estimação de Componentes de Médias	Método de Estimação de Componentes de Variância	Teste da Significância dos Efeitos
Modelo I (Fixo)	Quadrados Mínimos (LS)	Quadrados Mínimos: Análise de Variância (ANOVA)	Teste F de Snedecor; Teste de Wald
Modelo II (Aleatório)	BLP ou BLUP	Máxima Verossimilhança (ML) ou ML Residual (REML): Análise de Deviance (ANADEV)	Teste LRT via Qui - Quadrado
Modelo III (Misto)	BLUP	REML: Análise de Deviance (ANADEV)	Teste LRT via Qui-Quadrado
Modelo IV	James-Stein	Quadrados Mínimos: (OLS); Máxima Verossimilhança (ML)	Intervalo de Confiança
Modelo V	Bayes (MAP)	Moda a Posteriori (MAP) via MCMC	Intervalo Bayesiano de Credibilidade

Verifica-se uma sofisticação dos procedimentos quando se passa do modelo I para o modelo III e V. Uma ilustração de cálculos associados à análise de deviance é apresentada a seguir.

Na análise de modelos mistos com dados desbalanceados, os efeitos do modelo não são testados via testes F tal como se faz no método da análise de variância. Nesse caso, para os efeitos aleatórios, o teste cientificamente recomendado é o teste da razão de verossimilhança (LRT). Para os efeitos fixos, um teste F aproximado pode ser usado. Um quadro similar ao quadro da análise de variância pode ser elaborado. Tal quadro pode ser denominado de Análise de Deviance (ANADEV) e é estabelecido segundo os seguintes passos:

a) Obtenção do ponto de máximo do logaritmo da função de verossimilhança residual (Log L) para modelos com e sem o efeito a ser testado;
a) Obtenção da deviance D = -2 Log L para modelos com e sem o efeito a ser testado;
b) Fazer a diferença entre as deviances para modelos sem e com o efeito a ser testado, obtendo a razão de verossimilhança (LR);
c) Testar, via LRT, a significância dessa diferença usando o teste qui-quadrado com 1 grau de liberdade.

Considere como exemplo o seguinte experimento, conduzido no delineamento de blocos ao acaso com várias plantas por parcela. Tem-se então o seguinte modelo, y = u + g + b + gb + e, em que g refere-se ao efeito aleatório de genótipos, b refere-se ao efeito fixo de blocos, gb refere-se ao efeito aleatório de parcela e *e* refere-se ao resíduo aleatório dentro de parcela. A seguinte análise de deviance (ANADEV) pode ser realizada.

Efeito	Deviance	LRT(Qui-quadrado[d])	Comp.Var.	Coef. Determ.
Genótipos	647.1794[+]	6.5546**	0.032924*	h^2g = 0.0456*
Parcela	654.1289[+]	13.5041**	0.068492**	c^2parc = 0.0948**
Resíduo	-	-	0.6206	c^2res=0.8595
Modelo Completo	640.6248	-	-	c^2total=1.0000
Bloco	-	F = 7.0172**	-	-

Qui-quadrado tabelado: 3,84 e 6,63 para os níveis de significância de 5 % e 1 %, respectivamente..
+ Deviance do modelo ajustado sem os referidos efeitos
[d] Distribuição com 1 grau de liberdade.

Verifica-se que os efeitos de genótipos e de parcelas são significativos. Conseqüentemente, os respectivos componentes de variância são significativamente diferentes de zero, assim como os respectivos coeficientes de determinação (herdabilidade dos efeitos genotípicos – h^2g e coeficiente de determinação dos efeitos de parcela - c^2parc). O fator bloco, considerado de efeito fixo, foi testado via F de Snedecor. A análise de deviância é uma generalização (para os casos balanceado e desbalanceado) da clássica análise de variância.

1.6 Derivações Frequentistas e Bayesianas de Estimadores de Valores Genéticos

a. Minimização da soma de quadrados dos resíduos ou erros de estimação sob modelo de efeitos fixos e restrição U de não vicio (OLS).
b. Minimização da soma de quadrados ponderada (comtemplando heterocedasticia) dos resíduos sob modelo de efeitos fixos e restrição de não vicio (WLS).
c. Minimização da soma de quadrados ponderada (comtemplando heterocedasticia e erros correlacionados) dos resíduos sob modelo de efeitos fixos e restrição de não vicio (GLS).
d. Maximização da função de verossimilhança de y (ML; BLP empírico).
e. Minimização do erro quadrático médio de estimação sob modelo aleatório (BLP se os componentes de variância são conhecidos).
f. Maximização da acurácia: maximização da distribuição conjunta entre *g* e *y* (BLP se os componentes de variância são conhecidos).
g. Minimização do erro quadrático médio de estimação na classe U sob modelo misto (BLUP se os componentes de variância são conhecidos, Krigagem).
h. Maximização da acurácia na classe U: maximização (com respeito a *g* e *b*) da distribuição conjunta entre *g* e $(y - X\hat{b})$ (BLUP se os componentes de variância são conhecidos; BLP de *g* + GLS de *b*).
i. Maximização da função de verossimilhança restrita de $(y - X\hat{b})$ (REML; BLUP empírico).
j. Maximização da distribuição a posteriori de *g* dado *y* (MAP ou Bayes ou Média condicional a posteriori).
k. GWS: Maximização da acurácia na classe U: maximização da distribuição conjunta entre *g* e *m* (RR-BLUP e G-BLUP); m é um vetor dos efeitos de marcadores genéticos de DNA.
l. GWS: Maximização da distribuição a posteriori de *g* dado *m* (MAP ou Bayes ou Média condicional).

Existem duas formas frequentistas de derivação do BLUP: (i) pela minimização do erro quadrático médio de predição ($E\left[\sum_i (\hat{g}_i - g_i)^2\right]$) sob restrição de não vicio; (ii) pela maximização da função densidade de probabilidade conjunta do vetor de dados e do vetor de parâmetros. A forma (ii) é apresentada a seguir.

Modelo misto

$y = Xb + Zg + e$

Função Densidade de Probabilidade de y

$$f(y|Xb,V) = \frac{1}{2\pi^{(1/2)N}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y - Xb)'V^{-1}(y - Xb)\right\}$$

Função Densidade de Probabilidade Conjunta de *y* e *g*

$$f(y,g) = f(y|g).f(g)$$

$$= \frac{1}{2\pi^{(1/2)N}|R|^{1/2}} \exp\left\{-\frac{1}{2}(y - Xb - Zg)'R^{-1}(y - Xb - Zg)\right\}$$

$$\cdot \frac{1}{2\pi^{(1/2)q}|G|^{1/2}} \exp\left\{-\frac{1}{2}(g'G^{-1}g)\right\}$$

A função densidade de probabilidade conjunta de *y* e *g* é dada pelo produto entre a função densidade de probabilidade condicional de *y* dado *g* e a função densidade de probabilidade de *g*, ou seja, f(y,g) = f(y|g) . f(g). Maximizando essa função, por meio da derivação da mesma em relação a b e *g*, e tomando-se as derivadas identicamente nulas, obtêm-se as equações de modelo misto. É importante reafirmar que a função a ser maximizada é uma função densidade de probabilidade conjunta de *y* e dos parâmetros e não uma função de verossimilhança ($f(y|g)$). Detalhes dessa derivação são apresentados por Lopes et al. (1998) e Martins et al. (1997;1998).

A predição usando BLUP assume que os componentes de variância são conhecidos. Entretanto, na prática, são necessárias estimativas fidedignas dos componentes de variância (parâmetros genéticos) de forma a se obter o que se denomina BLUP empírico (Harville e Carriquiry, 1992). O procedimento recomendado para estimação de componentes de variância é o da máxima verossimilhança restrita (REML), desenvolvido por Patterson e Thompson (1971).

Teorema de Bayes (em termos de Eventos)

Probabilidade condicional: Se A e B são eventos em um dado espaço de probabilidade, a probabilidade condicional de um evento A dado o evento B, indicado por P[A|B] é definida por:

$$P[A|B] = \frac{P[A,B]}{P[B]} \text{ se } P[B] > 0,$$

Probabilidade Conjunta: a partir da fórmula da probabilidade condicional obtém-se a fórmula da probabilidade conjunta dada por

$P[A,B]=P[B] \,.\, P[A|B]=P[A] \,.\, P[B|A]$.

Teorema de probabilidade total: para um dado espaço de probabilidade se B_1, B_2, ..., B_n é uma coleção de eventos mutuamente disjuntos satisfazendo:

$\Omega = \bigcup_{j-1}^{n} B_j$ *e* $P[B_j] > 0$ *para* $j = 1, 2, ..., n$ *então*

$P[A] = \sum_{j-1}^{n} P[A|B_j] \cdot P[B_j] = P[A]$, em que Ω é o espaço amostral.

Teorema de Bayes: a partir da fórmula da probabilidade conjunta e da probabilidade total obtém-se:

$$P[B_k|A] = \frac{P[A,B]}{P[A]} = \frac{P[A|B_k] \cdot P[B_k]}{\sum_{j-1}^{n} P[A|B_j] \cdot P[B_j]} = \frac{\Pr \textit{obabilidad e Conjunta}}{\Pr \textit{obabilidad e Total}}.$$

Função Densidade de Probabilidade e Expectância

Uma variável aleatória contínua não possui uma função de probabilidade que associe probabilidades a cada ponto ou valores de seu domínio. Estas probabilidades são calculadas para intervalos de valores do domínio através de uma função densidade de probabilidade. A função f(Y) é uma função densidade de probabilidade desde que satisfaça às condições:

(i) $P\ (a<Y<b)=\int_a^b f(y)\ dy$ (ii) $\int_{-\infty}^{\infty} f(y)\ dy=1$

Uma variável com distribuição Normal ou Gaussiana com parâmetros μ (média) e σ^2 (variância), tem como função densidade de probabilidade:

$$f(y)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(y-u)^2\right\},\ y\in\Re,\ \mu\in\Re\ e\ \sigma>0$$

Formalmente, os momentos dos dados equivalem aos valores esperados de uma função de uma variável aleatória. Sendo Y uma variável aleatória e g(·) uma função com domínio e contradomínio reais, define-se expectância ou valor esperado g(·) da variável aleatória Y, a função E [g(Y)] dada por:

(i) $E[g(Y)]=\sum_Y g(Y)\cdot P_Y(y)$ se Y é uma variável aleatória discreta;

(ii) $E[g(Y)]=\int_{-\infty}^{\infty} g(Y)\ f_Y(y)\ dy$ se Y é uma variável aleatória contínua com função densidade de probabilidade $f_Y(y)$.

Assim, tem-se:

a) Se g (Y) = Y, então, E[g(Y)] = E(Y) = μ_Y: primeiro momento;
b) Se g (Y) = Y^2, então, E[g(Y)] = E(Y^2): segundo momento;
c) Se g (Y) = Y^3, então, E[g(Y)] = E(Y^3): terceiro momento;
d) Se g (Y) = Y^4, então, E[g(Y)] = E(Y^4): quarto momento;
e) Se g (Y) = (Y-o), então, E[g(Y)] = E(Y) = μ_Y: primeiro momento centrado em zero (média);
f) Se g (Y) = $(Y-\mu_Y)^2$, então, E[g(Y)] = E$(Y-\mu_Y)^2$=Var(Y): segundo momento centrado na média (variância).

Os momentos de uma variável aleatória ou de sua correspondente distribuição são as potências das esperanças. O r-ésimo momento de uma variável aleatória Y é usualmente indicado por M_r e definido por M_r = E(Y^r) se a esperança existe. O r-ésimo momento central de uma variável aleatória Y em torno de a é definido como E$[(Y-a)^r]$. Se a = μ_Y, tem-se o r-ésimo momento central de Y em torno da média μ_Y. Assim:

M_1 = E$[(Y-\mu_Y)]$ = o: primeiro momento central;
M_2 = E$[(Y-\mu_Y)^2]$ = Var (Y): segundo momento central.

A variância de uma variável aleatória Y com esperança E(Y) = μ_Y é definida por:

(i) $\sigma_Y^2 = Var(Y) = \sum_Y (Y - \mu_Y)^2 P_Y(y)$ se Y é discreta;

(ii) $\sigma_Y^2 = Var(Y) = \int_{-\infty}^{\infty} (Y - \mu_Y)^2 f_Y(y)\ dy$ se Y é contínua.

Função Densidade Marginal

Uma função densidade marginal de uma variável Y_1 com respeito à outra variável Y_2 refere-se aos valores assumidos por Y_1 independente dos valores assumidos por Y_2. Nesse caso, a distribuição marginal Y_1 é dada por $f(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\, dy_2$, donde se vê que y_2 é integrada (tendo eliminada a sua influência) na função. Assim, y_2 é considerada variável de distúrbio.

Função Densidade Condicional

Uma função densidade condicional de uma variável Y_1 com respeito à outra variável Y_2 refere-se aos valores assumidos por Y_1 quando Y_2 assume um valor constante. Nesse caso, a distribuição condicional é dada por $f(y_1|y_2) = f(y_1, y_2) / f(y_2)$, onde $f(y_2)$ é a densidade marginal da variável Y_2, a qual é fixada em um determinado valor.

A esperança condicional de Y_1 dado Y_2 é uma regressão de Y_1 em Y_2, dada por $E(Y_1|Y_2 = y_2) = \mu_1 + \beta(y_2 - \mu_2) = \mu_1 + (\sigma_{y1y2} / \sigma_{y2}^2)(y_2 - \mu_2) = \mu_1 + (conjunta_{\ y1y2} / marginal_{\ y2})(y_2 - \mu_2)$

Estimação Bayesiana

A estimação Bayesiana difere da estimação por máxima verossimilhança (ML) devido ao fato de se maximizar a distribuição *a posteriori* do parâmetro em vez da função de verossimilhança. Essa distribuição é dita condicional do parâmetro dadas as observações (y) e é proporcional ao produto da função de verossimilhança pela distribuição *a priori* do parâmetro. De maneira similar à ML, é possível também maximizar a função densidade *a posteriori* em relação aos parâmetros. Se a informação *a priori* encontra-se disponível a estimação Bayesiana deve ser preferível à ML.

O princípio bayesiano é atribuído postumamente (1763) a Thomas Bayes, que nunca publicou em vida um trabalho matemático. No entanto, a base desse princípio foi publicada antes por Saunderson (1683-1739), um cego professor de ótica, que publicou vários artigos matemáticos.

Ao invés de maximizar a distribuição *a posteriori*, uma alternativa é definir uma função de perda, como por exemplo as funções de perda linear e quadrática, as quais contemplam respectivamente as diferenças simples e quadráticas entre os valores estimados e os parâmétricos. Minimizar a função de perda linear equivale a maximizar a densidade *a posteriori* (obtendo a moda) e minimizar a função de perda quadrática equivale a maximizar a média da distribuição *a posteriori*. Se a distribuição a priori é não informativa (vaga) e/ou a quantidade de dados é muito grande (a

verossimilhança domina *a priori*), a estimação bayesiana converge para a estimação ML, ou seja, ambas são equivalentes.

O Teorema de Bayes, definido em termos de densidades de probabilidade, tem a seguinte formulação para a distribuição de uma variável aleatória contínua:

$$f(\theta|y) = \frac{f(y,\theta)}{f(y)} = \frac{f(y|\theta)\ f(\theta)}{\int_R f(y|\theta)\ f(\theta)\ d\ \theta}. \qquad \text{(1)}$$

θ: vetor de parâmetros

$f(\theta)$: função densidade de probabilidade da distribuição *a priori*, que é também a densidade marginal de θ. Esta função denota o grau de conhecimento acumulado sobre θ, antes da observação de y.

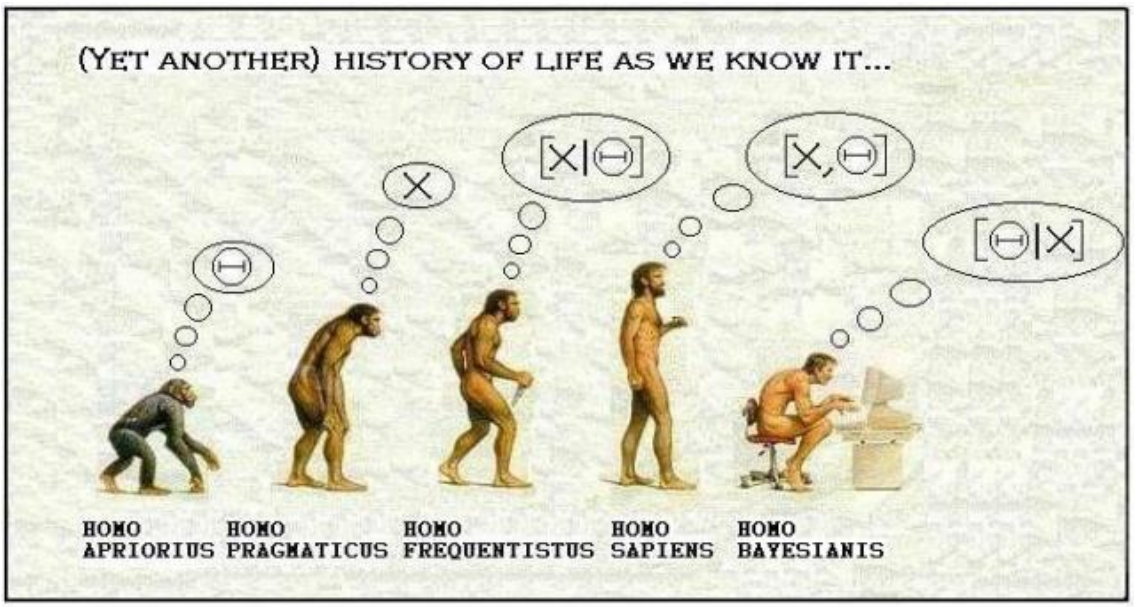
y : vetor de dados ou de informações obtidas por amostragem.

$f(y/\theta)$: função densidade de probabilidade da distribuição condicional de uma observação (y) dado θ (denominada função de verossimilhança ou modelo para os dados).

$f(y,\theta) = f(y/\theta)\ f(\theta)$: função densidade conjunta de y e θ.

$f(\theta/y)$: distribuição condicional de θ dado y, ou distribuição *a posteriori* (que é a base da estimação e predição bayesiana).

A Figura a seguir (em que y foi substituído por X) ilustra essas distribuições.



$f(y)=\int_R f(y,\theta)\ d\theta=\int_R f(y|\theta)\ f(\theta)\ d\theta=E_\theta\left[f(y|\theta)\right]$ - distribuição marginal ou preditiva de *y* com respeito a θ, onde R é a amplitude da distribuição de θ. E_θ significa esperança com respeito à distribuição de θ. (A integração da distribuição conjunta, no espaço paramétrico de θ, produz a marginal de *y*). A função $f(y)$ é denominada função de verossimilhança ponderada (por $f(\theta)$) sobre a distribuição de θ. A marginal de y é independente de θ, o qual é integrado para fora da função.

Como *f(y)* não é função de θ (ou seja, *f(y)* é constante para qualquer θ), a forma usual da formulação de Bayes é: $f(\theta/y)\ \alpha\ f(y/\theta)\ f(\theta)$, onde α indica proporcionalidade. Dessa forma, $f(\theta/y)$ não integra 1.

A expressão **(1)** advém das expressões $f(\theta,y) = f(y/\theta)\ f(\theta)$ e $f(\theta,y)=f(\theta/y)\ f(y)$, as quais são obtidas a partir do teorema da probabilidade condicional.

Em termos de estimação, enquanto para a estatística freqüentista podem existir vários estimadores para um determinado parâmetro, para a estatística bayesiana existe, em princípio, um único estimador, o qual conduz a estimativas que maximizam a função densidade de probabilidade *a posteriori*. Assim, inferências sobre θ são realizadas a partir da densidade *a posteriori* através da expressão geral $p(\theta|y)=\int_R f(\theta|y)\ d\theta$, onde *p* denota probabilidade (Gianola & Fernando, 1986).

Ao nível do –i-ésimo elemento do vetor θ, a esperança condicional de θ_i dado *y* é $\frac{\int_R \theta_i\ f(y|\theta)\ f(\theta)\ d\theta}{\int_R f(y|\theta)\ f(\theta)\ d\theta}$, o qual é o usual estimador bayesiano de θ_i.

Verifica-se que a predição dos valores genéticos ($\theta = g$), a partir dos dados fenotípicos (*y*), baseia-se na média condional ou regressão de *g* em *y*, dada por:

$E(g|y)=\int g\ f(y,g)\ dg/\int f(y,g)\ dg$, em que:

$f(y,g)$: função densidade da distribuição de probabilidade conjunta de *y* e *g*.

Com dados desbalanceados, independentemente da distribuição, o ordenamento dos candidatos com base em E(g|y) e a seleção daqueles com os maiores valores, maximiza a média dos indivíduos selecionados, conforme demonstrado por Fernando & Gianola (1986).

Em inferência bayesiana não existem parâmetros de efeitos fixos, mas apenas variáveis aleatórias. Tais variáveis são estimadas, diferentemente da abordagem frequentista, em que os efeitos aleatórios são preditos e os efeitos fixos e componentes de variância são estimados. Na inferência bayesiana os parâmetros têm uma distribuição de probabilidade enquanto na inferência frequentista (com fatores de efeitos fixos) os estimadores dos parâmetros é que têm uma distribuição de probabilidade.

Relação entre Blup e Estimadores Bayesianos

Além das distribuições (normais) adotadas para os efeitos aleatórios (g) no modelo linear clássico e para a verossimilhança do vetor de observações (y), a abordagem bayesiana requer atribuições para as distribuições a priori dos efeitos fixos e componentes de variância. A atribuição de distribuições a priori não informativas ou uniformes para os efeitos fixos e componentes de variância é uma forma de caracterizar um conhecimento a priori vago sobre os referidos efeitos e componentes (Gianola & Fernando, 1986; Silva et al., 2008; 2011).

Quanto à estimação dos efeitos fixos (efeitos de blocos completos, por exemplo) e de efeitos aleatórios (valores genéticos), tem-se que as médias das distribuições marginais a posteriori dos parâmetros de locação (efeitos fixos e aleatórios), dados os componentes de variância ou parâmetros de dispersão conhecidos, equivalem às soluções das equações do modelo misto do BLUP, desde que: sejam atribuídas prioris não informativas para os efeitos fixos, prioris normais para os efeitos aleatórios e verossimilhança normal para o vetor de observações.

Uma vez que a distribuição *a posteriori* resultante é simétrica e unimodal (normal), a moda, a mediana e a média são idênticas e uma grande classe de funções de perda comum (função de perda quadrática, função de perda absoluta ou função de perda uniforme) conduz ao mesmo estimador. Determinando a moda obtém-se o vetor médio da distribuição conjunta a posteriori, por maximização e não integração. Obtém-se então:

$$\begin{bmatrix} X'R^{-1}X+S^{-1} & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+A^{-1}\sigma_g^{-2} \end{bmatrix}\begin{bmatrix} E(b|y) \\ E(g|y) \end{bmatrix}=\begin{bmatrix} X'R^{-1}y+S^{-1}\ r_1 \\ Z'R^{-1}y+(A^{-1}\sigma_g^{-2})0 \end{bmatrix}$$, em que $r_1 = E(b)$ e $0 = E(g)$.

Essa derivação da metodologia BLUP, sob o enfoque bayesiano baseia-se na combinação de dois estimadores (fontes de informação) independentes. Neste caso, as equações resultantes são denominadas equações de modelo misto de Robertson (Resende e Rosa-Perez, 1999).

Tomando a distribuição a priori sobre os efeitos fixos como não informativa (expressa como $S \to \infty$ e então $S^{-1} \to 0$), tem-se que esta equação resultante equivale às equações do modelo misto do BLUP:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+A^{-1}\sigma_g^{-2} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix}=\begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Essa equação pode ser derivada também pela maximização de f (y, θ) para variações em θ (em que θ = b;g), sendo o estimador, neste caso, denominado máximo a posteriori (MAP). Sendo $P(g|y)$ = probabilidade de *g* dado *y*, o máximo *a posteriori* (MAP) de *g* é dado pela maximização de $P(g|y)$. Quando *g* e *y* seguem uma distribuição normal multivariada, o MAP de *g* é equivalente ao BLUP de *g*. A prova disso é apresentada a seguir.

Se Y ~N (μ, V), ou seja, P (Y) = $\frac{1}{(2\pi)^{n/2}|V|^{n/2}}\ e^{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)}$ em que n = ordem de *y*, tem-se usando o teorema de Bayes:

$$P(g|y) = \frac{P(y|g)\ P(g)}{P(y)}$$

$$\log P(g|y) = \log P(y|g) + \log P(g) - P(y)$$

$$= -\frac{n}{2}\log|R| - \frac{1}{2}(y - X\beta - Zg)'R^{-1}(y - X\beta - Zg) - \frac{1}{2}\log|G| - \frac{n}{2}g'A^{-1}g + \text{constante}$$

$$\frac{\partial}{\partial B}X'R^{-1}(y - X\hat{\beta} - Z\hat{g}) = 0 \qquad \Rightarrow X'R^{-1}X\hat{\beta} + X'R^{-1}Z\hat{g} = X'R^{-1}y$$

$$\frac{\partial}{\partial g}Z'R^{-1}(y - X\hat{\beta} - Z\hat{g}) - G^{-1}g = 0 \qquad \Rightarrow Z'R^{-1}X\hat{\beta} + (Z'R^{-1}Z + G^{-1})\ \hat{g} = Z'R^{-1}y$$

Esta última expressão é equivalente ao BLUP de *g*.

Relação entre Estimadores de Máxima Verossimilhança (ML) e Bayesianos

O objetivo do método ML é encontrar um conjunto de parâmetros que maximizam a verossimilhança de um modelo, dado uma coleção de observações. A verossimilhança para um determinado modelo pode ser escrito como uma função. Segundo os fundamentos de cálculo matemático, para encontrar o máximo dessa função, deve-se tomar a primeira derivada ou diferencial dessa função e igualar o resultado a zero. Isto propicia o conjunto de parâmetros que conduzem a função a um ponto crítico máximo, desde que não se tenha atingido um ponto de mínimo. Isto pode ser verificado usando o sinal da derivada segunda. Sinal positivo da derivada segunda indica concavidade para cima, ou seja, ponto de mínimo. Sinal negativo da derivada segunda indica concavidade para baixo, ou seja, ponto de máximo.

Considerando como uniforme a distribuição a priori dos parâmetros em b a serem estimados e maximizando (obtendo a moda) a distribuição a posteriori, o estimador resultante é equivalente ao de máxima verossimilhança – ML (Henderson, 1984; Gianola & Fernando, 1986). De fato, maximizando f (g,b) (mas considerando uma priori não informativa para b) com respeito a *g* e *b* obtém-se um estimador denominado de máxima verossimilhança, por Henderson et al. (1959), embora f(g,b) não seja uma função de verossimilhança e sim uma densidade a posteriori. Mesmo assim, pode ser obtido a partir das equações do modelo misto que $E(b\,|\,y) = (X'V^{-1}X)^{-1}\ X'V^{-1}y = \hat{b}$ e $E(g|y) = GZ'V^{-1}\ [y - Xb] = \hat{g}$ em que $\hat{b}$ é um estimador GLS e também ML de *b* e $\hat{g}$ é um estimador ML de E(g | b, y), equivalendo à média da distribuição condicional na qual *b* é fixado.

Implementação Prática da Análise Bayesiana

Os resultados de interesse gerados pela análise Bayesiana são, em geral, as distribuições marginais *a posteriori* dos parâmetros de interesse. Posteriormente, inferências baseadas na média, mediana, moda e desvios padrões destas distribuições são realizadas na prática.

O problema básico da implementação da análise Bayesiana refere-se à integração numérica. A integração (no espaço do parâmetro) da função densidade de probabilidade *a posteriori*, por exemplo:

$E\left[g\left(\theta\right)|y\right] = \int_{R_\theta} g\left(\theta\right) p(\theta|y)\, d\theta$, onde:

$g\left(\theta\right) = \theta$, para obtenção da média *a posteriori* e

$g\left(\theta\right) = (\theta - \mu)^2,\ \mu = E(\theta|y)$, para obtenção da variância *a posteriori* ou risco de Bayes, pode ser realizada através dos métodos (Gamerman, 1996): (*i*) analítico para aproximação de integral; (*ii*) automáticos ou de quadratura; (*iii*) simulação estocástica para obtenção de distribuições *a posteriori*, a qual é descrita em tópico seguinte.

1.7 Estimação de Componentes de Variância

Embora o problema central da avaliação genética seja a estimação de componentes de médias (valores genéticos), os quais são obtidos via integração (cálculo de esperança matemática) de funções, os componentes de variância são um problema tangencial à avaliação genética e são também essenciais em outras etapas do melhoramento genético. Os componentes de variância podem ser obtidos via integração ou derivação (maximização) de funções. Na Tabela 8 são apresentados os principais métodos de estimação de componentes de variância. Em cada linha da tabela o primeiro autor citado refere-se ao trabalho mais influente e os demais referem-se a trabalhos básicos e/ou teóricos que complementam o tema.

Tabela 8. Evolução dos métodos de estimação de componentes de médias (valores genéticos).

Método	Autores	Modelo	Estrutura de Variâncias	Distribuição das Variâncias
ANOVA	Henderson (1953); Fisher (1925)	Fixo, função para y	$e \sim N(0, I\sigma_e^2)$	-
ML	Hartley e Rao (1967); Fisher (1922)	Aleatório, função para y	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$	-
REML	Patterson e Thompson (1971); Thompson (1969; 1973)	Misto, função para (y-Xb)	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, A\sigma_g^2)$	-
BAYES- MCMC	Geman e Geman (1984); Gelfand e Smith (1990)	Aleatório, distribuição a posteriori	$e \sim N(0, I\sigma_e^2)$ $a \sim N(0, A\sigma_a^2)$	$\sigma_e^2 \sim \chi^{-2}(\nu_e, S_e^2)$ $\sigma_a^2 \sim \chi^{-2}(\nu_a, S_a^2)$ Uniforme se $\nu_i = -2;\ S_i^2 = 0$
G-REML ou REML/G-BLUP	Van Raden (2008); Misztal et al. (2010)	Misto, função para (y-Xb) com regressão em covariáveis (marcas) aleatórias ou G como matriz de parentesco genômico	$e \sim N(0, R = I\sigma_e^2)$ $g \sim N(0, G\sigma_g^2)$	-

A variação fenotípica é devida a efeitos genéticos e ambientais. Os efeitos genéticos podem ser decompostos em efeitos de um conjunto de genes de efeitos menores (poligenes) e efeitos atribuídos a genes maiores ou regiões genômicas específicas. A distinção entre esses três tipos de efeitos, bem como a decomposição da variação fenotípica total de um caráter em função desses três componentes, tem se tornado essencial aos programas de melhoramento genético de plantas e animais. Os efeitos ambientais podem ser desmembrados em independentes e correlacionados.

Os métodos padrões para estimação desses componentes de variância têm sido o da máxima verossimilhança residual (REML) e o da estimação Bayesiana (MCMC). Aplicados sobre dados fenotípicos combinados com informações de marcadores genéticos e de genealogia, esses métodos permitem a separação da variância genética associada a todo genoma daquela associada a regiões cromossômicas específicas, conduzindo à detecção de genes individuais. Quando aplicado usando apenas a informação de ligação gênica em análise dentro de família, geralmente não conduz a mapeamento suficientemente preciso para permitir resolução ao nível molecular. Entretanto, a inferência sobre o parentesco genético entre indivíduos usando as informações sobre desequilíbrio de ligação marcadores-QTL em toda a população, contribui para a melhoria da resolução.

Máxima Verossimilhança (ML)

O método da máxima verossimilhança baseia-se na obtenção do ponto de máximo de uma função de verossimilhança (que é a função densidade de probabilidade conjunta dos pontos amostrais). E este máximo é obtido por derivação da função de verossimilhança (L) em relação ao parâmetro de interesse. Assim, o estimador ML maximiza a verossimilhança do parâmetro dado a função densidade de probabilidade e o conjunto de dados. O ponto de máximo da função de verossimilhança é mais facilmente encontrado quando se toma o logaritmo natural dessa função. Isto porque, com essa transformação, o produtório em $L = \ell(\theta;y)$ transforma-se em somatório, fato que torna os cálculos mais tratáveis. No presente texto, as denominações Log e Log_e denotam a mesma coisa, ou seja, o logaritmo natural ou na base e.

O método ML foi desenvolvido por Fisher (1922), mas somente após cerca de 45 anos, Hartley e Rao (1967) apresentaram a especificação matricial de um modelo misto e a derivação de equações ML para várias classes de modelos. Os trabalhos de Henderson (1953) usando quadrados mínimos tiveram grande impacto no desenvolvimento dos métodos de estimação de componentes de variância a partir de dados desbalanceados, estimulando principalmente os trabalhos de Hartley e Rao. Embora viciado, o procedimento ML é computacionalmente mais simples que o método REML (descrito a seguir) e, em determinadas situações, apresenta eficiência satisfatória. O vício pode ser considerável se o número de equações independentes (posto de X, em que X é a matriz de incidência dos efeitos fixos), para os efeitos fixos, for relativamente grande em relação ao número (N) de observações. Quando o posto de X é pequeno em relação a N, os métodos ML e REML conduzem a resultados similares, conforme verificado por Resende et al. (1996) e Duarte e Vencovsky (2001).

Máxima Verossimilhança Restrita (REML)

O método REML foi desenvolvido e melhorado pelo pesquisador Robin Thompson e co-autores na Inglaterra. Tal método (Patterson & Thompson, 1971) surgiu a partir de esforços na obtenção de melhores estimadores de componentes de variância para dados não ortogonais e desbalanceados (Thompson, 1969). Posteriormente, foi estendido para modelos multivariados (Thompson,1973) e melhorado em termos do algoritmo de estimação via informação média (AI-REML) (Johnson & Thompson, 1995), visando a incorporação em softwares de excelência como o GENSTAT e o ASREML (Gilmour, Thompson e Cullis, 1995).

O método REML propicia uma correção ao ML, eliminando o seu vício. No método REML, somente a porção da verossimilhança que é invariante aos efeitos fixos (especificados no vetor β) é maximizada. Assim, o REML mantém as demais propriedades do ML, é não viciado e permite também a imposição de restrições de não negatividade. Dessa forma, o REML é o procedimento ideal de estimação de componentes de variância em modelos mistos. No método REML, os componentes de variância são estimados sem serem afetados pelos efeitos fixos do modelo e os graus de liberdade referentes à estimação dos efeitos fixos são considerados, produzindo estimativas não viciadas (Resende, 2007).

O método REML divide os dados em duas partes: contrastes dos efeitos fixos; e contrastes dos erros (isto é, todos os contrastes com esperança zero) os quais contém informações somente sobre os componentes de variância. Apenas os contrastes dos erros são então usados para estimar os componentes de variância, uma vez que eles contêm todas as informações disponíveis sobre os parâmetros de variância. Isto é feito pela projeção dos dados no espaço residual ou espaço vetorial dos contrastes dos erros. Os dados projetados têm Log L dado por $-2RL = [N - r(X)]\log 2\pi - \log|X'X| + \log|XV^{-1}X| + \log|V| + (y - X\hat{b})'V^{-1}(y - X\hat{b})$, em que N é o número de dados e r(X) é o posto da matriz de incidência dos efeitos fixos. Os componentes de variância são então estimados pela maximização do logaritmo da função RL dos dados projetados.

O Log L dos dados originais é dado por $-2L = N\log 2\pi + \log|V| + (y - Xb)'V^{-1}(y - Xb)$. A função RL tem termos adicionais em relação a L. O único termo adicional relevante para a estimação de componentes de variância é $\log|XV^{-1}X|$, o qual efetivamente remove os graus de liberdade usados na estimação dos efeitos fixos. Essa diferença entre RL e L reflete exatamente a diferença entre REML e ML (Resende, 2007). Quando o modelo inclui também outros efeitos fixos, além da média geral, o método REML deve ser usado em vez do ML.

Sob o enfoque frequentista o REML é derivado por meio da marginalização da verossimilhança através dos efeitos fixos. Pelo enfoque Bayesiano o REML é obtido por meio da integração através dos efeitos fixos e outros efeitos aleatórios.

1.8 Estimação Bayesiana de Componentes de Variância e relação com ML e REML

No contexto dos modelos lineares mistos, os valores genéticos (θ_1=g) são preditos simultaneamente à estimação dos efeitos fixos (θ_2=b) e dos componentes de variância (θ_3= σ_i^2). Na abordagem bayesiana, a avaliação genética pode ser obtida, de maneira geral, pela construção da densidade a posteriori f (θ_1, θ_2, θ_3|y) e, se necessário, pela integração de f (θ_1, θ_2, θ_3|y) em relação a θ_2 e θ_3. Estes (θ_2 e θ_3) são denominados parâmetros de *nuisance* e, por isso, devem ser integrados fora, exceto θ_2 em alguns casos, onde o mesmo constitui-se em uma parte integrante da função de mérito total (neste caso, a função de mérito depende da combinação linear de θ_1 e θ_2).

A obtenção de θ_1 requer a integração ou o conhecimento de θ_2 e θ_3. Henderson (1973) propôs o método BLUP para situações em que θ_3 é conhecido e θ_2 não o é. Para situações em que θ_3 não é conhecido, este autor sugeriu que o procedimento de máxima verossimilhança (ML) propiciaria estimativas razoáveis. Conforme Gianola & Fernando (1986), argumentos bayesianos, que não requerem normalidade e linearidade, permitem validar a intuição de Henderson.

A distribuição de θ_1, θ_2, e θ_3, dado y é proporcional a f (θ_1, θ_2, θ_3|y) α f (y|θ_1, θ_2, θ_3) . f (θ_1, θ_2, θ_3). Concentrando o interesse em θ_1 (o vetor de valores genéticos), deve-se integrar θ_2 e θ_3 por meio de $f(\theta_1 \mid y) = \int_{R_{\theta_2}} \int_{R_{\theta_3}} f(\theta_1 \mid \theta_2,\ \theta_3,\ y) \cdot f(\theta_2, \theta_3 \mid y) \cdot d\theta_2\, d\theta_3$. Tomando a distribuição conjunta a posteriori de forma que a maioria da densidade esteja na moda $(\hat{\theta}_2,\ \hat{\theta}_3)$, tem-se: $f(\theta_1 \mid y) \doteq f(\theta_1 \mid \theta_2 = \hat{\theta}_2,\ \theta_3 = \hat{\theta}_3,\ y)$.

Usando prioris não informativas para θ_2 e θ_3, tem-se que $\hat{\theta}_2$ e $\hat{\theta}_3$ são precisamente estimadores ML de θ_2 e θ_3, pois neste caso f(θ_2, θ_3|y) α f(y|θ_2, θ_3), ou seja a densidade de θ_2 e θ_3 dado y é proporcional à função de verossimilhança, de forma que a moda da posteriori conjunta corresponde ao máximo da função de verossimilhança, produzindo estimadores ML (Resende, 2000).

Uma abordagem alternativa para inferência sobre θ_1 consiste em obter $f(\theta_1,\ \theta_2 \mid y) \doteq f(\theta_1,\ \theta_2 \mid \theta_3 = \hat{\hat{\theta}}_3,\ y)$, onde $\hat{\hat{\theta}}_3$ refere-se à moda da densidade marginal de θ_3, dado y. Para obtenção de $\hat{\hat{\theta}}_3$ deve-se integrar θ_2 em $f(\theta_2,\ \theta_3 \mid y) \alpha f(y \mid \theta_2,\ \theta_3)$ e então maximizar f(θ_3|y). Usando-se uma priori não informativa para θ_3, sob normalidade $\hat{\hat{\theta}}_3$ é um estimador de máxima verossimilhança restrita (REML) para θ_3 (Harville, 1977). Assim, se o interesse reside na inferência conjunta para θ_1 e θ_2 basta usar $f(\theta_1,\ \theta_2 \mid y) \doteq f(\theta_1,\ \theta_2 \mid \theta_3 = \hat{\hat{\theta}}_3, y)$, que sob normalidade é equivalente à solução das equações de modelo misto com θ_3 substituído pelas estimativas REML de θ_3 (desde que se tenha usado prioris não informativas para θ_2 e θ_3) (Resende, 1999). Utilizando-se distribuições a priori não informativas para os efeitos fixos e componentes de variância, as modas das distribuições marginais a posteriori dos componentes de variância correspondem às estimativas obtidas por REML.

Inferências sobre componentes de variância devem ser baseadas em $f(\theta_3|y)$ α $f(y|\theta_3)$. $f(\theta_3)$, em que θ_3 contém variâncias e, portanto, $f(\theta_3|y)$ é definida na amplitude (o, ∞) para cada um dos elementos de θ_3, de forma que nunca surgem problemas de estimativas negativas de componentes de variância (Box & Tiao, 1973). $f(\theta_3|y)$ é obtida integrando-se θ_1 em $f(\theta_1, \theta_2, \theta_3|y)$ produzindo $f(\theta_2, \theta_3|y)$ e integrando-se θ_2 nesta última. Neste caso, $f(\theta_2, \theta_3|y)$ conduz aos estimadores ML de θ_2 e θ_3 e $f(\theta_3|y)$ conduz a um estimador REML de θ_3. Segundo Gianola & Fernando (1986), isto (eliminação das influências de θ_2 ou dos efeitos fixos) mostra precisamente porque REML deve ser preferido em relação a ML, ou seja, estes argumentos são mais fortes do que os apresentados por Patterson & Thompson (1971), que enfatizaram a propriedade de vício do ML.

Além da possibilidade do uso de informação a priori, eliminação de parâmetros de *nuisance* ou de distorção, a abordagem Bayesiana permite a integrada estimação – predição – decisão e a análise exata de amostras de tamanho finito (Resende, 1997). Assim, é uma maneira inteligente (*clever*) de fazer inferência. Outros procedimentos tradicionais de inferência são considerados ingênuos (*naive*) por alguns autores.

1.9 Estimação Bayesiana via MCMC

Dentre as classes de algoritmos para aproximar as integrais, a simulação estocástica baseada nos métodos de Monte Carlo é largamente indicada e utilizada para integração multivariada. Os métodos de Monte Carlo referem-se a processos de aproximação de valores esperados (integrais com respeito a uma distribuição de probabilidade) por meio de amostras, podendo ser referidos também como um caso especial de simulação de um processo estocástico.

Em genética quantitativa, para implementação prática da análise Bayesiana, uma das maiores dificuldades técnicas é a marginalização. A obtenção de distribuições marginais por processos analíticos é praticamente impossível (Sorensen e Gianola, 2002). Assim, a obtenção da distribuição marginal a posteriori (marginalização da distribuição conjunta a posteriori) tem sido obtida pelo método da amostragem de Gibbs (GS) através da amostragem e atualização das distribuições condicionais.

O método da amostragem de Gibbs pertence à classe de métodos, denominada Monte Carlo – Cadeias de Markov, a qual é sustentada em propriedades das Cadeias de Markov. O nome Gibbs advém da distribuição de Gibbs, que é muito utilizada na área de Física Estatística ou Mecânica Estatística. O amostrador de Gibbs explorando as distribuições condicionais completas através de algoritmo iterativo foi proposto inicialmente por Geman & Geman (1984) para aplicações na área de processamento de imagens. Entretanto, somente em 1990, este trabalho foi divulgado para toda a comunidade da área de estatística por Gelfand & Smith (1990) que publicaram em periódico da área de estatística, trabalho comparando o amostrador de Gibbs com outros processos de simulação estocástica.

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 da distribuição conjunta *a posteriori* para os efeitos e componentes de variância; (iv) obtenção das distribuições condicionais completas *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, portanto, métodos MCMC, como o amostrador de Gibbs, têm sido utilizados para obter amostras das distribuições marginais a posteriori por meio das distribuições condicionais completas *a posteriori* já citadas.

Geralmente são usadas distribuições *a priori* conjugadas pois, nesse caso, as distribuições *a posteriori* resultantes pertencem as mesmas famílias de distribuições das prioris. Assim, se a priori assume-se que os valores genéticos g apresentam distribuição normal, se terá na posteriori amostras de g também provenientes de uma distribuição normal.

Para ilustrar a aplicação da técnica da amostragem de Gibbs na avaliação genética será considerado o modelo individual univariado, conforme Resende e Rosa-Perez (1999) e Resende (2000).

Modelo

$y = Xb + Zg + e$, onde:
y : vetor de dados, de ordem n.
b : vetor de efeitos fixos, de ordem p.
g : vetor de valores genéticos aditivos, de ordem q.
e : vetor de erros, de ordem n.
X, Z : matrizes de incidência que associam b e g aos dados (y).

Na inferência bayesiana a formulação do modelo é denominada hierárquica ou em níveis. O primeiro nível refere-se à especificação da distribuição condicional dos dados em relação aos parâmetros, a denominada função de verossimilhança. O segundo nível da hierarquia refere-se à especificação das distribuições a priori dos parâmetros da distribuição condicional dos dados.

Definição da distribuição para a verossimilhança

Considera-se, inicialmente, que a distribuição condicional dos dados, dados b, g e σ_e^2 é normal multivariada: $y|\beta,\ g,\ \sigma_e^2 \sim N\ (Xb+Zg,\ I\ \sigma_e^2)$, onde I é a matriz identidade e σ_e^2 a variância residual. Essa igualdade advem do fato de que $e \sim N\ (0,\ I\ \sigma_e^2)$ e, fazendo-se $e = y - Xb - Zg$, esse novo residuo tem distribuição $\sigma_e^2 \sim N\ (Xb+Zg,\ I\ \sigma_e^2)$, decorrente da mudança na média de 0 para $(Xb+Zg)$.

Distribuições a priori

Considerando o modelo quantitativo infinitesimal, tem-se que a distribuição de g é também normal multivariada: $g|A,\ \sigma_g^2 \sim N(O,\ A\,\sigma_g^2)$, onde A é a matriz de parentesco genético aditivo e σ_g^2 é a variância genética aditiva na população base.

Os parâmetros de interesse para inferências são: $b,\ g,\ \sigma_g^2\ e\ \sigma_e^2$. Para conduzir a análise Bayesiana torna-se necessário especificar as distribuições a priori para $b,\ \sigma_g^2\ e\ \sigma_e^2$ (a distribuição de g já foi especificada).

Como priori para b pode-se assumir p(b) α constante, que especifica aproximadamente a noção de conhecimento a priori vago para b. Esta distribuição a priori é imprópria, mas pode-se tornar própria, desde que se especifique os limites superior e inferior para p(b).

As distribuições a priori dos componentes de variância ($\sigma_e^2\ e\ \sigma_g^2$) poderiam ser uniforme da forma $p\ (\sigma_i^2)\ \alpha$ constante, $0 \le \sigma_i^2 < \sigma^2_{i\ max}\ (i = e,\ g)$, onde, de acordo com o conhecimento acumulado sobre o caráter, $\sigma^2_{i\ max}$ seria o valor máximo que σ_i^2 poderia assumir, a priori. Alternativamente, poderia ser especificada uma priori mais informativa para os componentes de variância, considerando uma distribuição qui-quadrado escalada invertida, da forma: $p\ (\sigma_i^2|\upsilon_i,\ S_i^2)\ \alpha\ (\sigma_i^2)^{-((\upsilon_i/2)+1)} \exp\left[\frac{-\upsilon_i\ S_i^2}{2\ \sigma_i^2}\right]$ $(i = e,\ g)$, onde υ são os graus de liberdade da distribuição qui-quadrado e S_i^2, o valor inicial da variância. Esta distribuição reduz-se a uma distribuição uniforme imprópria se $\upsilon_i = -2$ e $S_i^2 = 0$.

Uma distribuição a priori f(θ) é imprópria quando a integral sobre todos os possíveis valores de θ não converge: $\int f(\theta)\ d\,\theta \to \infty$. Entretanto, o interesse principal reside na distribuição a posteriori e como esta é, em geral, própria mesmo quando a priori não o é, a eventual impropriedade das distribuições a priori não é importante.

Distribuição conjunta a posteriori

Definidas estas distribuições, pode-se agora escrever a distribuição conjunta a posteriori dos parâmetros do modelo.

$p\ (b,\ g,\ \sigma_g^2,\ \sigma_e^2|y)\ \alpha\ p\ (b,\ g,\ \sigma_g^2,\ \sigma_e^2)\ p\ (y|b,\ g,\ \sigma_g^2,\ \sigma_e^2)$

$= p\ (b)\ p\ (g|\sigma_g^2)\ p\ (\sigma_g^2)\ p\ (\sigma_e^2)\ p\ (y|b,\ g,\ \sigma_g^2,\ \sigma_e^2)$, em que se omitiu o condicionamento nos hiperparâmetros (parâmetros que auxiliam na especificação da priori) e na conhecida matriz de parentesco A.

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

$$p\ (b,\ g,\ \sigma_g^2,\ \sigma_e^2|y)\ \alpha\ \sigma_e^{2-\left(\frac{n+\upsilon_e}{2}+1\right)}\ \exp\left[-\frac{(y-Xb-Zg)'(y-Xb-Zg)+\upsilon_e\ S_e^2}{2\ \sigma_e^2}\right]$$

$$\sigma_g^{2-\left(\frac{q+\upsilon_a}{2}+1\right)}\ \exp\left[-\frac{(g'A^{-1}g+\upsilon_g\ S_g^2}{2\ \sigma_g^2}\right]$$

Desejando atribuir distribuição a priori uniforme para $\sigma_g^2\ e\ \sigma_e^2$, basta fazer $\upsilon_i=-2\ e\ S_i^2=0\ (i=g,\ e)$ na expressão acima.

Distribuições condicionais a posteriori

Para implementação do GS, deve-se derivar todas as distribuições condicionais a posteriori a partir da distribuição conjunta a posteriori apresentada acima.

Denominando-se $Xb+Zg=W\theta,\ onde\ W=[X\ Z]\ e\ \theta'=[b'\ g']$, tem-se que a matriz dos coeficientes das equações de modelo misto é dada por C = W' W + Σ, onde $\sum=\begin{bmatrix}0 & 0\\ 0 & A^{-1}\sigma_e^2/\sigma_g^2\end{bmatrix}$. A distribuição condicional a posteriori de θ é:

$\theta\ |\ \sigma_g^2,\ \sigma_e^2,\ y\sim N\ (\hat{\theta},\ C^{-1}\sigma_e^2),\ em\ que\ \hat{\theta}\ é\ dado\ por\ C\hat{\theta}=W'y$, ou seja, pelas equações de modelo misto.

Como exemplo, a derivação da distribuição condicional a posteriori para b_i (o i-ésimo elemento do vetor b) conduz a

$$b_i|b_{-i},\ g,\ \sigma_g^2,\ \sigma_e^2,\ y\sim N\ (\hat{b}_i,(X'_i\ X_i)^{-1}\sigma_e^2),$$
$$em\ que:\ \hat{b}_i=(X'_i\ X_i)^{-1}X'_i\ (y-X_{-i}b_{-i}-Zg)$$

X_{-i} e b_{-i} referem-se a X e b excluindo-se o elemento i.

E a distribuição condicional a posteriori de g_i é:

$g_i|b,\ g_{-i},\ \sigma_g^2,\ \sigma_e^2,\ y\sim N\ (\hat{g}_i,(z'_i\ z_i+A_{i,i}^{-1}\alpha)^{-1}\sigma_e^2)$ e pode ser escrita também como $g_i|b,\ g_{-i},\ \sigma_g^2,\ \sigma_e^2,\ y\sim N\ (\hat{g}_i,PEV_i)$.

Marginalização das distribuições condicionais por amostragem dos parâmetros de locação

Consiste em amostrar das condicionais a posteriori acima, para cada elemento de b e g.

Marginalização das distribuições condicionais por amostragem dos parâmetros de dispersão

Tendo amostrado todos os parâmetros de locação do modelo, deve-se computar:

$$SS_e^{(1)}=(y-Xb^{(1)}-Zg^{(1)})'(y-Xb^{(1)}-Zg^{(1)})$$
$$SS_g^{(1)}=(g^{(1)})'\ A^{-1}\ g^{(1)}$$

A primeira iteração do amostrador é completada, retirando-se os componentes de variância, usando $SS_g^{(1)}$ e $SS_e^{(1)}$:

$\sigma_g^2|b,\ g,\ \sigma_e^2, y \sim SS_g^{(1)}\ \chi_{q-2}^{-2}$

$\sigma_e^2|b,\ g,\ \sigma_g^2, y \sim SS_e^{(1)}\ \chi_{n-2}^{-2}$

A segunda iteração inicia-se através de atualizações das equações de modelo misto com $\alpha = \sigma_e^2 / \sigma_g^2$, onde σ_e^2 e σ_g^2 são os valores amostrados acima.

As bases para essas expressões vêm da distribuição qui-quadrado dada por uma razão entre variâncias:

$\chi_{v_0}^2 = \frac{SS}{\sigma^2} = \frac{S_0^2 v_0}{\sigma^2}$, em que v_0 é um hiperparâmetro referente ao grau de confiança no componente de variância a priori S_0^2. Dessa expressão tem-se que $SS = S_0^2 v_0$ e que $\sigma^2 = \frac{S_0^2 v_0}{\chi_{v_0}^2}$, que é a distribuição (qui-quadrado invertida escalada) ou densidade a priori para o componente de variância σ^2. Assim, $\sigma^2 \sim SS\ \chi_{v_0}^{-2}$, conforme usado acima e derivado de $\sigma_e^2|b,\ g,\ \sigma_g^2,\ y \sim \tilde{\upsilon}_e\ \tilde{S}_e^2\ \chi_{\tilde{\upsilon}_e}^{-2}$ e $\sigma_g^2|b,\ g,\ \sigma_e^2,\ y \sim \tilde{\upsilon}_g\ \tilde{S}_g^2\ \chi_{\tilde{\upsilon}_g}^{2}$. Associado a uma variável qui-quadrado invertida escalada tem-se as seguintes média e variância: $E(\sigma^2|S_0^2 v_0) = \frac{S_0^2 v_0}{v_0 - 2}$ e $E(\sigma^2|S_0^2 v_0) = \frac{2(S_0^2 v_0)^2}{(v_0-2)^2(v_0-4)}$.

Algoritmo GS

Em termos mais simples, o algoritmo GS pode ser apresentado de forma resumida:

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 $\bar{y}$ como único efeito fixo, pode-se calcular $\bar{y}$ como a média aritmética das observações e $g_i = h^2(y_i - \bar{y})$. Devem ser fornecidos os valores iniciais para $\bar{y}_i,\ g_i,\ \sigma_e^2,\ \sigma_g^2$ e $\alpha = \sigma_e^2 / \sigma_g^2$.
2. Gerar valores para os efeitos fixos. Sendo o único efeito fixo, a média geral, tem-se: $\hat{\bar{y}} = \bar{y} + rnd\ \sigma_e / (n)^{1/2}$
3. Gerar valores para os efeitos aleatórios: $\hat{g} = g_i + rnd\ [(1 - r_{\hat{g}g}^2)\ \sigma_g^2]^{1/2}$, onde $r_{\hat{g}g}$ é a acurácia dada por $r_{\hat{g}g}^2 = (1 - PEV_i / \sigma_g^2)^{1/2}$, onde PEV_i é o iésimo elemento da inversa da matriz dos coeficientes das EMM multiplicado por σ_e^2.
4. Calcular a soma de quadrados do resíduo (SSE) e a variância residual σ_e^2.

Considerando que a distribuição a priori para a variância residual é a inversa de uma qui-quadrado, tem-se:

$SSE = \sum(y_i - \hat{\bar{y}} - \hat{g}_i)^2$

$\sigma_e^2 = \frac{SSE}{X_n^2}$

5. Gerar um valor para a variância dos efeitos aleatórios de valores genéticos.

$$\sigma_g^2 = \frac{\hat{g}' A^{-1} \hat{g}}{X_q^2}$$

6. Calcular o novo valor do parâmetro

$$\hat{\alpha} = \frac{\sigma_e^2}{\sigma_g^2}$$

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

Diagnóstico de Convergência

Para a inferência bayesiana sobre os parâmetros de interesse pode empregar-se a técnica da amostragem de Gibbs. O principal aspecto deste procedimento refere-se ao fato de as inferências basearem-se na distribuição marginal a posteriori dos parâmetros, sendo que a marginalização da distribuição conjunta a posteriori é obtida via o amostrador de Gibbs através de amostragens e atualizações das distribuições condicionais. A abordagem bayesiana baseia-se então na construção da distribuição marginal a posteriori de um parâmetro de interesse tratando-o como uma variável aleatória e aplicando cálculo de probabilidades. Este procedimento implica problemas multidimensionais, uma vez que todos os outros parâmetros do modelo devem ser integrados (eliminados), fato que raramente é possível usando os métodos numéricos padrões.

O procedimento iterativo da amostragem de Gibbs refere-se a uma técnica de integração estocástica que cria uma cadeia de Markov, que é uma distribuição (conjunta a posteriori) estacionária associada à distribuição a posteriori de interesse. Tomando-se amostras, iterativamente, das distribuições condicionais a posteriori, com contínua atualização, obtém-se a distribuição conjunta a posteriori em equilíbrio e, após um número de iterações suficientemente grande, a última amostra desta seqüência e qualquer amostra subseqüente é uma amostra da distribuição marginal requerida. Este resultado implica que cada coordenada do vetor de amostras retiradas, $\theta^n = [b^n \; g^n \; \sigma_g^{2(n)} \; \sigma_e^{2(n)}]$, é uma amostra da distribuição marginal a posteriori apropriada. Em resumo, antes do equilíbrio amostra-se da distribuição condicional completa e após o equilíbrio amostra-se da distribuição marginal $f(\theta_i|y)$.

As cadeias de Markov estão inseridas no contexto da teoria dos processos estocásticos, teoria esta, definida como a parte dinâmica da teoria de probabilidades, onde se estuda uma coleção de variáveis aleatórias, com respeito a sua interdependência e comportamento limite. Para a inferência bayesiana, é de maior relevância o estudo do comportamento assintótico da cadeia, quando o número de iterações tende a ∞, uma vez que a inferência deve ser baseada na distribuição (a posteriori) estacionária, ou seja, em equilíbrio.

À medida em que o número de iterações aumenta, a cadeia se aproxima da condição de equilíbrio. Dessa forma, é necessário considerar a convergência em uma determinada iteração cuja distribuição esteja próxima da distribuição em equilíbrio (atingido teoricamente quando $n \rightarrow \infty$), ou seja, após um número suficientemente grande de iterações.

A forma básica de obter uma amostra de tamanho m da posteriori é produzir m cadeias independentes (geradas a partir de m valores iniciais diferentes) e, após a convergência, retirar os valores da última iteração de cada cadeia. Outra opção consiste em retirar m amostras da mesma cadeia após a convergência, visto que se estará amostrando da distribuição a posteriori em equilíbrio. Neste último caso, é importante relatar que as amostras sucessivas não são independentes, de forma que se torna necessário descartar várias iterações entre cada duas amostras a serem salvas. Como o processo é markoviano, a dependência diminui com o aumento da distância entre iterações, obtendo-se, assim, independência entre as amostras salvas.

Considerando a segunda opção, no contexto do diagnóstico da convergência, tornam-se relevantes as quantidades: M = número de iterações pré-convergência ou período de descarte ou período de aquecimento da cadeia (*burn in*); N = número de iterações após a convergência; K = número de iterações entre amostras sucessivas ou intervalo entre amostras (*thin*). O tamanho total da cadeia é dado por T = M + N.

O valor de K pode ser determinado calculando as autocorrelações na série de valores gerados e verificando a partir de qual ponto pode-se considerar as autocorrelações como nulas. Uma vez que o valor de K é muito menor que o de M, os métodos baseados em uma única cadeia (mais longa) são preferidos computacionalmente.

Outra forma de análise de convergência refere-se à estimação do erro de Monte Carlo, que é uma estatística associada ao erro de estimação de determinado parâmetro devido ao número de amostras utilizadas na cadeia de Gibbs, sendo que este erro é inversamente proporcional ao tamanho da cadeia. Este erro pode ser calculado pela variância dos parâmetros amostrados sucessivamente a cada intervalo dividida pelo número de amostras salvas, sendo que a raiz quadrada deste erro fornece uma aproximação para o desvio padrão do erro associado ao comprimento da cadeia.

Devido ao fato de que valores aleatórios são utilizados inicialmente como realização do conjunto de parâmetros, é necessário um período de descarte de amostras até que as amostras de GS possam ser consideradas como provenientes da distribuição conjunta a posteriori, ou seja, da distribuição em equilíbrio estacionário. Em geral, tem sido utilizado o esquema tradicional de cadeia longa (única) de Gibbs, onde o processo de reamostragem é contínuo. Assim, de maneira geral, um grande (da ordem de 10.000 a 1.000.000) número de ciclos tem sido utilizado, sendo descartadas as primeiras amostras (da ordem de poucos milhares) e amostras de cada parâmetro são salvas a cada pequeno (da ordem de 50 a 100) número de iterações. O intervalo entre amostras salvas é necessário como forma de obtenção de amostras independentes, visto que amostras sucessivas apresentam correlação serial. O número total de amostras salvas é utilizado para cômputo das estimativas pontuais e intervalares de interesse.

1.10 Métodos numéricos e softwares para REML/BLUP e MCMC

A implementação computacional da metodologia de modelos mistos baseia-se fortemente em métodos numéricos, notadamente, em álgebra linear numérica, visando à obtenção iterativa das soluções das equações de modelo misto (obtenção do BLUP) e, cálculo numérico para a maximização/ minimização de funções de várias variáveis, visando à obtenção das estimativas REML.

Vários algoritmos computacionais para a obtenção de componentes de variância por ML e REML têm sido desenvolvidos tais como o MS (Method of Scoring de Fisher), o EM (Expectation-Maximization, de Dempster et al., 1977), o DF-REML (Derivative-free Restricted Maximum Likelihood, de Graser et al., 1987 e o AI-REML (Average Information-REML de Johnson & Thompson, 1995). Dentre estes, os mais usados são o EM e o AI-REML. O algoritmo EM é muito estável, numericamente, apresentando convergência mesmo que os valores iniciais não tenham sido totalmente adequados. Entretanto, uma inconveniência do algoritmo EM é a lentidão para as estimativas próximas ao limite do espaço paramétrico (por exemplo, quando uma variância tende a zero). Se valores iniciais positivos forem utilizados, a convergência para valores não negativos é garantida.

O algoritmo EM atua por meio da obtenção da esperança (por integração) e maximização (derivação) da função de verossimilhança dos dados, sucessivamente. Nos modelos ao nível de indivíduos, em que, freqüentemente, a ordem das equações de modelo misto excedem o número de observações, a obtenção de estimativas por meio de primeira derivada pelo método EM requer a inversão da matriz dos coeficientes das equações de modelo misto, aumentando muito o esforço computacional. Os métodos de Newton-Raphson e de Fisher apresentam convergência quadrática, ao passo que o algoritmo EM apresenta convergência linear, sendo, portanto, mais lento.

Os algoritmos para obtenção de estimativas REML podem ser agrupados de acordo com a ordem das derivadas usadas. Assim, têm-se: (i) não derivativo (DF-REML); (ii) baseado em derivadas parciais de primeira ordem (EM-REML); (iii) baseado em derivadas parciais de primeira e segunda ordens (AI-REML). O algoritmo AI é um procedimento derivativo melhorado, o qual fundamenta-se no uso dos métodos de Newton, que usam as derivadas primeira e segunda da função de verossimilhança. Tal algoritmo baseia-se na utilização da informação advinda da média das derivadas segundas observadas e esperadas da função de verossimilhança, de forma que o termo que contém os traços dos produtos da matriz inversa é cancelado, restando uma expressão mais simples para computação. Técnicas de matrizes esparsas são empregadas no cálculo dos elementos da inversa da matriz dos coeficientes, os quais são necessários para as derivadas primeiras da função de verossimilhança. Este algoritmo é também denominado Quasi-Newton (Gilmour et al., 1995), o qual aproxima a matriz hessiano (matriz de derivadas segundas) pela média das informações observadas e esperadas. A informação observada é uma medida da curvatura da função (ou do seu log) de verossimilhança e a informação esperada é a própria informação de Fisher.

Johnson e Thompson (1995) e Gilmour, Thompson e Cullis (1995) apresentaram o algoritmo de Informação Média (AI), o qual baseia-se no uso de uma matriz de informação alternativa. Visto que as matrizes de IO e IE são difícil computação (pois envolvem a segunda derivada), tais autores propuseram o uso da matriz de informação média, a qual contempla uma média das matrizes IO e IE. O cálculo da matriz AI é muito mais simples do que o cálculo de qualquer uma das duas (IO e IE) isoladamente. Isto porque, quando é feita a média das derivadas segunda observadas e esperadas, o termo envolvendo traços de produtos da matriz inversa, são cancelados, permanecendo uma expressão de simples computação.

O método esperança - maximização com parâmetros estendidos (PX-EM) é mais recente (Foulley e Van Dyk, 2000) e também o mais eficiente juntamente com o AI. Esse método baseia-se na normalização dos efeitos aleatórios e aumenta muito a velocidade de convergência quando comparado ao EM tradicional. Atualmente é utilizado na implementação dos *softwares* Wombat (antigo Dfreml), ASREML e Selegen-REML/BLUP. No ASREML e Wombat é usado em associação com o AI.

Os métodos baseados em cadeias de Markov/método Monte Carlo (MCMC), muito usados em inferência bayesiana, podem também ser usados no contexto da inferência verossimilhança. O método estatístico REML e os métodos numéricos (NR, FS, EM, DF, AI e PX-EM) até aqui apresentados são denominados métodos exatos. Esses métodos são exatos no sentido de que não são baseados em amostragens de distribuições de probabilidade. Os métodos estatísticos bayesianos baseiam-se em amostragem e, nesse sentido, não são denominados métodos exatos. Os métodos numéricos empregados na abordagem bayesiana como a amostragem de Gibbs pertencem a uma classe de métodos denominada cadeias de Markov e Monte Carlo (MCMC). No entanto, para usar os métodos MCMC, não há necessidade de se empregar os fundamentos bayesianos. O fundamento dos métodos MCMC é de que, devido às dificuldades para se calcular as PEV associadas aos efeitos dos fatores aleatórios, essas são substituídas por amostragens. Assim, podem ser usados também associados ao algoritmo EM. Segundo Thompson (2002), nem sempre é claro qual abordagem computacional é mais eficiente: exata, amostragem de Gibbs bayesiana ou algo intermediário. A dependência da PEV na estimação de componentes de variância é ilustrada a seguir.

Henderson (1986) apresentou equações para a estimação de componentes de variância por EM. Essas equações envolvem a computação de formas quadráticas para os fatores aleatórios e sub-equações para as variâncias dos erros de predição (PEV) dos efeitos de todos os fatores aleatórios. Tomando os traços dos produtos das formas quadráticas pelas PEV obtém-se p+1 equações para p parâmetros ou componentes de variância. Somando-se as duas equações referentes ao fator aleatório dos efeitos genéticos aditivos, o sistema de equações pode ser resolvido para os p componentes de variância associados aos p fatores aleatórios. A seguir maiores detalhes são apresentados.

Para a estimação dos componentes de variância são necessárias duas formas quadráticas para o vetor de erros preditos ($\hat{e}$) e uma forma quadrática para o vetor de valores genéticos preditos ($\hat{g}$). A forma quadrática para $\hat{g}$ é dada por $\hat{g}'Q\hat{g}$, sendo a matriz associada igual a $Q \equiv A^{-1}\sigma_g^{-4}$. As duas formas quadráticas para $\hat{e}$ são dadas por $\hat{e}'P_g\hat{e}$ e $\hat{e}'P_e\hat{e}$. As duas matrizes associadas são iguais a $P_g = R^{-1}DR^{-1}$ e $P_e = R^{-1}R_eR^{-1}$. A matriz R pode ser rescrita como $R = D\sigma_g^2 + R_e\sigma_e^2$ em que $R_e = I$.

Essas formas quadráticas devem ser igualadas às suas esperanças matemáticas para que se obtenha equações resultantes para $\hat{\sigma}_g^2$ e $\hat{\sigma}_e^2$. Para encontrar essas esperanças deve-se observar que:

$E(\hat{g}'Q\hat{g}) = tr[Q \text{ var}(\hat{g})]$,
$E(\hat{e}'P_g\hat{e}) = tr[P_g \text{ var}(\hat{e})]$ e
$E(\hat{e}'P_e\hat{e}) = tr[P_e \text{ var}(\hat{e})]$.

Verifica-se assim que, para encontrar os valores esperados necessitam-se das PEV dos efeitos aleatórios, aqui denominadas Var($\hat{g}$) e Var($\hat{e}$) e essas são funções lineares de $\hat{\sigma}_g^2$ e $\hat{\sigma}_e^2$.

Segundo Schaeffer (1999), a amostragem de Gibbs é muito similar ao método iterativo de Gauss-Seidel, exceto que quando cada solução para os efeitos são obtidas, adiciona-se uma quantidade aleatória baseada na distribuição condicional *a posteriori* de sua variância. Para usar a amostragem de Gibbs, há necessidade apenas de um programa de resolução das equações de modelo misto, um bom gerador de números aleatórios e tempo computacional para processar um imenso número de amostras. Thompson (2002) relata um procedimento de aumento de dados para reduzir o esforço computacional na estimação de componentes de variância, porém sem adicionar tanto *noise* em *a*. O procedimento envolve o ajuste de dois modelos $y - Z\tilde{g} = Xb + e$ e $y - X\tilde{b} = Zg + e$. No primeiro modelo ajusta-se $\hat{b}$ e se obtém $\tilde{b} = \hat{b} + amostragem$. No segundo modelo, ajusta-se y para $\tilde{b}$, estima-se σ_g^2 e σ_e^2, ajusta-se $\hat{g}$ e obtém-se $\tilde{g} = \hat{g} + amostragem$. Então ajusta-se y para $Z\tilde{g}$ e o procedimento é repetido. Após um período de aquecimento, as médias $\overline{\sigma}_g^2$ e $\overline{\sigma}_e^2$ fornecem estimativas para σ_g^2 e σ_e^2, assim como no procedimento de amostragem de Gibbs. Isto evita adicionar tanto *noise* em $\tilde{g}$ quando σ_g^2 e σ_e^2 são estimados. A amostragem de Gibbs é uma forma de tornar o REML computacionalmente possível para grande conjuntos de dados e modelos complexos.

O método numérico de Gauss-Seidel para a resolução iterativa de sistemas de equações lineares é descrito a seguir empregando um pequeno exemplo.

Seja o sistema de equações lineares:

$$\begin{cases} 4X_1 + X_2 + X_3 = 5 \\ -2X_1 + 5X_2 + X_3 = 0 \\ 3X_1 + X_2 + 6X_3 = -6,5 \end{cases}$$

As soluções para as três incógnitas X_1, X_2 e X_3 são dadas por:

$$X_1^k = \frac{(5 - X_2^{k-1} - X_3^{k-1})}{4}; \quad X_2^k = \frac{(0 + 2X_1^k - X_3^{k-1})}{5}; \quad X_3^k = \frac{(-6,5 - 3X_1^k - X_2^k)}{6},$$

em que k, refere-se à k-ésima iteração.

Partindo-se de um vetor inicial X^o = (0, 0, 0), tem-se a 1ª iteração:

$$X_1^1 = \frac{(5-0-0)}{4} = \frac{5}{4}; \quad X_2^1 = \frac{(0+2\cdot 5/4-0)}{5} = \frac{1}{2}; \quad X_3^1 = \frac{(-6,5-3\cdot 5/4-1/2)}{6} = -1,7967.$$

Na 2ª iteração, tem-se:

$$X_1^2 = \frac{(5-1/2+1,7967)}{4} = 1,58; \; X_2^2 = \frac{(0+2\cdot 1,58-(-1,7967))}{5} = 0,992; \; X_3^2 = \frac{(-6,5-3\cdot 1,58-0,992)}{6} = 2,03$$

O procedimento prossegue até que o menor valor de $\left| X^k - X^{k-1} \right| \leq \varepsilon$, em que ε é o erro desejado (geralmente $\leq 10^{-5}$).

O algoritmo esperança – maximização com aproximação estocástica (SAEM) foi apresentado por Jaffrezic et al. (2007) como uma forma eficiente de computação e inferência em modelos não lineares mistos. Nessa situação complexa, geralmente são usados procedimentos aproximados de máxima verossimilhança e também métodos bayesianos. O método SAEM surge como uma opção de rápida convergência em relação aos algoritmos EM Monte Carlo e bayesiano. Outra vantagem é que o mesmo não requer a especificação de distribuições *a priori* e é bastante robusto à escolha dos valores iniciais no processo iterativo. A idéia é reciclar os valores simulados de uma iteração, na próxima iteração do algoritmo EM, fato que acelera consideravelmente a convergência.

A escolha dos algoritmos matriciais quanto a esparsidade das matrizes depende da situação, e os principais métodos para cálculo da inversa de matrizes esparsas foram descritos por Takahashi et al. (1973), Zollenkof (1971). Esses métodos calculam somente os elementos da inversa que pertencem ao padrão de esparsidade da matriz original. Mesmo assim, o custo computacional para o cálculo da inversa esparsa é de duas a três vezes maior do que para cálculo de determinantes. O cálculo de uma inversa esparsa aumenta os requerimentos computacionais para avaliação de verossimilhanças. Thompson et al. (1994) apresentaram métodos para encontrar os elementos da matriz esparsa, os quais reduzem esses requerimentos. Um resumo dos Métodos Numéricos para REML é apresentado na Tabela 9.

Tabela 9. Métodos numéricos para REML.

Método Numérico para REML	Autores	Ordem da Derivação
Newton-Raphson (NR)	Newton	Derivadas parciais de primeira e segunda ordens
Escores de Fisher (FS)	Fisher	Derivadas parciais de primeira ordem
Esperança - Maximização (EM)	Dempster et al. (1977)	Derivadas parciais de primeira ordem
Livre de Derivadas (DF)	Graser et al. (1987)	Não derivativo
Informação Média (AI)	Gilmour, Cullis e Thompson (1995)	Derivadas parciais de primeira e segunda ordens
Esperança - Maximização com Parâmetros Estendidos (PX-EM)	Foulley e Van Dyk (2000)	Derivadas parciais de primeira ordem
Cadeias de Markov e Monte Carlo (MCMC)	Gelfand e Smith (1990)	-
Esperança - Maximização Estocástico (SAEM)	Jaffrezic et al. (2007)	-

Os softwares para REML/BLUP fenotípico e genômico mais utilizados no Brasil são apresentados na Tabela 10. Códigos para ajustes de alguns modelos lineares generalizados mistos para variáveis normais e binomiais pelo ASREML são apresentados por Resende (2000).

Tabela 10. Softwares para REML/BLUP e MCMC.

Software	Autores	Método Numérico para REML	Inversão Esparsa	Procedimentos
ASREML e GENSTAT	Gilmour, Cullis e Thompson (1995)	Informação Média (AI)	AS	REML e BLUP
DFREML/WOMBAT	Meyer (1991)	Esperança – Maximização (EM) e AI	-	REML e BLUP
REMLF90 e BLUPF90	Misztal (1995)	EM Acelerado	Takahashi	REML e BLUP Blup Genômico
SELEGEN-REML/BLUP	Resende (1994)	EM Acelerado	Zollenkopf	REML e BLUP
SAS	Littell et al. (1996)	-	-	REML e BLUP
SELEGEN GENÔMICA	Resende (2007)	-	Zollenkopf	Blup Genômico
QxPack	Perez-Enciso e Misztal (2004)	-	-	Análise de QTL REML e BLUP Blup Genômico
ASREML				Blup Genômico
GS3	Legarra et al. (2011)	-	-	Blup Genômico IBLASSO Bayes Cpi MCMC
GENOME WIDE PREDICTION	Meuwissen (2009)	-	-	Blup Genômico BayesA BayesB
BLR	Perez et al. (2010)	-	-	Blasso
rr-BLUP	Endelman (2011)	-	-	REML e BLUP Blup Genômico

1.11 Testes de Hipóteses e Parcimônia de Modelos

Os testes de hipóteses referentes aos efeitos fixos e aleatórios no contexto dos modelos mistos bem como os critérios para a comparação de modelos são apresentados na Tabela 11.

Tabela 11. Testes de hipóteses referentes aos efeitos fixos e aleatórios e critérios para a comparação de modelos.

Testes de Hipóteses	Efeitos	Distribuição Assintótica	Calculo
t	Fixos	t	$t = (\overline{Y}_i - \overline{Y}_j)/(\hat{\sigma}/n^{1/2})$
F	Fixos e Aleatórios	F	F =[Var(trat) + Var(Residual)] / Var(Residual)
LRT	Aleatórios	Qui-quadrado	LRT = $(-2\text{Log L})_{p+1}$ - $(-2\text{Log L})_p$
$t^2 = F$	Fixos	F	F = [Var(trat) + Var(Residual)] / Var(Residual)
WALD n pequeno = F	Fixos	F	$W = \theta^2 / \text{Var}(\theta)$
WALD n grande = LRT	Aleatórios	Qui-quadrado	$W = \theta^2 / \text{Var}(\theta)$
AIC	Aleatórios	-	AIC = -2 log L + 2 p
BIC	Aleatórios	-	BIC = -2 log L + p log v v = N – r(x)
AICc	Aleatórios		AIC = -2 log L + 2 p + [2p(p+1)/(n-p-1)]

A significância da diferença no ajuste de diferentes modelos aos dados pode ser testada usando o Teste da Razão de Verossimilhança de Wilks (LRT), definido por: $\lambda = 2[Log_e L_{p+1} - Log_e L_p]$. Assim, basta comparar λ [2 vezes a diferença (modelo com maior número de parâmetros – modelo com menor número p de parâmetros) de Log_e L associados a dois modelos ajustados] com o valor da função densidade de probabilidade (Tabela de χ^2) para determinado número de graus de liberdade e probabilidade de erro. O número de graus de liberdade é definido pela diferença no número de parâmetros ou componentes de variância entre modelos.

Tal teste envolve duas vezes a redução no Log L resultante da retirada de t termos aleatórios, quantidade esta distribuída como uma χ^2_t. Assim, para a verificação da significância de um efeito aleatório, tem-se que LRT ~ χ^2_1. Entretanto, Stram e Lee (1994) sugerem uma correção por meio da multiplicação do P valor associado a χ^2_1 por 0,5, ou seja, sugerem o uso de uma distribuição $\chi^2_{0,5}$. Esta correção é, sobretudo, indicada para teste no limite do espaço paramétrico, quando o P valor aproximado para a estatística de teste d (duas vezes a redução no Log L) é $0{,}5(1 - P(\chi^2_1 \leq d))$, em que P denota probabilidade. Nesse caso (mistura de distribuições com 1 e 0 graus de liberdade), o valor tabelado de qui-quadrado para o nível de significância de 5 % é 2.79.

Quando dois modelos aninhados são ajustados, aquele com mais parâmetros apresenta maior log L. Entretanto, esse não é necessariamente o melhor modelo. Isto significa que não se pode comparar diretamente os Log L quando o número de parâmetros varia entre modelos. Além do LRT, outro critério para a seleção de modelos é o Critério de Informação de Akaike (AIC), o qual penaliza a verossimilhança pelo número de parâmetros independentes ajustados. Por esse critério, qualquer parâmetro extra deve aumentar a verossimilhança por ao menos uma unidade para que o mesmo entre no modelo. O AIC é dado por AIC = -2 log L + 2 p, em que p é o número de parâmetros estimados. Menores valores de AIC refletem um melhor ajuste global (Akaike, 1974). Assim, os valores de AIC são calculados para cada modelo e aquele com menor valor de AIC é escolhido como melhor modelo. Há uma equivalência assintótica entre a escolha de modelos pelo critérios AIC e validação cruzada (Stone, 1977; Fang, 2011).

A comparação de modelos hierárquicos, mas com mesma estrutura de efeitos fixos, é realizada pelo LRT ou (análise de *deviance*), AIC, BIC e AICc. 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. 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.

O primeiro termo do AIC pode ser interpretado como uma medida de ajuste do modelo e o segundo termo como uma penalização. Desse modo, no caso em que se compara modelos com o mesmo número de parâmetros, necessita-se comparar apenas o Log L. A vantagem do AIC é que as comparações não se limitam a modelos com estrutura hierárquica de fatores, fato que faz do AIC uma ferramenta genérica para a seleção de modelos. Pode ser usado, por exemplo, para a comparação entre modelos com erros apresentando diferentes distribuições. O AICc é uma modificação que penaliza mais a adição de parâmetros quando o tamanho n da amostra é pequeno.

Outra abordagem é o Critério de Informação Bayesiano (BIC) de Schwarz (1978), o qual é dado por BIC = -2 log L + p log v, em que v = N − r(x) é o número de graus de liberdade do resíduo. O BIC é calculado para cada modelo e aquele com menor valor é escolhido como melhor modelo. Pode ser usado quando os modelos não possuem estrutura hierárquica. No entanto, os modelos devem ter a mesma estrutura de efeitos fixos. Logicamente, tanto o LRT, o AIC e o BIC dependem da mesma quantidade básica -2 log L.

A diferença entre as deviances de dois modelos com efeitos fixos diferentes não propicia um teste estatístico adequado. Isto deve-se ao fato de que a verossimilhança residual (função de y − Xb) é que é maximizada e não a

verossimilhança dos dados originais (função de y). A verossimilhança residual refere-se à verossimilhança dos dados após projeção no espaço residual e, portanto, dois diferentes modelos quanto aos efeitos fixos referem-se a duas diferentes projeções e, conseqüentemente, correspondem a diferentes conjuntos de dados nos quais os mesmos fatores aleatórios são estimados.

No contexto da estimação por máxima verossimilhança existem três testes assintoticamente equivalentes, dado a estimação do modelo, restrito ou reduzido ($\tilde{\theta}$) e sem restrição ou sem redução ($\hat{\theta}$):

Teste de Wald: procura medir a distância entre $\tilde{\theta}$ e $\hat{\theta}$.
Teste LRT: ocupa-se da distância entre $\log L(\tilde{\theta})$ e $\log L(\hat{\theta})$.
Teste do Multiplicador de Lagrange (LM) ou Escore Eficiente: compara as tangentes nos pontos $\tilde{\theta}$ e $\hat{\theta}$. O *Multiplicador de Lagrange* visa solucionar um problema de maximização (otimização) condicionada.

$L(\tilde{\theta})$ e $L(\hat{\theta})$ são valores da função de verossimilhança no ponto de máximo com e sem restrição. Se a restrição for verdadeira os valores da função de verossimilhança avaliada em $\tilde{\theta}$ e $\hat{\theta}$ são próximos, revelando que os dados dão suporte à restrição ou redução.

1.12 Modelos Computacionais BLUP

Considerando um vetor *y* de observações individuais, os seguintes modelos estatísticos equivalentes podem ser especificados:

(1) $y = Xb + e_1$: modelo com interesse apenas nos efeitos fixos (MEF).

$\hat{e} = y - Xb$: resíduos cheios = genéticos + ambientais aleatórios; equivalem aos valores genéticos desregressados.

(2) $y = Xb + Z(g_p/2 + g_m/2 + g_d) + e_2$: modelo reduzido de valores genéticos aditivos ou modelo individual reduzido (MIR).

$\hat{e} = (y - X\hat{b} - 0{,}5\ \hat{g}_p - 0{,}5\ \hat{g}_m)$: muita utilidade na seleção genômica ampla (GWS) = resíduo do MIR: corrigido para os genitores e desregressado.

(3) $y = Xb + Zg + e_2$: modelo de valores genéticos aditivos individuais ou modelo individual (MI).

$\hat{g}$: pouca utilidade direta na GWS.
$\hat{g}_d = (\hat{g} - 0{,}5\ \hat{g}_p - 0{,}5\ \hat{g}_m)$: muita utilidade na GWS = valor genético corrigido para os genitores.
$\hat{g}_d / h_d^2$: valor genético desregressado e corrigido para os genitores, em que $h_d^2 = (1/2\ h^2)/(1/2\ h^2 + (1-h^2))$ é a herdabilidade da segregação mendeliana e h^2 é a herdabilidade individual.

(4) $y = Xb + Z_m(g_m/2) + e_4 = Xb + Z_mf + e_4$: modelo de genitores femininos ou modelo gamético (MG).

$\hat{e} = (y - X\hat{b} - 0{,}5\ \hat{g}_p)$: resíduo do MG: corrigido para um genitor e desregressado.

(5) $y = Xb + X_m(g_m/2) + X_p(g_p/2) + Zg + e_5$: modelo ajustando genitores como de efeitos fixos. (Igual ao (3) se $r^2_{g\hat{g}} \rightarrow 1$).

$\hat{g} = \hat{g}_d$

$\hat{g} / h^2_d$: valor genético desregressado e corrigido para os genitores.

No modelo (1), o interesse reside apenas sobre os efeitos fixos (b) e todos os efeitos aleatórios (genético aditivo, genético de dominância, epistático e ambientais) são agrupados no resíduo aleatório e_1. O modelo (3) é o próprio modelo de valores genéticos aditivos individuais (g) e, o resíduo e_2 contempla os efeitos aleatórios de dominância alélica, epistasia e ambientais. No modelo (2), o valor genético aditivo individual (a) é dividido em 3 partes: (i) metade do valor genético aditivo da mãe = $g_m/2$; (ii) metade do valor genético aditivo do pai = $g_p/2$; (iii) segregação mendeliana ou desvio genético em relação à média dos valores genéticos aditivos dos genitores = g_d. O modelo (4) é expresso em termos da metade do valor genético aditivo dos genitores femininos ou do efeito de famílias f, sendo que e_4 compreende o somatório de $g_p/2$, g_d e e_2 . Nestes modelos, X, Z e Z_m são matrizes de incidência para b, a e $g_m/2$, respectivamente.

O modelo de interesse prático ao melhoramento refere-se ao (3) ou modelo individual (MI). Entretanto, tal modelo é o mais complexo computacionalmente, com número de equações para g igual ao número de descendentes mais o número de genitores em avaliação. O modelo individual reduzido - MIR- produz resultados idênticos ao MI, porém com um menor esforço computacional, podendo-se trabalhar com um número de equações igual ao número de genitores, obtendo-se as predições para g_p e g_m e, posteriormente, as predições para g_d e, conseqüentemente, para g. Um resumo dos modelos computacionais BLUP é apresentado na Tabela 12.

Tabela 12. Modelos Computacionais BLUP.

Nome	Modelo
Modelo Individual (Animal)	$y = Xb + Zg + e_2$
Modelo de Genitor (Reprodutor)	$y = Xb + Z_m(g_m/2) + e_4$
Modelo Individual (Animal) Reduzido	$y = Xb + Z(g_p/2 + g_m/2 + g_d) + e_2$
Modelo Individual (Animal) com Grupos Genéticos (r)	$y = Xb + Pr + Zg + e_2$

1.13 Modelos BLUP Univariados Multi-Efeitos

O BLUP univariado pode ser ajustado incluindo diferentes fatores de efeitos. Um resumo dos modelos computacionais univariados multi-efeitos para o BLUP é apresentado na Tabela 13.

Tabela 13. Modelos BLUP univariados multi-efeitos.

Nome	Modelo
Modelo de Repetibilidade com Ambiente Permanente (p)	$y = Xb + Zg + Tp + e$
Modelo com Efeito de Ambiente Comum (c)	$y = Xb + Zg + Tc + e$
Modelo com Interação Genótipos x Ambientes(ge)	$y = Xb + Zg + Tge + e$
Modelo com Efeito Materno (m)	$y = Xb + Zg + Zm + Tp + e$
Modelo com Efeito de Dominância (d)	$y = Xb + Zg + Td + e$
Modelo com Efeito de Heterose (h)	$y = Xb + Zg + Th + e$
Modelo com Efeitos Epistáticos (gg)	$y = Xb + Zg + Zgg + e$

*** X, Z e T são matrizes de incidência.**

1.14 Modelos BLUP Multivariados

O BLUP multivariado pode ser ajustado usando diferentes parametrizações e técnicas. Um resumo dos modelos multivariados para o BLUP é apresentado na Tabela 14.

Tabela 14. Modelos BLUP multivariados.

Modelo	Objetivo
Modelo Multivariado	Análise Simultânea de Variáveis, posto completo.
Componentes Principais sob Modelos Mistos (PCAM)	Análise Simultânea de Variáveis, posto reduzido.
Modelos Fator Analíticos Mistos (FAMM)	Análise Simultânea de Variáveis, posto reduzido. Interação Genótipos x Ambientes
Modelos de Normas de Reação via Regressão Aleatória	Interação Genótipos x Ambientes

A análise multivariada apresenta grande utilidade na formulação de índices de seleção (Resende et al., 1990; Lopes, 2005). A associação das técnicas de análise multivariada e de modelos mistos é importante para a análise de múltiplos caracteres, múltiplos experimentos e, em alguns casos, medidas repetidas. Para o caso de múltiplos caracteres, o uso da PCAM é mais adequado. Para múltiplos experimentos, a técnica FAMM é mais indicada. Isto porque a análise de componentes principais enfatiza a identificação de variáveis que explicam o máximo da variação total multivariada, fato que é relevante para o caso de múltiplos caracteres. Por outro lado, a análise de fatores enfatiza a atribuição da covariância entre variáveis a fatores comuns. Isto é relevante quando as variáveis referem-se a ambientes ou experimentos e todos os ambientes são alvo da análise e não apenas aqueles que mais contribuem para a variação total. Também, a covariância ou correlação entre ambientes atribuídas a fatores comuns automaticamente considera a similaridade e dissimilaridade entre ambientes, o que é uma propriedade interessante nesse contexto. Uma descrição detalhada e exemplo de aplicação da técnica FAMM na análise de múltiplos experimentos com interação g x e é apresentada por Resende e Thompson (2004).

Componentes principais sob modelos mistos (PCAM)

O método PCAM reuni as técnicas de análise multivariada e de modelos mistos e produz uma análise direta e em um só passo, no nível genético. Esta análise simultânea tem grande aplicação na análise de múltiplos caracteres e de medidas repetidas. A metodologia de modelos mistos padrão pode ser usada para estimar autovalores e autovetores diretamente sem a necessidade de se estimar a matriz de covariância (Σ) completa. A principal diferença para o modelo multivariado misto tradicional refere-se ao fato de que os parâmetros a serem estimados fazem parte da matriz de incidência dos efeitos genéticos aleatórios, conduzindo à estimação sob posto reduzido.

Outra vantagem dessa abordagem refere-se ao fato de que a estimação direta da estrutura de covariância garante que a matriz de covariância será positiva definida, fato que não é garantido por outros métodos de estimação de ∑. Assim, a inclusão de caracteres adicionais na análise contribui para aumentar a precisão na estimação ao invés de desestabilizar as estimativas. Também PCA's podem ser estimados tanto no nível genético quanto ambiental, desdobrando a tradicional PCA fenotípica. A seguir, é apresentada uma extensão dos modelos mistos para incorporar a análise de componentes principais.

Modelo Misto Tradicional

$$y = Xb + Zg + e$$

PCA sob Modelo Misto (PCAM)

$y = Xb + Z(Q \otimes I_g)(Q^{-1} \otimes I_g)g + \varepsilon = Xb + Z^* g^* + \varepsilon$, em que: $Q = V_p$ e $g_j^* = Q' g_j$.

Os valores genéticos do indivíduo j para os caracteres originais é dado por $\hat{g}_j = Q\hat{g}_j^*$. I_g é a matriz identidade com ordem igual ao número g de genótipos. Sob esse modelo, a matriz de covariância genética é dada por $\sum = \Lambda\,\Lambda'$, em que $\Lambda\,\Lambda' = VD_\alpha V'$, D_α é a matriz diagonal dos p autovalores e V é a matriz dos autovetores. Escolhendo-se V e D_α referentes apenas à dimensão p, esse modelo misto é reduzido e ajusta somente os primeiros componentes principais. Assim, na técnica PCAM, a estrutura de covariância é simplificada para $\sum^* = \Lambda_p\,\Lambda_p' = V_p D_{\alpha p} V_p'$ em que p indica uma das dimensões dessas matrizes (número de colunas).

Análise de fatores sob modelos multiplicativos mistos (FAMM)

A estrutura da matriz de covariância ou correlação envolvendo v caracteres está associada a v(v+1)/2 elementos. Visando simplificar a estrutura dessa matriz, sumarizar a informação multivariada e reduzir a dimensionalidade do problema, decomposições dessas matrizes, baseadas em seus autovalores e autovetores, são usadas com base em diferentes parametrizações produzindo as técnicas de componentes principais e da análise de fatores.

Entretanto, tais procedimentos são baseados na estimação completa da matriz de covariância ou de correlação com todos os seus v(v+1)/2 elementos. Um procedimento estatístico mais atrativo refere-se a estimar os componentes principais e os fatores diretamente, restringindo a estimação apenas àqueles mais importantes. Esse procedimento não requer a estimação prévia da matriz de covariância e de correlação e é sobretudo relevante no contexto dos modelos mistos e de dados desbalanceados. Nesse caso, torna-se necessária uma reparametrização dos modelos mistos tradicionais (Resende e Thompson, 2004). A seguir é apresentada uma extensão dos modelos mistos para incorporar a análise de fatores.

Modelo misto tradicional
$y = Xb + Zg + e$

Modelo misto fator analítico (FAMM)
$y = Xb + Z[(\Lambda \otimes I_g)f + \delta] + e$, em que: $a = [(\Lambda \otimes I_g)f + \delta]$

Sob esse modelo, a matriz de covariância genética é dada por $\sum = \Lambda\,\Lambda' + \psi$, em que $\Lambda\,\Lambda' = VD_\alpha V'$, D_α é a matriz diagonal dos m autovalores e V é a matriz dos autovetores. Escolhendo-se V e D_α referentes apenas à dimensão p, esse modelo misto é reduzido e ajusta somente os p fatores. Na técnica FAMM, a estrutura de covariância é simplificada para $\sum = \Lambda_p \Lambda_p + \Psi'$. Definem-se as seguintes quantidades: f é o vetor de escores fatoriais para os indivíduos nos fatores; δ é o vetor de erros representando a falta de ajuste do modelo fatorial; Λ é a matriz dos carregamentos dos fatores nas variáveis; Ψ é a matriz diagonal de variâncias específicas $Var(\delta_i)$ (Resende e Thompson, 2004).

A metodologia de modelos mistos padrão pode ser usada para estimar autovalores e autovetores diretamente sem a necessidade de se estimar Σ completa. A principal diferença para o modelo multivariado misto tradicional refere-se ao fato de que os parâmetros a serem estimados fazem parte da matriz de incidência dos efeitos genéticos aleatórios. Como a distribuição de $[(\Lambda \otimes I_g)f$ é singular, isto conduz à estimação sob posto reduzido, restrições devem ser impostas aos parâmetros do modelo fator analítico (Thompson et al., 2003).

1.15 Modelos BLUP Espaciais e de Competição (Efeitos Associativos) (SCM)

O BLUP sob modelos espaciais e de competição (SCM) pode ser ajustado usando diferentes parametrizações. Um resumo desses modelos é apresentado na Tabela 15.

Tabela 15. Modelos espaciais e de competição para o BLUP.

Nome	Modelo	Estrutura de Variâncias
Modelo Geoestatístico (Exponencial)	$y = Xb + Zg + e$ $e \sim N(0, \Sigma)$ $g \sim N(0, A\sigma_g^2)$	$\Sigma = \sigma_\varepsilon^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$ $\sum_c(\Phi_c) = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$
Modelos Autoregressivos	Idem acima	Idem acima
Modelos Ante-Dependência	Ver texto	Ver texto
Modelos ARIMA	Ver texto	Ver texto
Modelos Associativos de Competição	$y = Xb + Z\tau + NZ\phi + e$	$G^* = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$
Modelos Associativos e Espaciais de Competição	$y = Xb + Z\tau + NZ\phi + \xi + \eta$	$G^* = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$ $\Psi = \sigma_\varepsilon^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)] + I\sigma_\eta^2$

Modelos Espaciais

As variáveis com comportamento espacial são denominadas variáveis regionalizadas e mostram características intermediárias entre as variáveis verdadeiramente casuais ou aleatórias e aquelas completamente determinísticas, exatas ou matemáticas. A estatística clássica trata de variáveis aleatórias ao passo que a estatística espacial aborda estas variáveis mistas.

Tais variáveis regionalizadas apresentam uma aparente continuidade no espaço. A continuidade geográfica se manifesta pela tendência de a variável apresentar valores muito próximos (dependentes) em dois pontos vizinhos e muito diferentes em pontos distantes. Assim, não são realizações de uma variável aleatória, pois são correlacionadas. Gráficos contemplando a variabilidade espacial dos experimentos, denominados variogramas, ilustram o padrão de dependência espacial.

Algumas estatísticas permitem sumarizar as informações contidas nos diagramas e descrever a continuidade espacial. Estas estatísticas são: (i) o coeficiente de correlação entre valores separados por uma dada distância, ou seja, o coeficiente de autocorrelação, também denominado autocorrelação serial ou autocorrelação espacial; (ii) a covariância entre valores separados por uma distância (autocovariância); (iii) momento de inércia ou semivariância. Variogramas, correlogramas e covariograma para a descrição da continuidade espacial podem ser obtidos a partir da semivariância, autocorrelação e autocovariância, respectivamente, associados a diferentes distâncias em uma determinada direção.

A variabilidade espacial pode ser estudada basicamente por meio de duas classes de métodos: os métodos de análise de séries temporais e os métodos geoestatísticos. Por meio dos métodos de análise de séries temporais, tem sido usado de um modelo auto-regressivo de primeira ordem (AR1) para modelar os resíduos em uma dimensão do espaço e o uso do método REML para estimar os parâmetros do modelo. Em um modelo AR1, a autocorrelação $[\rho(Y_i , Y_j)]$ entre as observações Y_i e Y_j é uma função potência da distância entre as observações, de forma que $\rho(Y_i , Y_j) = \rho^{|i-j|}$, em que i e j referem-se às coordenadas espaciais e ρ é o coeficiente de autocorrelação.

Um modelo auto-regressivo de primeira ordem indica que somente a correlação entre observações imediatamente vizinhas são diretamente especificadas. Correlações entre vizinhos mais distantes surgem somente como consequências dessas correlações de primeira ordem. Modelos de ordem mais elevada (por exemplo um AR2) podem ser especificados, nos quais observações não adjascentes podem apresentar dependência direta, além daquela indireta contemplada pelo modelo AR1 (Resende e Sturion, 2001).

O modelo AR1 pode ser estendido para considerar a variabilidade em duas dimensões do espaço considerando processos (AR1⊗AR1) separáveis em duas direções: linhas e colunas. Neste modelo, a autocorrelação é dada por: $\rho\left(Y_{i,j}\ ,\ Y_{k,\ell}\right)= \rho_{lin}^{[i-k]}\ \rho_{col}^{[j-\ell]}$ para observações com coordenadas *i*, *j* e *k*, ℓ referentes a linhas e colunas, respectivamente (Cullis e Gleeson, 1991; Cullis et al., 1998).

Estes últimos modelos consideram os erros por meio de um processo auto-regressivo integrado de médias móveis (ARIMA (p, q, d)) que pode ser aplicado a duas dimensões: linhas e colunas. Tal modelo estendido é da forma ARIMA (p_1, d_1, q_1) x ARIMA (p_2, d_2, q_2). Estes modelos são denominados modelos com erros nas variáveis e consideram um efeito de tendência (ξ) mais um erro η independente ou efeito pepita. Assim, o vetor de erros é particionado em e = ξ + η. Os modelos de análise tradicionais não incluem o componente ξ.

O modelo é da forma $y = Xb + Zg + \xi + \eta$, em que ξ é o vetor aleatório de erros correlacionados e η é o vetor aleatório de erros não correlacionados. A variância dos resíduos é dada por Var(e) = Var (ξ + η)= Ψ, em que $\Psi = \sigma_\xi^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]+I\sigma_\eta^2$, sendo σ_ξ^2 a variância devida a tendência e σ_η^2 é a variância dos resíduos não correlacionados (Resende e Sturion, 2003). As matrizes $\sum_c(\Phi_c)$ *e* $\sum_r(\Phi_r)$ referem-se a matrizes de correlação auto-regressivas de primeira ordem com parâmetros de autocorrelação Φ_c e Φ_r e ordem igual ao número de colunas e número de linhas, respectivamente. Assim, ξ é modelado como um processo auto-regressivo separável de primeira ordem (AR1 x AR1) com matriz de covariância $Var(\xi) = \Sigma = \sigma_\xi^2[\sum_c(\Phi_c)\otimes\sum_r(\Phi_r)]$. As matrizes de correlação auto-regressivas são da forma:

$$\sum_c(\Phi_c) = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}.$$

Em um modelo com efeito de ambiente comum, as equações de modelo misto para o BLUP são dadas por:

$$\begin{bmatrix} X'X & X'Z & X'W & X'I \\ Z'X & Z'Z+A^{-1}\lambda_1 & Z'W & Z'I \\ W'X & W'Z^* & W'W+I\lambda_2 & W'I \\ I'X & I'Z^* & I'W & I'I+H^{-1}\lambda_3 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \\ \tilde{c} \\ \tilde{\xi} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ I'y \end{bmatrix}, \text{ em que:}$$

$$\lambda_1 = \frac{\sigma_\eta^2}{\sigma_g^2}; \qquad \lambda_2 = \frac{\sigma_\eta^2}{\sigma_c^2} \qquad \lambda_3 = \frac{\sigma_\eta^2}{\sigma_\xi^2}.$$

A e H são as matrizes de correlação para os efeitos g e ξ, respectivamente. A inversão de H é dada por $H^{-1} = [\sum_c^{-1}(\Phi_c) \otimes \sum_r^{-1}(\Phi_r)]$. A estimação da variância do erro correlacionado via REML pode ser dada por $\hat{\sigma}_\xi^2 = [\tilde{\xi}' H^{-1} \tilde{\xi} + \hat{\sigma}_\eta^2 \; tr \; (H^{-1}C^{44})]/N$, em que C^{44} advém da inversa da matriz dos coeficientes e N é o número total de dados.

Comparando-se as magnitudes de ξ e η pode-se inferir se a variável é predominantemente determinística ($\tilde{\xi} > \tilde{\eta}$) ou aleatória ($\tilde{\eta} > \tilde{\xi}$). Para o LRT, a comparação entre um modelo espacial e um não espacial deve considerar 3 graus de liberdade, referentes às estimativas dos parâmetros de variância ξ e parâmetros de autocorrelação Φ_c e Φ_r (ρ_c e ρ_r).

A geoestatística consiste basicamente de variografia e krigagem. A variografia usa variogramas para caracterizar e modelar a variação espacial. A krigagem usa a variação modelada para predizer valores, tais quais os BLUPs de erros ou pontos correlacionados. O variograma usa semivariâncias e pode ser usado em ambos os métodos de análise espacial: geoestatística e modelos de séries temporais. Pela geoestatística, o modelo padrão para ajuste de uma função ao variograma experimental em ensaios de campo é o exponencial.

Os procedimentos geoestatísticos consideram a heterogeneidade espacial de forma direta por meio da inclusão dos efeitos de tendência e correlação residual na modelagem da matriz de covariância residual. Como o modelo associado ao variograma é exponencial, os resíduos podem ser interpretados como uma realização de um processo auto-regressivo de primeira ordem (AR1). Isto faz sentido uma vez que o modelo AR1 projeta a auto-correlação para lags distantes, como uma função potência da distância entre plantas. O modelo exponencial faz o mesmo. Entretanto, os modelos geoestatísticos muitas vezes assumem isotropia (mesmo padrão de variação nas duas dimensões), o que pode ser inadequado para modelar a estrutura de variâncias nos experimentos de campo. Há equivalência entre a modelagem geoestatística exponencial e o modelo separável AR1 x AR1 para experimentos de campo. Em função desta equivalência e da facilidade em ajustar modelos anisotrópicos (variação diferenciada em duas dimensões) pela modelagem ARIMA, esta tem sido preferida. Adicionalmente, a separabilidade resulta em maior eficiência computacional em termos de tempo.

Modelos Espaciais na Análise de SNPs

Com a disponibilidade de marcadores SNPs a predição de valores genéticos por meio da seleção genômica ampla (GWS) consiste na substituição da matriz de correlação genética A entre indivíduos, obtida via pedigree pela matriz de correlação genética G entre indivíduos, obtida via marcadores. No caso, a matriz W de incidência dos marcadores nos indivíduos tem elementos dados por $2p_i$, (1 - $2p_i$) e (2 - $2p_i$) ou -1, 0 e 1, para os genótipos marcadores mm, Mm e MM, respectivamente, em que p_i é a frequência de um dos alelos do loco marcador i. A seguir é demonstrada a equivalência entre os modelos A-BLUP e G-BLUP.

Modelo A-BLUP

$y = Xb + Zg + \varepsilon$; $Var(g) = A\sigma_g^2$

Modelo Equivalente G-BLUP

$g = Wm$

$y = Xb + ZWm + \varepsilon$; $Var(Wm) = WI\sigma_m^2 W' = WW'\sigma_m^2$

, em que *m* é o vetor de efeitos genéticos (substituição alélica) dos marcadores.

Assim, $Var(g) = Var(Wm)$ e, portanto, $A\sigma_g^2 = WW'\sigma_m^2$. Desenvolvendo tem-se:

$A = WW'\sigma_m^2 / \sigma_g^2 = WW'\sigma_m^2 / [2\sum_i^n p_i(1-p_i)\sigma_m^2] = WW' / [2\sum_i^n p_i(1-p_i)]$ e

$A = WW' / [2\sum_i^n p_i(1-p_i)]$, pois $\sigma_g^2 = [2\sum_i^n p_i(1-p_i)]\sigma_m^2$ (Falconer, 1989).

Uma prova da validade da expressão $G = A = WW' / [2\sum_i^n p_i(1-p_i)]$ é apresentada a seguir:

Códigos na matriz W

Códigos	Códigos Centrados	Códigos Centrados com p_i = 0.5	Numerador do coeficiente de parentesco de Wright entre Irmãos Completos
0	$0 - 2p_i$	-1	0.0
1	$1 - 2p_i$	0	0.5
2	$2 - 2p_i$	1	1.0

Cálculo da matriz G = A

	Matriz W				Matriz WW'		Matriz G	
Individuo	Marca 1	Marca 2	Marca 3	Marca 4			$2\sum_i^n p_i(1-p_i) = 2$	
Individuo 1	-1	0	0	1	2	2	1	1
Individuo 2	-1	0	0	1	2	2	1	1

Numerador do coeficiente de parentesco de Wright entre clones = 1

Verifica-se que os dois indivíduos são idênticos (clones) considerando os 4 locos marcadores, apresentando correlação genética igual a 1 na matriz G. Com infinitos locos marcadores, G tende a A. G também contempla o parentesco médio nos vários locos mas, sob GWS com seleção de marcadores, são considerados especificamente os locos que controlam o caráter em questão. E se o numero de locos que controlam o caráter é finito, G é muito diferente de A.

Com heterogeneidade de variância entre SNPs e sendo D uma matriz diagonal ($diag(D) = \tau_i$, sendo τ_i o componente de variância associado ao loco marcador i) contemplando essa heterogeneidade, a modelagem da estrutura de variância torna-se (se $m \sim (0, D\sigma_m^2)$):

$Var(g) = Var(Wm) = WD\sigma_m^2 W' = WDW'\sigma_m^2$,

em que WDW' é uma matriz de incidência ponderada quadrática.

E a igualdade entre as matrizes de correlação genética entre indivíduos torna-se

$A\sigma_g^2 = WDW'\sigma_m^2$

$A = WDW'\sigma_m^2 / \sigma_g^2$

$A = G = WDW' / [2\sum_i^n p_i(1-p_i)]$.

Se m for parametrizado como $m \sim (0, D)$, tem-se $G = WDW' / \sigma_g^2 = WDW' / \{[2\sum_i^n p_i(1-p_i)]\sigma_m^2\}$. Em ambos os casos, a matriz G substitui a matriz A nas equações de modelo misto.

Com $m \sim (0, D)$ e quando W_p contém elementos centrados e padronizados dados por $w_{ij_p} = \frac{(w_{ij} - 2p_i)}{[2p_i(1-p_i)]^{1/2}}$, tem-se $G = WDW' / \sigma_g^2 = WDW' / (n\sigma_m^2)$.

Essa modelagem gera um método G-BLUP com heterogeneidade de variância e produz resultados similares aos obtidos pelo método BayesA (ver tópico 1.18 e capítulo 6).

Com heterogeneidade de frequências alélicas entre SNPs (contemplada em uma matriz diagonal D_p), a parametrização torna-se $G = WD * W'$, em que $D* = DD_P$, sendo $diag(D_P) = 1/[n2p_i(1-p_i)]$.

Considerando a correlação entre efeitos de SNPs dentro de cromossomos devido ao desequilíbrio de ligação entre eles, modelos espaciais podem ser adotados. Nesse caso, a matriz D deve ser substituída por uma matriz de correlação autoregressiva (AR1) contemplando essa covariância espacial.

Assim, $Var(g) = Var(Wm) = WD\sigma_m^2 W' = WDW'\sigma_m^2$ deve ser rescrita como $Var(g) = Var(Wm) = W\Sigma_{mc}\sigma_{mc}^2 W' = W\Sigma_{mc}W'\sigma_{mc}^2$ em que σ_{mc}^2 é a variância correlacionada de marcadores e Σ_{mc} é uma matriz de correlação autoregressiva de primeira ordem com parâmetro de autocorrelação ρ. Para o caso de 4 marcas, Σ_{mc} é dada por

$$\Sigma_{mc} = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}.$$

Um modelo autorregressivo com variâncias heterogêneas (ARH) também pode ser ajustado. Nesse caso, tem-se $Var(g) = Var(Wm) = W\Sigma_{mch}W'$ e para 3 marcas a estrutura de covariância é:

$$\Sigma_{mch} = \begin{bmatrix} \sigma^2_{m_1} & \sigma_{m1}\sigma_{m2}\rho^1 & \sigma_{m1}\sigma_{m3}\rho^2 \\ Sim. & \sigma^2_{m_2} & \sigma_{m2}\sigma_{m3}\rho^1 \\ & & \sigma^2_{m_3} \end{bmatrix}.$$

Se parte da variância entre SNPs é correlacionada e parte é independente ou não correlacionada, tem-se a estrutura Var(m) = Var (m_c + m_{nc})= Ψ, em que $\Psi = \Sigma_{mc}\sigma^2_{mc} + I\sigma^2_{mnc}$, em que σ^2_{mnc} é a variância de marcadores não correlacionada. No caso, tem-se $Var(g) = Var(Wm) = W\Psi W'$. Para o caso de 4 marcas, Ψ é dada por

$$\Psi = \begin{bmatrix} (\sigma^2_{mc}+\sigma^2_{mnc}) & \rho^1\sigma^2_{mc} & \rho^2\sigma^2_{mc} & \rho^3\sigma^2_{mc} \\ \rho^1\sigma^2_{mc} & (\sigma^2_{mc}+\sigma^2_{mnc}) & \rho^1\sigma^2_{mc} & \rho^2\sigma^2_{mc} \\ \rho^2\sigma^2_{mc} & \rho^1\sigma^2_{mc} & (\sigma^2_{mc}+\sigma^2_{mnc}) & \rho^1\sigma^2_{mc} \\ \rho^3\sigma^2_{mc} & \rho^2\sigma^2_{mc} & \rho^1\sigma^2_{mc} & (\sigma^2_{mc}+\sigma^2_{mnc}) \end{bmatrix}.$$

Outra estrutura de correlação que pode ser usada é associada a modelos ante-dependência estruturados (SAD), em que a estrutura da matriz de covariância é:

$$\Sigma_{mSAD} = \begin{bmatrix} \sigma^2_{m_1} & \sigma_{m1}\sigma_{m2}\rho_1 & \sigma_{m1}\sigma_{m3}\rho_1\rho_2 \\ Sim. & \sigma^2_{m_2} & \sigma_{m2}\sigma_{m3}\rho_2 \\ & & \sigma^2_{m_3} \end{bmatrix}$$

Modelos SAD nos métodos BayesA e BayesB foram aplicados por Yang e Tempelman (2012). Maiores detalhes sobre modelos espaciais na análise genômica são apresentados no tópico 6.26.

Modelos de Competição (Associativos ou de Interação Social)

Em um modelo de interferência ou de interação social, a parcela ou individuo i tem um efeito direto τ_i nele e um efeito indireto ϕ_i no individuo vizinho. A competição genotípica pode ser considerada sob a ótica desse modelo. Esse modelo é da forma: $y = Xb + Zg + e = Xb + Z\tau + NZ\phi + e$, em que:

$Zg = Z\tau + NZ\phi$.

τ: vetor dos efeitos genéticos diretos dos indivíduos (genótipos).

ϕ : vetor dos efeitos centrados de tratamentos (genótipos) sobre os vizinhos (efeitos indiretos ou associativos), os quais são genéticos e não fenotípicos. São também denominados efeitos genéticos sociais.

N: matriz de incidência de vizinhança, de dimensão n x n, composta por 0 e 1.

Pode ser visto explicitamente no modelo genético social que os efeitos de competição referem-se a efeitos genéticos (dependem da matriz Z) e não a efeitos residuais. Devido a essa razão, o uso somente da abordagem auto-regressiva para os resíduos pode ser inapropriada para contemplar a competição entre indivíduos ou entre parcelas.

O componente ϕ_i pode ser positivo ou negativo, dependendo da agressividade do genótipo. Se negativo (para genótipos agressivos), o valor absoluto de ϕ_i deve ser subtraído de τ_i por meio de $\tau_i^* = \tau_i - v\phi_i$, propiciando os efeitos de genótipos para uso em plantios ou planteis puros, em que v é o número de vizinhos considerados. Se positivo (genótipos sensíveis), ϕ_i será somado na expressão $\tau_i^* = \tau_i + v\phi_i$.

A competição e a tendência espacial podem ser incluídas em um modelo espacial. O modelo é da forma: $y = Xb + Z\tau + NZ\phi + \xi + \eta$. A competição é modelada como parte da estrutura de tratamentos e a tendência em uma ou duas dimensões é modelada como parte da estrutura dos erros.

Resende e Thompson (2003) e Resende et al. (2005) usaram esse mesmo modelo e assumiram τ_i e ϕ_i como efeitos aleatórios. Nesse caso, existe uma covariância entre τ_i e ϕ_i. A matriz de covariância entre eles equivale a:

$G = \begin{pmatrix} g_{\tau\tau} & g_{\tau\phi} \\ g_{\tau\phi} & g_{\phi\phi} \end{pmatrix}$, em que $g_{\tau\tau}$ é o componente de variância para os efeitos genotípicos diretos, $g_{\phi\phi}$ é o componente de variância associado aos efeitos genotípicos indiretos sobre os vizinhos (numerador da herdabilidade dos efeitos de competição) e $g_{\tau\phi}$ é a covariância entre os efeitos diretos no próprio indivíduo e indiretos sobre os vizinhos e é também o numerador da correlação genética entre a produtividade e a agressividade das indivíduos, dada por $r_{\tau\phi} = g_{\tau\phi} / (g_{\tau\tau} g_{\phi\phi})^{1/2}$.

Esta correlação é, em geral, negativa, evidenciando que os melhores indivíduos são beneficiados nos experimentos e criações comunitárias. A seleção deve então ser baseada em $\hat{\tau} + \hat{\phi}$, em que $\hat{\phi}$ é negativo nas indivíduos mais agressivos. A seleção pode basear-se também em $\hat{\phi}$ visando a identificação de genótipos adequados a plantios adensados como, por exemplo, no melhoramento do cafeeiro e do dendezeiro.

Modelo idêntico ao apresentado passou a ser usado também no melhoramento animal (Van Vleck e Cassady, 2005; Arango et al., 2005; Muir, 2005). Atualmente, esses mesmos modelos vem sendo enfatizados novamente no melhoramento florestal (Brotherstone et al., 2011; Bijima, 2011; Costa e Silva et al., 2012).

1.16 Modelos BLUP Longitudinais (Regressão Aleatória Multivariada e Normas de Reação)

Dados longitudinais ou medidas repetidas ao longo do tempo são a regra no melhoramento de animais e plantas perenes. O BLUP sob modelos com medidas repetidas pode ser ajustado usando diferentes parametrizações da estrutura de correlação dos fenótipos ao longo do tempo. Esse assunto é tratado por Meyer (2005) e Mrode (2005), dentre outros. Um resumo desses modelos é apresentado na Tabela 16.

Tabela 16. Modelos BLUP Longitudinais.

Modelo	Objetivo	Modelos
Modelos de Regressão Aleatória via Polinômios Ortogonais de Legendre	**Modelagem de medidas repetidas no tempo**	Ver texto
Modelos de Regressão Aleatória via Splines	**Modelagem de medidas repetidas no tempo**	Ver texto
Modelos Processo Caráter e Autoregressivos	**Modelagem de medidas repetidas no tempo**	Ver texto
Modelos Ante-Dependência Estruturados (SAD)	**Modelagem de medidas repetidas no tempo**	Ver texto
Modelos de Simetria Composta	**Modelagem de medidas repetidas no tempo**	Ver texto

Regressão Aleatória Multivariada

Para caracteres associados a curvas de crescimento em função do tempo ou da idade de avaliação, os modelos de regressão aleatória multivariados (RRM) devem ser adotados considerando dois conjuntos de regressão dos fenótipos do caráter em função das idades mensuradas. O primeiro conjunto diz respeito à regressão fixa para os indivíduos pertencentes à mesma classe de efeitos fixos e o segundo contempla efeitos aleatórios que descrevem os desvios de cada indivíduo em relação à regressão fixa. As regressões fixas e aleatórias são representadas por funções contínuas.

Um modelo de regressão aleatória multivariado pode ser ajustado para os efeitos aleatórios genético aditivo e ambiente permanente cujas covariáveis relacionadas aos tempos ou idades podem ser descritas por polinômios de Legendre. Esse modelo é dado por y = Xb + Zg + Tp + e, em que p é o vetor dos efeitos de ambiente permanente com matriz de incidência T. Expresso de outra forma, o modelo é dado por $y = Xb + \Phi_g g + \Phi_p p + e$, em que Φ_g e Φ_p são matrizes de incidência (de covariáveis) para os coeficientes polinomiais dos efeitos genético aditivo e de ambiente permanente, respectivamente.

As distribuições dos coeficientes de regressão aleatória são dadas por: $g \sim N(0, A \otimes K_g)$, sendo A a matriz de parentesco entre os indivíduos e K_g uma matriz de dimensão (k_g +1) x (k_g+1) de covariâncias entre coeficientes de regressão aleatória para os efeitos genéticos aditivos; $p \sim N(0, I_n \otimes K_p)$, sendo I_n uma matriz identidade de ordem n e K_p uma matriz de dimensão (k_p+1) x (k_p+1) de covariâncias entre coeficientes de regressão aleatória para os efeitos de ambiente permanente. Com seleção genômica, os modelos de regressão aleatória multivariados devem usar, em lugar de A, a matriz de parentesco genômico, dada por $G = (WW')/k = (WW')/[2\sum_i^n p_i(1-p_i)]$.

O modelo de regressão aleatória pode ser dado por:

$$y_{ij} = Fij + \sum_{m=0}^{k_\beta - 1} \beta_m \phi_m(a_{ij}^*) + \sum_{m=0}^{k_g - 1} g_{im} \phi_m(a_{ij}^*) + \sum_{m=0}^{k_p - 1} p_{im} \phi_m(a_{ij}^*) + e_{ij}$$

em que:

y_{ij} variável observada no j^{ésima} idade do i^{ésimo} indivíduo;

F_{ij} conjunto de efeitos fixos;

β_m : m^{ésimo} coeficiente de regressão de efeito fixo da curva média da variável na população;

g_{im} e p_{im}: m^{ésimos} coeficientes de regressão aleatória referentes aos efeitos genético aditivo e de ambiente permanente, respectivamente, para o i^{ésimo} indivíduo;

k_g e k_p : ordens das funções de covariâncias utilizadas para descrever, respectivamente, os efeitos genético aditivo e de ambiente permanente;

a_{ij}^* : idade j do indivíduo i;

$\phi_m(a_{ij}^*)$: polinômios de Legendre avaliados para a_{ij}^*, referentes a regressão de efeito fixo e aos efeitos aleatórios genético e de ambiente permanente, considerando as ordens das funções de covariâncias k_d e k_p, respectivamente;

e_{ij} : efeito aleatório residual.

O modelo matricial equivalente $y = Xb + \Phi_g g + \Phi_p p + e$ é caracterizado a seguir:

$$\Phi_g = \begin{bmatrix} \Phi_{1g} & 0 & 0 & 0 & 0 \\ 0 & \Phi_{2g} & 0 & 0 & 0 \\ 0 & 0 & \Phi_{3g} & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \Phi_{Ng} \end{bmatrix}$$

, em que g e p são os vetores dos coeficientes de regressão aleatória referentes aos efeitos genético aditivo e de ambiente permanente, respectivamente.

A matriz Φ_{jg} para o indivíduo j contém os elementos $\phi_m(a_{ij}^*)$ ou polinômios de Legendre avaliados para a_{ij}^* (idade padronizada i para o individuo j) e é dada por Φ_{jg} = MΛ. A matriz Λ apresenta dimensão k_g x k_g , em que k_g refere-se à ordem da função de covariância utilizada. A matriz M (de dimensão t x k_g, em que t é o número de idades avaliadas no indivíduo j) contém os valores de idade padronizados.

Os polinômios de Legendre são denotados por $P_n(x)$. Definindo $P_0(x) = 1$, o polinômio *n*+1 é descrito pela fórmula de recorrência:

$$P_{n+1}(x) = \frac{1}{n+1}\left((2n+1)xP_n(x) - nP_{n-1}(x)\right).$$

Assim, $P_1(x) = x$,

$P_2(x) = \frac{1}{2}\left(3xP_1(x) - 1P_0(x)\right) = \left(\frac{3}{2}x^2 - \frac{1}{2}\right)$,

$$P_3(x) = \frac{x - \frac{5}{3}x^3}{-\frac{2}{3}} = \frac{5}{2}x^3 - \frac{3}{2}x$$

$P_4(x) = \frac{1}{8}\left(35x^4 - 30x^2 + 3\right)$ e assim sucessivamente.

Na forma "normalizada" tem-se: $\phi_n(x) = \left(\frac{2n+1}{2}\right)^{0,5} P_n(x)$ e então tem-se a série de polinômios ortogonais: $\phi_0(x) = \left(\frac{1}{2}\right)^{0,5} P_0(x) = 0,7071$

$\phi_1(x) = \left(\frac{3}{2}\right)^{0,5} P_1(x) = 1,22467x$

$\phi_2(x) = \left(\frac{5}{2}\right)^{0,5} \left(\frac{3}{2}x^2 - \frac{1}{2}\right) = -0,7906 + 2,3717x^2$ e assim por diante.

Em resumo podem ser apresentados da seguinte maneira:

n	$P_n(x)$
0	1
1	x
2	$\frac{1}{2}(3x^2-1)$
3	$\frac{1}{2}(5x^3-3x)$
4	$\frac{1}{8}(35x^4-30x^2+3)$
5	$\frac{1}{8}(63x^5-70x^3+15x)$
6	$\frac{1}{16}(231x^6-315x^4+105x^2-5)$
7	$\frac{1}{16}(429x^7-693x^5+315x^3-35x)$
8	$\frac{1}{128}(6435x^8-12012x^6+6930x^4-1260x^2+35)$
9	$\frac{1}{128}(12155x^9-25740x^7+18018x^5-4620x^3+315x)$
10	$\frac{1}{256}(46189x^{10}-109395x^8+90090x^6-30030x^4+3465x^2-63)$

Empregando polinômios na forma normalizada, tem-se que os elementos de Λ são dados por $\phi_n(x) = \left(\frac{2n+1}{2}\right)^{0,5} P_n(x)$. Considerando k_g = 6 a ordem da função de covariância utilizada, a matriz Λ (de dimensão k_g x k_g) dos coeficientes dos polinômios de Legendre é definida da seguinte forma:

$$\Lambda = \begin{bmatrix} 0,7071 & 0 & -0,7906 & 0 & 0,7955 & 0 \\ 0 & 1,2247 & 0 & -2,8062 & 0 & 4,3973 \\ 0 & 0 & 2,3717 & 0 & 7,9550 & 0 \\ 0 & 0 & 0 & 4,6771 & 0 & -20,5206 \\ 0 & 0 & 0 & 0 & 9,2808 & 0 \\ 0 & 0 & 0 & 0 & 0 & 18,4685 \end{bmatrix}$$

A matriz M, considerando a avaliação de 6 idades no indivíduo j é dada por

$$M = \begin{bmatrix} 1 & a_1 & a_1^2 & a_1^3 & a_1^4 & a_1^5 \\ 1 & a_2 & a_2^2 & a_2^3 & a_2^4 & a_2^5 \\ 1 & a_3 & a_3^2 & a_3^3 & a_3^4 & a_3^5 \\ 1 & a_4 & a_4^2 & a_4^3 & a_4^4 & a_4^5 \\ 1 & a_5 & a_5^2 & a_5^3 & a_5^4 & a_5^5 \\ 1 & a_6 & a_6^2 & a_6^3 & a_6^4 & a_6^5 \end{bmatrix}.$$

A quantidade a_t refere-se à idade padronizada para o intervalo -1 ; 1 e é dada por $a_t = -1 + 2(a_t - a_{\min})/(a_{\max} - a_{\min})$.

Para o caso das idades 60, 150, 300, 420, 500 e 620 dias, o vetor *a* das idades padronizadas é dado por *a'*=

[-1.0000 -0.6786 -0.1429 0.2857 0.5714 1.0000].

A matriz M equivale então a M =

1	-1	1	-1	1	-1
1	-0.6786	0.4605	-0.3125	0.2120	-0.1439
1	-0.1429	0.0204	-0.0029	0.0004	-0.0001
1	0.2857	0.0816	0.0233	0.0067	0.0019
1	0.5714	0.3265	0.1866	0.1066	0.0609
1	1	1	1	1	1

Finalmente a matriz $\Phi_{jg} = \Phi_{jp}$ para o indivíduo j é dada por $\Phi_{jg} = M\Lambda$.

Φ_{jg} =

0.7071	-1.2247	1.5811	-1.8709	2.1213	-2.3452
0.7071	-0.8311	0.3016	0.4427	-0.9002	0.7711
0.7071	-0.1750	-0.7422	0.3874	0.6369	-0.5707
0.7071	0.3499	-0.5971	-0.6928	0.2086	0.8133
0.7071	0.6998	-0.0162	-0.7307	-0.8125	-0.1918
0.7071	1.2247	1.5811	1.8709	2.1213	2.3452

Com t = k tem-se o caso de ajuste completo (full fit) e o modelo de regressão aleatória reproduz exatamente o modelo multicaracterístico. Assim, a matriz de covariância genética (Σ_g) do modelo multicaracterístico é exatamente reconstituída por $\Sigma_g = \Phi_g K_g \Phi_g'$. Em um modelo multivariado tem-se $y = Xb + Zg + \varepsilon$, com $Var(y) = Z\Sigma_g Z' + R$, em que $\Sigma_g = A \otimes \Sigma_{g0}$ e $R = I \otimes R_0$, sendo $\Sigma_{g0} = \Phi_g K_g \Phi_g'$, $Var(Zg) = Var(\Phi_g g)$. E sendo $R_0 = \Phi_p K_p \Phi_p' + I\sigma_e^2$, tem-se que $Var(\varepsilon) = Var(\Phi_p p + e)$.

Na prática, usa-se um modelo com ajuste reduzido, ou seja, tem-se k < t e $\Sigma_g = \Phi K_g \Phi' + \nu$, em que ν é um desvio em relação ao modelo multivariado total com as t idades. Bons ajustes conduzem a ν desprezíveis.

As equações de modelo misto são dadas por

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}\Phi_g & X'R^{-1}\Phi_p \\ \Phi_g'R^{-1}X & \Phi_g'R^{-1}\Phi_g + A^{-1} \otimes K_g^{-1} & \Phi_g'R^{-1}\Phi_p \\ \Phi_p'R^{-1}X & \Phi_p'R^{-1}\Phi_g & \Phi_p'R^{-1}\Phi_p + I^{-1} \otimes K_p^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ \Phi_g'R^{-1}y \\ \Phi_p'R^{-1}y \end{bmatrix}$$

A matriz de covariância (K_g) entre os efeitos genéticos aleatórios, desconsiderando as relações de parentesco e para o caso de um ajuste linear equivale a:

$K_g = Var\begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix} = \begin{pmatrix} Var(\alpha_0) & Cov(\alpha_0,\alpha_1) \\ Cov(\alpha_0,\alpha_1) & Var(\alpha_1)\end{pmatrix} = \begin{pmatrix} g_{\alpha_0\alpha_0} & g_{\alpha_0\alpha_1} \\ g_{\alpha_0\alpha_1} & g_{\alpha_1\alpha_1}\end{pmatrix}$, em que para o individuo j o vetor g é dado por $g_j = [\alpha_{0j} \quad \alpha_{1j}]'$ e α_{0j} e α_{1j} são o intercepto e a inclinação do individuo em função da idade. E o vetor no tempo t é dado por $\phi_i = \begin{bmatrix}1 \\ t\end{bmatrix}$.

Voltando ao modelo inicial tem-se:

$y_t = Xb + g_t + p_t + e_t$

$y_t = Xb + (\Phi_g g)_t + (\Phi_p p)_t + e_t$

$y_t = Xb + (\Phi_g \alpha_g)_t + (\Phi_p p)_t + e_t$

Para o caso de um ajuste linear, o efeito aleatório do valor genético no tempo t é dado por: $g_t = \alpha_0 + \alpha_1 t$. Matricialmente, tem-se: $g_t = \phi_i \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix} = \begin{bmatrix}1 & t\end{bmatrix}\begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix}$.

De acordo com a ordem k do ajuste os efeitos genéticos aditivos são dados por:

Ordem k do Ajuste	g_t
1	α_0
2	$\alpha_0 + \alpha_1 t$;
3	$\alpha_0 + \alpha_1 t + \alpha_2 t^2$

Com k = 3, $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ e $\Phi_i = \begin{bmatrix}1 \\ t \\ t^2\end{bmatrix}$.

As variâncias genética e fenotípica são dependentes da idade, ou seja, podem aumentar ou diminuir com a idade. Tem-se que:

$\hat{\sigma}^2_{g(i)} = \phi_i' \hat{K}_g \phi_i = g_{\alpha_0\alpha_0} + 2t_i g_{\alpha_0\alpha_1} + t_i^2 g_{\alpha_1\alpha_1}$: variância genética na idade *i*.

$\hat{\sigma}^2_{y(i)} = \hat{\sigma}^2_{g(i)} + \hat{\sigma}^2_e$: variância fenotípica na idade *i*.

$\hat{\sigma}_{g(ij)} = \phi_i' \hat{K}_g \phi_j = g_{\alpha_0\alpha_0} + (t_i + t_j) g_{\alpha_0\alpha_1} + t_i t_j g_{\alpha_1\alpha_1}$: covariância genética entre as idades *i* e *j*.

$r_{g(ij)} = \dfrac{\hat{\sigma}_{g(ij)}}{\hat{\sigma}_{g(i)}\ \hat{\sigma}_{g(j)}}$: é a correlação genética entre as idades i e j.

Para o caso de um ajuste linear, um modelo sem efeito de ambiente permanente pode ser escrito como (Resende e Rosa-Perez, 1999; Resende et al, 2001):

$y = Xb + Z_o \alpha_0 + Z_1 \alpha_1 + e$, em que:

Z_o : matriz de incidência para α_0 , contendo *o* e *1's*.

Z_1 : matriz associando α_1 a y, contendo zero e valores de idade.

As equações de modelo misto podem ser formuladas:

$$\begin{bmatrix} X'X & X'Z_0 & X'Z_1 \\ Z_0'X & Z_0'Z_0 + A^{-1}\lambda_{00} & Z_0'Z_1 + A^{-1}\lambda_{01} \\ Z_1'X & Z_1'Z_0 + A^{-1}\lambda_{01} & Z_1'Z_1 + A^{-1}\lambda_{11} \end{bmatrix} \begin{pmatrix} \hat{b} \\ \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix} = \begin{pmatrix} X'y \\ Z_0'y \\ Z_1'y \end{pmatrix}, \text{ em que:}$$

$$K_g^{-1}\sigma_e^2 = \begin{pmatrix} \lambda_{00} & \lambda_{01} \\ \lambda_{01} & \lambda_{11} \end{pmatrix}, \text{ sendo } K_g = \begin{pmatrix} g_{\alpha_0\alpha_0} & g_{\alpha_0\alpha_1} \\ g_{\alpha_0\alpha_1} & g_{\alpha_1\alpha_1} \end{pmatrix}.$$

Modelos de Normas de Reação

O modelo de normas de reação refere-se ao estudo da interação genótipos x ambientes em termos de resposta fenotípica à variação em um gradiente ambiental representado por diferentes locais. Têm-se os seguintes modelos equivalentes:

$y = Xb + \Gamma\ell + Z_0 g + \Delta g\ell + e$

$y = Xb + \Gamma\ell + Z_0\alpha + Z_1\beta + e$

em que ℓ e $g\ell$ são os efeitos de ambientes ou locais e da interação genótipos x locais, com matrizes de incidência Γ e Δ, respectivamente.

No segundo modelo os efeitos de genótipos e da interação genótipos x locais são expressos como combinação genótipos-locais ($Z_0 g + \Delta g\ell = Z_0\alpha + Z_1\beta$), permitindo inferir sobre o desempenho de cada genótipo em cada local. Para isso define-se α e β como vetores dos coeficientes de regressão aleatória referentes aos efeitos genéticos de intercepto para cada genótipo e da inclinação para cada genótipo em função da ambiente. Define-se ainda:

Z_o : matriz de incidência para α , contendo *o* e *1's*.

Z_1 : matriz associando β a y, contendo zero e valores de médias por local.

As equações de modelo misto são:

$$\begin{bmatrix} X'X & X'Z_0 & X'Z_1 \\ Z_0'X & Z_0'Z_0 + A^{-1}\lambda_{00} & Z_0'Z_1 + A^{-1}\lambda_{01} \\ Z_1'X & Z_1'Z_0 + A^{-1}\lambda_{01} & Z_1'Z_1 + A^{-1}\lambda_{11} \end{bmatrix} \begin{pmatrix} \hat{b} \\ \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} X'y \\ Z_0'y \\ Z_1'y \end{pmatrix}, \text{ em que:}$$

$$N_r^{-1}\sigma_e^2 = \begin{pmatrix} \lambda_{00} & \lambda_{01} \\ \lambda_{01} & \lambda_{11} \end{pmatrix}, \text{ sendo } N_r = Var\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}.$$

Os efeitos aleatórios, α e β, correspondentes a cada genótipo são assumidos com distribuição normal de média nula e matriz de covariância dada por:

$N_r = Var\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}$, em que σ_α^2, σ_β^2 e $\sigma_{\alpha\beta}$ são a variância genética do intercepto, componente de variância da inclinação da norma de reação e covariância entre efeitos genéticos de intercepto e de inclinação, respectivamente.

A herdabilidade em função do gradiente ambiental é estimada por: $h_g^2 \mid \ell = \frac{\sigma_g^2 \mid \ell}{\sigma_g^2 \mid \ell + \sigma_e^2} = \frac{\sigma_g^2 \mid \ell}{\sigma_\alpha^2 + \sigma_\beta^2 \ell^2 + 2\sigma_{\alpha,\beta}\ell + \sigma_e^2}$, pois $\sigma_g^2 \mid \ell = Var(\alpha + \beta\ell) = \sigma_\alpha^2 + \sigma_\beta^2 \ell^2 + 2\sigma_{\alpha,\beta}\ell$.

No ambiente médio ($\ell = 0$): $h_g^2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}$ é o coeficiente de herdabilidade e σ_e^2 é a variância residual.

A correlação entre intercepto e inclinação das normas de reação são dadas por: $r_{\alpha,\beta} = \frac{\sigma_{\alpha,\beta}}{\sqrt{\sigma_\alpha^2 \sigma_\beta^2}}$. Essa correlação, quando tende a 1, indica que os indivíduos de maior valor genético são também mais responsivos (com grande adaptabilidade) à melhoria do ambiente. Essa é uma situação favorável.

O valor genético dos genótipos no ambiente ℓ é dado pela soma do intercepto α com o produto do coeficiente de inclinação β pelo valor do nível ambiental ℓ, da seguinte maneira: $g_i \mid \ell = \alpha_i + \beta_i \ell$. De maneira genérica, para todo o vetor g, tem-se $g \mid \ell = Z_0 \alpha_i + Z_1 \beta$.

As correlações entre valores genéticos em dois ambientes i e k são dadas por:

$r_{g|\ell_i, g|\ell_k} = \frac{\sigma_{g|\ell_i, g|\ell_k}}{\sqrt{\sigma^2_{g|\ell_i} \sigma^2_{g|\ell_k}}}$, em que $\sigma_{g|\ell_i, g|\ell_k} = \sigma_\alpha^2 + \sigma_\beta^2 \ell_i \ell_k + \sigma_{\alpha,\beta}(\ell_i + \ell_k)$ e $\sigma_g^2 \mid \ell_i = Var(\alpha + \beta\ell_i) = \sigma_\alpha^2 + \sigma_\beta^2 \ell_i^2 + 2\sigma_{\alpha,\beta}\ell_i$.

Os modelos de normas de reação podem ser ajustados via modelos de regressão aleatória por REML ou por via Bayesiana usando MCMC. Inclusive, modelos de regressão quadrática, cúbica e de maiores graus podem também ser avaliados. E pela abordagem Bayesiana, comparações entre modelos podem ser realizadas via AIC ou BIC usando uma deviance a posteriori. O método BIC usa o número efetivo de parâmetros, o qual é menor do que o numero bruto, devido ao parentesco.

1.17 Casos Especiais: GLMM, GEE, HGLMM, PL, PLS e SALP

A seguir são descritos casos especiais de modelos mistos envolvendo: analise de dados categóricos (GLMM), dados categóricos multivariados (GEE), modelos lineares mistos generalizados hierárquicos (HGLMM), perfil de verossimilhança (PL) e análise de sobrevivência para longevidade e precocidade (SALP). Um resumo sobre GLMM e GEE é apresentado na Tabela 17.

Tabela 17. Caracterização de modelos lineares generalizados mistos (GLMM) e equações de estimação generalizada (GEE) .

Classe de Modelos	Dimensão do Modelo	Função Associada à Variável Aleatória Discreta	Classificação do Modelo quanto aos Efeitos	Método de Estimação	Algoritmo Numérico
Modelos Lineares Generalizados (GLM)	Univariada	Verossimilhança	Fixo	Máxima Verossimilhança (ML)	Quadrados Mínimos Ponderados Iterativos (IWLS)
Modelos Lineares Generalizados (GLM)	Multivariada	Quase-Verossimilhança	Fixo	Equações de Estimação Generalizada (GEE)	Quadrados Mínimos Ponderados Iterativos (IWLS)
Modelos Lineares Generalizados Mistos (GLMM)	Univariada	Verossimilhança Residual	Misto	Máxima Verossimilhança Residual (REML)	Vários
Modelos Lineares Generalizados Mistos (GLMM)	Multivariada	Quase-Verossimilhança	Misto	Pseudo Máxima Verossimilhança ou REML Condicional	Vários

Modelos Lineares Generalizados Mistos (GLMM)

Variáveis não normais e não contínuas, como aquelas com distribuição binomial e outras variáveis categóricas, não são bem descritas por modelos estatísticos lineares. Para estas variáveis discretas, os modelos não lineares podem ser mais apropriados. A classe de modelos lineares generalizados permite a generalização ou flexibilização dos modelos lineares clássicos de variáveis contínuas, de forma que toda a estrutura para a estimação e predição em modelos lineares normais pode ser estendida para os modelos não lineares. Os modelos lineares clássicos são casos especiais de modelos lineares generalizados.

Estes modelos generalizados foram desenvolvidos para análise de dados associados a distribuições pertencentes à família exponencial com um parâmetro. A idéia de modelos lineares generalizados é permitir maior flexibilidade de análise. Tal idéia relaxa a suposição de que Y segue distribuição normal e permite que esta siga qualquer distribuição que pertença à família exponencial na forma canônica. As generalizações ocorrem em duas direções: (i) permitem que a esperança μ, de Y seja uma função monotonicamente diferençável do preditor linear $\eta = \sum x_i\beta_i$ de forma que $\mu = f(\eta) = f(\sum x_i\beta_i)$; (ii) ou, por inversão, $g(\mu) = \eta$, em que g é a função de ligação, a qual liga a média ao preditor linear. A incorporação da função de ligação nas equações de modelos lineares mistos para e estimação de componentes de variância e de efeitos fixos e predição de variáveis aleatórias gera a denominação de modelo não

linear devido à relação não linear que existe entre a escala latente e a probabilidade de um indivíduo pertencer a uma determinada categoria da variável discreta.

Para dados binomiais, $0 \leq \mu \leq 1$, funções de ligação tal qual a logito são utilizadas para satisfazer esta restrição natural. As transformações são importantes para: (i) estender a amplitude da variável analisada de (0,1) para a reta real; (ii) fazer a variância constante através da amplitude dos efeitos fixos (na escala da variável latente contínua). A função de ligação descreve, então, a relação existente entre o preditor linear (η) e o valor esperado μ de Y . No modelo linear clássico, tem-se $\eta = \mu$ que é chamada de ligação identidade, e esta ligação é adequada no sentido em que ambos η e μ podem assumir valores na reta real.

As distribuições a serem assumidas para a escala da variável latente e correspondentes funções de ligação devem ser capazes de transformar o intervalo (0,1) em $(-\infty, \infty)$. Neste sentido, as distribuições logística, normal padrão e Gumbel (ou distribuição de valor extremo) para a variável latente e suas correspondentes funções de ligação denominadas logito, probito e complemento log-log são apropriadas para o modelo binomial. Maiores detalhes sobre a estimação e predição em modelos lineares generalizados mistos via REML/BLUP são apresentados por Resende e Biele (2002).

Equações de Estimação Generalizada (GEE)

Análises estatísticas univariadas de variáveis discretas são realizadas eficientemente via a classe de modelos lineares generalizados. Nesse caso, uma função de verossimilhança é maximizada iterativamente analisando uma variável linearizada (transformação de y para a escala linear), usando modelos lineares normais ponderados. Modelos mistos normais ponderados podem ser ajustados via REML.

Para o caso multivariado, a estatística clássica tem se limitado a técnicas descritivas não paramétricas tal qual a análise de componentes principais ou a modelos paramétricos baseados em normalidade. Em muitas aplicações, principalmente na área de estatística médica, muitos problemas de estimação associados a variáveis discretas não podem ser abordados usando a estatística multivariada tradicional. Para o caso de variáveis não normais, uma forma geral para a distribuição multivariada não existe. Isto conduz ao fato de que uma verdadeira função de verossimilhança, que baseia-se em normalidade, não está disponível. Uma função alternativa é a quase-verossimilhança, a qual tem propriedades similares às da verossimilhança verdadeira. Essa função de quase-verossimilhança pode ser maximizada usando a técnica das equações de estimação generalizada (GEE) criada por Zeger et al. (1988). Por essa técnica, a estimação pode ser realizada por meio do método numérico ou algoritmo de quadrados mínimos ponderados iterativos (IWLS). Então, a técnica GEE encontra seu principal uso na análise multivariada de variáveis discretas. É então um desdobramento da classe de modelos lineares generalizados (GLM) em que se incorporam as correlações entre variáveis ou entre medidas repetidas. Pode ser aplicada a modelos de efeitos fixos e a modelos de efeitos mistos.

Uma diferença fundamental entre uma verossimilhança verdadeira e uma quase-verossimilhança abordada via equações de estimação é referente aos modelos de trabalho. Esses, no primeiro caso, tratam a verossimilhança como uma função objetivo para estimação e comparação de modelos. E no caso da quase-verossimilhança somente uma equação escore é especificada e resolvida para produzir uma estimativa. Essa abordagem da equação de estimação (EE) focaliza apenas o parâmetro de interesse e não toda a estrutura de probabilidade das observações. Uma vantagem da verossimilhança verdadeira refere-se à possibilidade de comparação de modelos via deviance e AIC. Uma abordagem alternativa de estimação associada à quase verossimilhança refere-se ao procedimento da pseudo-verossimilhança, o qual permite a comparação de modelos via LRT e AIC.

A análise de modelos lineares generalizados pode ser gerada via equações de estimação, via pseudo verossimilhança ou via REML ou IWLS (abordagem de verossimilhança verdadeira), mas as filosofias subjacentes são diferentes. Uma distinção essencial é que o teste da razão de verossimilhança não está disponível na abordagem EE (Resende, 2007).

A função objetivo denominada quase-verossimilhança apresenta duas características marcantes:

(i) Em contraste com a verossimilhança completa ou verdadeira, nenhuma estrutura de probabilidade é especificada, mas somente as funções da média e variância. Assim, essa abordagem pode ser denominada semi-paramétrica, em que os demais parâmetros, exceto aqueles de interesse, são deixados livres. Especificando apenas a média e a variância, a forma da distribuição permanece totalmente livre.

(ii) Com essa modelagem limitada, a amplitude de inferências possíveis é também limitada. Basicamente, apenas uma estimativa pontual do parâmetro é obtida. A construção de intervalos de confiança e a realização de testes de hipóteses assumem normalidade assintótica das estimativas, produzindo uma inferência do tipo Wald. Também, a comparação de modelos é limitada.

Modelos Lineares Mistos Generalizados Hierárquicos (HGLMM)

Nos modelos lineares mistos generalizados tradicionais assume-se que os resíduos podem não apresentar distribuição normal, mas, os demais efeitos aleatórios do modelo seguem a distribuição normal. Entretanto, essa suposição nem sempre é adequada. Um exemplo é a situação em que os dados seguem distribuição de Poisson e a função de ligação especificada para os resíduos é a logarítmica. Nesse caso, uma suposição mais apropriada para os demais fatores aleatórios é uma distribuição gama com função de ligação logarítmica. Modelos em que uma distribuição de probabilidade e uma função de ligação podem ser especificados para cada fator aleatório são denominados modelos lineares mistos generalizados hierárquicos (HGLMM). Como os fatores aleatórios nem sempre são de classificação hierárquica, uma denominação alternativa é modelos lineares mistos generalizados estratificados. HGLMM's são bem descritos por Lee et al. (2007). Um preditor BLUP para HGLMM's foi apresentado por Lee e Ha (2010). Para HGLMM's não normais o

BLUP linear pode não ser eficiente. Os autores apresentaram uma combinação do BLUP com modelos Tweedie de dispersão baseados em distribuição exponencial.

Verossimilhança Perfilada (PL)

A definição de verossimilhança contempla modelos multi-paramétricos. Entretanto, muitas vezes o interesse reside em apenas um subconjunto de parâmetros, sendo os demais denominados parâmetros de perturbação (*nuisance*) e participam do modelo apenas para ajudar a descrever melhor a variabilidade. Um caso típico é quando o interesse reside nos componentes de variância e os efeitos fixos são considerados *nuisance*. Nesse caso, é necessário um método para concentrar a verossimilhança em um só parâmetro ou grupo de parâmetros por meio da eliminação do parâmetro de *nuisance*.

A abordagem de verossimilhança para eliminar parâmetros de *nuisance* refere-se a substituir tais parâmetros por suas estimativas de máxima verossimilhança para cada valor fixo do parâmetro de interesse. A verossimilhança resultante é então denominada ***verossimilhança perfilada ou concentrada***. A abordagem bayesiana elimina todos os parâmetros não interessantes, integrando-os fora da distribuição. Entretanto, a função de verossimilhança não é uma função densidade de probabilidade (ou seja, não integra 1) e não obedece leis de probabilidade. Assim, integrar um parâmetro em uma função de verossimilhança não tem sentido. No entanto, existe uma analogia entre integração na abordagem bayesiana e o conceito de perfil de verossimilhança modificado relatado na seqüência.

Existe um método genérico de transformação de dados y para (v, w) de forma que a distribuição marginal de v e a distribuição condicional de v dado w depende apenas do parâmetro de interesse. Isso caracteriza o que é denominado ***verossimilhança marginal*** e ***verossimilhança condicional***, respectivamente. No entanto, verossimilhanças marginais e condicionais exatas nem sempre estão disponíveis ou são difíceis de derivar. Uma aproximação para essas pode ser obtida modificando-se o perfil de verossimilhança tradicional para se obter o perfil de verossimilhança modificado.

Não é possível usar o método REML ordinário para o modelo de competição por exemplo, uma vez que o coeficiente de competição aparece em ambos, na média e variância de y (pois ambos, tanto a variável quanto a covariável são o mesmo caráter). Entretanto, uma generalização do REML pode ser aplicada para estimação dos parâmetros do modelo. Essa generalização envolve o ajustamento da verossimilhança perfilada (por meio do escore perfilado ajustado) para o parâmetro de interesse em uma classe geral de modelos. Tal ajustamento pode ser feito pelo método de McCullagh e Tibshirani (1990), o qual remove o vício das estimativas de máxima verossimilhança, conforme realizado por Resende e Thompson (2003).

A inferência na presença de parâmetros de *nuisance* é um problema difícil em estatística. Sob a perspectiva da verossimilhança, a abordagem mais simples refere-se à eliminação (via maximização) dos referidos parâmetros para valores fixos dos parâmetros de interesse e então construir o que é denominado verossimilhança

perfilada. Em outras palavras, tal solução refere-se à substituição dos parâmetros de *nuisance* na função de verossimilhança por suas estimativas de máxima verossimilhança obtidas sob valores fixados dos parâmetros de interesse. Isto produz a verossimilhança perfilada. Essa é então tratada como uma função de verossimilhança ordinária para estimação e inferência sobre os parâmetros de interesse. Infelizmente, com grande número de parâmetros de *nuisance*, esse procedimento pode produzir estimativas ineficientes e inconsistentes. Os problemas inerentes ao uso de verossimilhanças perfiladas são a geração de estimativas viciadas dos parâmetros e otimistas dos desvios padrões.

Modificações na verossimilhança perfilada com o objetivo de aliviar esses problemas foram propostas. A verossimilhança perfilada modificada é intimamente relacionada à verossimilhança perfilada condicional na qual é sugerido um teste de razão de verossimilhança construído a partir da distribuição condicional das observações dadas as estimativas de máxima verossimilhança dos parâmetros de *nuisance*.

Máxima Parcimônia (MP)

Parcimônia é um princípio filosófico proposto pelo inglês William Ockam no século XIV e pode ser enunciado como: se existe mais de uma explicação para uma dado fenômeno, deve-se adotar aquela mais simples. O método de máxima parcimônia é muito empregado em análises de seqüências moleculares com o propósito de reconstrução de árvores filogenéticas como uma alternativa ao método de máxima verossimilhança.

O princípio da MP é que a hipótese mais simples deve ser a escolhida dentre todas as hipóteses possíveis de reconstrução filogenética. Em outras palavras, a árvore que apresentar o menor número de passos (mudanças de estado de caráter ou mutação) será a árvore mais parcimoniosa e deve ser escolhida para inferência.

Em termos estatísticos, esse princípio da simplificação de modelos indica que: modelos devem ter o mínimo possível de parâmetros; modelos lineares devem ser preferidos em relação aos não lineares; modelos baseados em poucas suposições devem ser preferidos em relação aos baseados em muitas suposições; modelos de simples explicação devem ser preferidos em relação aos de explicação complexa.

Einstein modificou ligeiramente o princípio de Occam e afirmou: um modelo deve ser tão simples quanto possível, mas não o mais simples. Também Oscar Wilde (escritor e poeta Irlandês) disse: a verdade é raramente pura, e nunca simples.

Quadrados Mínimos Parciais (PLS)

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. 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.

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.

Análise de Sobrevivência para Longevidade e Precocidade (SALP)

Na prática do melhoramento genético muitas vezes o caráter de interesse refere-se ao tempo ou número de dias ou meses para que determinado indivíduo atinja a produtidade ou peso desejável. Nesse caso, a seleção objetiva precocidade (menor tempo para atingir o valor desejável) ou longevidade (maior tempo ou vida produtiva). A seleção para longevidade e precocidade é de interesse em animais e fruteiras e para precocidade é interessante para espécies florestais.

Modelos de análise de sobrevivência para longevidade e precocidade (SALP) têm sido aplicados nessas espécies. Como o tempo é uma variável discreta e alguns indivíduos não atingem a produtividade desejada no período avaliado ou são descartados antes, os modelos usados em análise de sobrevivência para dados censurados têm sido empregados. O modelo em que os tempos t são independentes e seguem a distribuição de Weibull tem sido utilizados. Esse modelo é da forma $P(t|x_i, g_j) = \exp\{-g_j[\mu + b'x_i]t^{\gamma}$ em que: $P(t|x_i, g_j)$ é a probabilidade de um individuo *j* com vetor de efeitos fixos especificados por uma matriz de incidência X atingir a produtividade desejada após o o tempo *t*; u é uma constante; b é o vetor de coeficientes desconhecidos associados aos efeitos fixos x; g_j é o efeito genético aleatório associado ao genitor j; γ é o parâmetro de forma da distribuição Weibull.

Em termos de risco (λ) o modelo é dado por $\lambda(t|x_i, g_j) = g_j \exp[\mu + b'x_i]\gamma t^{\gamma-1}$. Risco no caso refere-se à propensão em atingir a produtividade desejada. Uma função de risco $\lambda(t|x_i, g_j)$ que cresce rapidamente e função de sobrevivência $P(t|x_i, g_j)$ que decresce rapidamente (menor tempo) através do tempo identifica um individuo precoce e interessante ao melhoramento.

Usando a denominação da área de análise de sobrevivência g_j é um efeito aleatório denominado fragilidade. Sob um modelo de sobrevivência Weibull com fragilidade gama tem-se que g_j segue uma distribuição gama com parâmetros (1 / Var(g), 1 / Var(g)), donde E(g)=1 e $Var(g) = I\sigma_g^2$. O modelo pode ser implementado via MCMC em que uma cadeia estocástica de valores dos parâmetros é assumida como contenedora de amostras da específica distribuição de probabilidade já em equilíbrio após períodos de descarte de amostras.

Outra abordagem aplicável nessa área são os modelos semiparamétricos como o modelo de riscos proporcionais de Cox.

1.18 Métodos Estatísticos para GWS

Os Métodos Estatísticos para GWS são apresentados na Tabela 18.

Tabela 18. 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
		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 Componentes Independentes	

Detalhes desses métodos são apresentados por Resende et al. (2011) e também no tópico 6.22. A seguir ilustra-se a questão dos métodos de regressão linear e não linear usando para isso o método 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-quadrada inversa e escalonada conforme a seguir:

$\beta_i \left| \sigma^2_{\beta i} \right. \sim N(0, \sigma^2_{\beta i})$

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

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

Assumido $\beta_i \sim N(0, \sigma^2_{\beta_i})$, em que $\sigma^2_{\beta_i}$ é tomado de uma distribuição qui-quadrado invertida, segundo o enfoque Bayesiano, 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 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. O método RR-BLUP são funções lineares dos dados e regressam as estimativas com o mesmo erro padrão (mesmas freqüências 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 distribuição a priori t escalada 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.

Em resumo, no modelo linear os efeitos de marcas são assumidos com distribuição normal e regressam as marcas de mesmas freqüências alélicas pela mesma quantidade. O modelo Bayesiano é não linear e os efeitos menores são regressados mais do que os maiores efeitos usando para isso informação a priori sobre a esperada distribuição dos efeitos de QTL (distribuição t no caso do BayesA).

1.19 Procedimento Estatístico para Comparação de Duas Metodologias

Conforme visto no tópico 1.1, um método ótimo de estimação/predição deve apresentar mínimo erro quadrático médio (EQM), o qual é dado por EQM = Vício² + Precisão = Vício² + PEV. Assim, um estimador de mínimo EQM apresenta vício nulo ou baixo e alta precisão (baixa variância do erro de predição – PEV ou $Var(\hat{g}-g)$). Em ausência de vício, EQM = PEV (Resende, 2008).

Algebricamente tem-se:

$EQM = E\left[(\hat{g}-g)^2\right] = E\left[\hat{g}^2\right] - 2E\left[\hat{g}\right]g + g^2$, em que g é tratada como uma constante determinística;

$PEV = E\left[\hat{g}^2\right] - E\left[\hat{g}\right]^2$;

$Vício^2 = b^2 = (E\left[\hat{g}\right] - g)^2 = EQM - PEV = E\left[\hat{g}\right]^2 - 2E\left[\hat{g}\right]g + g^2$;

$Vício = b = (E\left[\hat{g}\right] - g)$;

$Vício_i = b_i = \frac{1}{n}\sum_{j=1}^{n}(\hat{g}_{ij} - g_i)$ se pelo menos n = 2 repetições forem empregadas para a obtenção de $\hat{g}_i$.

O erro quadrático médio de estimação ou predição equivale à distância Euclideana média entre os estimadores e os correspondentes parâmetros. Minimizar o erro quadrático médio significa maximizar a acurácia. Um estimador acurado apresenta menor diferença quadrática entre valores verdadeiros (g) e estimados ($\hat{g}$). A acurácia ($r_{\hat{g}g}$) é definida como correlação entre g e $\hat{g}$ e seu quadrado ($r^2_{\hat{g}g}$) é um coeficiente de determinação denominado confiabilidade. O valor estimado equivale ao verdadeiro mais o erro de predição $(\hat{g}-g)$, ou seja, $\hat{g} = g + (\hat{g}-g)$.

A acurácia e a precisão guardam entre si as seguintes relações, na classe de estimadores não viesados:

- Acurácia ($r_{\hat{g}g}$)

$r_{\hat{g}g} = [1 - PEV / \sigma^2_g]^{1/2}$

- Precisão (PEV)

$PEV = Var(\hat{g}-g) = (1 - r^2_{\hat{g}g})\ \sigma^2_g$, em que σ^2_g é a variância de g.

Assim, o método ideal de estimação pode ser viciado em pequeno grau, pois o que importa é minimizar a soma $(Vício)^2 + PEV$. Na classe dos estimadores/preditores não viciados, a precisão é dada pelo parâmetro variância do erro de predição (PEV) e a estratégia de minimizar PEV conduz também à maximização da acurácia. Mas, de maneira geral (relaxando a necessidade de não vício), o que deve ser minimizado é o EQM, buscando a admissibilidade.

A comparação entre duas metodologias estatísticas ou dois vetores contendo variáveis quantitativas pode ser realizada por meio da comparação de seus EQMs e a

identidade entre as duas pode ser inferida com base na identidade de seus EQMs. Geralmente um modelo ($\hat{g}$) é comparado com a distribuição paramétrica (g). Mas, em muitas situações práticas, uma metodologia alternativa ($\hat{g}$) é comparada com uma metodologia padrão ou de referência (g) por meio de seu erro em relação à essa referência. Sendo $EQM = (Vício)^2 + PEV$, seus componentes devem ser testados estatisticamente visando inferir sobre a identidade entre duas metodologias. Uma abordagem para isso foi apresentada por Leite e Oliveira (2002). Esses autores propõem os seguintes testes para os três componentes:

(i) teste t para o erro médio $\bar{e} = \sum_{i=1}^{n}\left(\frac{(\hat{g}_i - g_i)}{g_i}\right)/n$ contra zero:

Estatística de teste: $t_{\bar{e}} = \left(\frac{(\bar{e}-0)}{s_{\bar{e}}}\right)$, em que $s_{\bar{e}} = s_e/\sqrt{n}$ e s_e é a estimativa do desvio padrão do erro, ou seja, $s\left(\frac{(\hat{g}_i - g_i)}{g_i}\right)$;

Hipótese H_o sob normalidade: $H_0 : \bar{e} = 0$.

Regra de Decisão: se $t_{\bar{e}} \geq t_{\alpha}(n-1)$, rejeita-se H_o, em que (n-1) são o número de graus de liberdade.

(ii) teste simultâneo de $\beta_0 = 0$ e $\beta_1 = 1$ para avaliar a significância do vício:

Segundo o modelo $\hat{g} = \beta_0 + \beta_1 g + e$, tem-se:

Estatística de teste: $F = \frac{(\beta-\theta)'(g^{*\prime}g^{*})(\beta-\theta)}{2QM\,\mathrm{Re}siduo}$, em que $\beta = \begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix}$; $\theta = \begin{bmatrix}0 \\ 1\end{bmatrix}$;

$g^* = \begin{bmatrix}1 & g_1 \\ 1 & g_2 \\ . & . \\ . & . \\ . & . \\ 1 & g_n\end{bmatrix}$ e $(g^{*\prime}g^{*}) = \begin{bmatrix}n & \sum g_i \\ \sum g_i & \sum g_i^2\end{bmatrix}$.

Hipótese H_o sob normalidade: $H_0 : \beta' = [0 \;\; 1]$.

Regra de Decisão: se $F(H_0) \geq F_{\alpha}(2, n-2)$, rejeita-se H_o, em que (n-2) são o número de graus de liberdade.

(iii) teste se a correlação difere de 1:

Sob $H_0 : \bar{e} = 0$ verdadeira, se $r_{\hat{g}g} \geq (1 - |\bar{e}|)$, a correlação entre as duas metodologias não difere de um. A correlação pode também ser testada pelo teste t diretamente contra 1. Mas isso envolveria a necessidade de um teste de hipótese a mais.

Um erro médio $\bar{e} = \sum_{i=1}^{n}\left(\frac{(\hat{g}_i - g_i)}{g_i}\right)/n$ pode não ser significativamente diferente de zero mas, o estimador pode ser viesado. Em estudos de simulação, o viés pode ser calculado pela expressão $(Vício)^2 = EQM - PEV$. Na prática, os valores paramétricos são desconhecidos mas, a significância do viés pode ser avaliada pelo teste simultâneo de $\beta_0 = 0$ e $\beta_1 = 1$ na expressão $\hat{g} = \beta_0 + \beta_1 g + e$. Mesmo com alta correlação ou determinação (baixa PEV devida ao alto $r^2_{\hat{g}g}$) é possível obter $\beta_0 \neq 0$ e $\beta_1 \neq 1$. $\beta_0 \neq 0$ indica diferença sistemática ou vício envolvendo dois vetores a serem comparados. $\beta_1 \neq 1$ indica erro ou diferença proporcional entre os dois vetores, conforme pode ser visto na expressão $\beta_1 = \frac{Cov(g,\hat{g})}{Var(g)} = r_{g\hat{g}}\sqrt{\frac{Var(\hat{g})}{Var(g)}}$, a qual revela que o coeficiente de regressão é função da correlação e também da diferença proporcional entre as variâncias associadas aos dois métodos.

Coeficientes de regressão abaixo de 1 indicam que os valores preditos são subestimados e apresentam variabilidade aquém da esperada e, acima de 1, indicam que os valores preditos apresentam variabilidade além da esperada. Coeficientes de regressão próximos de 1 indicam que as predições são não viesadas e são efetivas em predizer as reais magnitudes das diferenças entre os indivíduos em avaliação.

Não vício é importante quando se testa identidade entre modelos. Na classe dos estimadores/preditores não viciados não há necessidade de se testar $\beta_0 = 0$ e $\beta_1 = 1$. As regras de decisão são apresentadas no Quadro a seguir, conforme Leite e Oliveira (2002).

Regras de decisão na Comparação entre Duas Metodologias.

Situação	F(Ho): viés	$t_{\bar{e}}$: erro médio	$r_{\hat{g}g}$: componente da PEV	Decisão
1	não significativo	não significativo	$r_{\hat{g}g} \geq (1 - \lvert\bar{e}\rvert)$	$\hat{g} = g$
2	não significativo	não significativo	$r_{\hat{g}g} \leq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
3	não significativo	significativo	$r_{\hat{g}g} \geq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
4	não significativo	significativo	$r_{\hat{g}g} \leq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
5	significativo	não significativo	$r_{\hat{g}g} \geq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
6	significativo	não significativo	$r_{\hat{g}g} \leq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
7	significativo	significativo	$r_{\hat{g}g} \geq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$
8	significativo	significativo	$r_{\hat{g}g} \leq (1 - \lvert\bar{e}\rvert)$	$\hat{g} \neq g$

Fonte: Leite e Oliveira (2002).

Acurácia Seletiva via Inferência Bayesiana

No enfoque frequentista a acurácia é dada por $r_{\hat{g}g} = [1 - PEV / \sigma_g^2]^{1/2}$ em que PEV é relacionado à variância do estimador $\hat{g}$. Em inferência bayesiana computa-se a variância do próprio parâmetro que é assumido como uma variável aleatória. Assim, essa fórmula não é válida no contexto bayesiano. Propõe-se aqui usar a seguinte expressão para o cômputo da acurácia via estimação bayesiana: $\tilde{r}_{\hat{g}g} = [1 - s(\tilde{g}) / \tilde{g}]$, em que $s(\tilde{g})$ é o desvio padrão do valor genético estimado ($\tilde{g}$). Nota-se uma similaridade entre $r_{\hat{g}g}$ e $\tilde{r}_{\hat{g}g}$, sendo que $r_{\hat{g}g}$ envolve componentes quadráticos (PEV / σ_g^2, por isso existe a raiz quadrada na fórmula) e $\tilde{r}_{\hat{g}g}$ envolve componentes lineares ($s(\tilde{g}) / \tilde{g}$, por isso não existe a raiz quadrada na fórmula).

Em $r_{\hat{g}g}$ é computada a redução proporcional na correlação perfeita (igual a 1) dada pela razão entre a variação dos valores estimados em torno do valor verdadeiro (PEV) e a própria variação entre os valores verdadeiros (σ_g^2). Em $\tilde{r}_{\hat{g}g}$ a redução proporcional na correlação perfeita (igual a 1) é dada pela razão entre o erro padrão do valor verdadeiro realizado ($s(\tilde{g})$) e o próprio valor verdadeiro realizado ($\tilde{g}$). Se $s(\tilde{g})$ como proporção de $\tilde{g}$ tende a zero, a acurácia tende a 1. Se essa proporção afasta-se de zero, a acurácia afasta-se de 1. Por $\tilde{g}$ tratar-se do próprio parâmetro, deveria apresentar $s(\tilde{g})$ igual a zero. A medida que este afasta-se de zero penaliza-se $\tilde{r}_{\hat{g}g}$.

As duas abordagens podem ser comparadas por meio de $r_{\hat{g}g}$ e $\tilde{r}_{\hat{g}g}$. Se $\tilde{r}_{\hat{g}g} > r_{\hat{g}g}$, isso indica que provavelmente as distribuições dos parâmetros atribuídas pela abordagem Bayesiana foram mais adequadas do que aquelas associadas ao modelo tradicional.

1.20 Procedimento BLUP Melhorado: I-BAYES-BLUP

O BLUP tradicional é adequado quando: não existem genes maiores segregando na população; uma população base ideal foi formada (em equilíbrio, com endogamia F = o e sem indivíduos aparentados); toda a genealogia e conjunto completo de dados são usados, desde a população base; não existem erros no pedigree. Se a população base ideal não foi formada e/ou o pedigree não está completo e livre de erros e/ou nem todo o conjunto de dados é usado, surgem problemas tais quais: a variação em F entre genitores (os quais diferem em heterozigose) e desequilíbrio de ligação em fase gamética dentro de família não são levados em consideração; as predições de valores genéticos obtidas são viesadas pelos efeitos da seleção; os componentes de variância da população base são pobremente estimadas; o parentesco entre os indivíduos não é estimado corretamente. Segundo Endelman e Jannink (2012), as suposições sobre a população base ideal raramente se verificam na prática.

Um método para a estimação de componentes de variância e valores genéticos delineado para aumentar a eficiência do REML/BLUP fenotípico foi introduzido por Resende, Silva e Viana (2012). O método é denominado I-BAYES-BLUP (Improved Bayesian BLUP) ou BBM (BLUP Bayesiano Melhorado) e visa capturar os diferentes graus de variação dentro de famílias da geração atual e a correlação genética entre famílias, devidos à esses fatores.

Tais fatores produzem diferentes níveis de variação nas relações de parentesco entre pares de indivíduos dentro de diferentes famílias e então diferentes parentescos médios dentro de cada família. A captura dessa variabilidade possibilita a estimação da variação genética contribuída especificamente por cada família da geração atual e propicia uma melhor partição da variabilidade genética entre e dentro de famílias, permitindo estimar uma variação genética específica para cada família. Como consequência, uma melhor estimativa do componente do valor genético, denominado efeito da segregação mendeliana, é obtida.

O procedimento é geral e equivale ao próprio BLUP tradicional quando não existe heterogeneidade de variância genetica dentro de famílias. Então, é recomendado para uso amplo, quando os tamanhos de família são grandes o suficiente (no mínimo 10 indivíduos por família) para obtenção de estimativas precisas. O método é superior quando ocorre pelo menos um dos seguintes fatos: o modelo infinitesimal (genes de iguais efeitos) não se aplica; o pedigree é incompleto e/ou com presença de erros; uma população base ideal não foi formada; o conjunto de dados é incompleto (não contempla todas as medições desde a população base); existe grau diferencial de variância genetica dentro de família.

O método proposto ameniza o caso do uso do BLUP considerando apenas dados da geração atual e genealogia contemplando apenas as duas últimas gerações, por meio do uso da identidade em estado (IBS) e não por descendência (IBD). O procedimento considera também (pelo menos em parte) a variação no sistema reprodutivo entre genitores (sistema misto de reprodução, envolvendo simultaneamente autogamia e alogamia).

O procedimento envolve os seguintes passos: (1) ajuste apenas dos efeitos de genitores (aqueles indivíduos com progênie) como covariáveis aleatórias, de maneira similar ao modelo animal reduzido, entretanto, usando uma abordagem Bayesiana que admite variancias genéticas aditivas específicas (σ^2_{gi}) para cada família i; (2) cálculo das variancias genéticas aditivas dentro de família ($\hat{\sigma}^2_{gwi} = (1-\hat{\rho}_{gi})\sigma^2_{gi}$), específicas para cada família i; (3) cálculo das variancias fenotípicas dentro de família ($\hat{\sigma}^2_{ywi} = \hat{\sigma}^2_{gwi} + \sigma^2_e$), específicas para cada família i; (4) cálculo das herdabilidades individuais dentro de família ($h^2_{wi} = \frac{\hat{\sigma}^2_{gwi}}{\hat{\sigma}^2_{ywi}}$), específicas para cada família i; (5) estimação do efeito da segregação mendeliana usando uma formula específica para cada família i ($\hat{g}_{wij} = (y_{ij} - X\hat{b}_{ij} - 0.5\hat{g}_{pi})h^2_{wi}$), em que b é um vetor de efeitos fixos; (6) soma dos efeitos dos genitores ($\hat{g}_{pi}$) com os efeitos da segregação mendeliana de cada individuo por meio de $\hat{g}_{ij} = 0.5\hat{g}_{pi} + \hat{g}_{wij} = 0.5\hat{g}_{pi} + (y_{ij} - X\hat{b}_{ij} - 0.5\hat{g}_{pi})h^2_{wi}$.

O método é um modelo animal reduzido, melhorado por meio de uma combinação ou mistura das abordagens Bayesiana e BLUP tradicional. Pelo BLUP tem-se a estrutura de covariância para os efeitos genéticos g: $g \sim N(0,\Sigma) = g \sim N(0, A\sigma^2_g)$. Pela abordagem Bayesiana tem-se $g \sim N(0, G_{BAYES})$, em que $\Sigma = G_{BAYES}$ é estimada diretamente como uma matriz de covariância não estruturada, contemplando a heterogeneidade de variância genética dentro de família (e diferentes endogamias F dos genitores) em sua diagonal e o parentesco entre os genitores fora da diagonal. Alternativamente, G_{BAYES} pode ser ajustada como uma matriz diagonal $G_{BAYES} = G_{BAYES-Diag}$, usando a matriz de parentesco entre os genitores (A) simultaneamente via modelagem de g como $g \sim N(0, A\, G_{BAYES})$. As variâncias e covariâncias genéticas componentes de G_{BAYES} são assumidas como provenientes de uma distribuição Whishart e estimadas por meio do pacote bayesm do R via função rhierLinearModel (Rossi et al., 2005; 2012).

O I-BAYES-BLUP modela uma estrutura de variância similar a $\Sigma = A\sigma^2_{gi}$, porém, usa a identidade em estado (IBS), de forma que é semelhante a $\Sigma = G\sigma^2_{gi}$, em que G é a matriz de parentesco genômico baseada em IBS e não em IBD como a A. Assim, G_{BAYES} se aproxima de $G_{BAYES} = G\sigma^2_{gi}$, tendendo a captar intrinsecamente IBS, via uma regressão implícita, porém paramétrica. Essa equivalência é razoável uma vez que a matriz de parentesco IBS usa a população corrente como população base, ou seja, está associada à estimação da variação genética na população corrente (Endelman e Jannink, 2012).

Conforme Powell et al. (2010) e Endelman e Jannink (2012), a meta do geneticista não é estimar probabilidades IBD e sim estimar covariância genética entre indivíduos, a qual é fundamentalmente uma propriedade de estado (IBS) e não IBD. Assim, Σ depende de probabilidades IBS, as quais não invocam uma população base ideal. Dessa forma, o método I-BAYES-BLUP é uma boa alternativa para fazer uso desse novo conceito, usando apenas dados fenotípicos.

2 Análise Genômica

Esforços na área de pesquisa com marcadores genéticos em prol do melhoramento têm se dividido em duas linhas: a detecção de marcadores associados a QTL e mapeamento desses; uso dos marcadores nos programas de seleção genética via seleção auxiliada por marcadores (MAS) e seleção genômica ampla (GWS), também denominada seleção genômica (GS). Esse capítulo aborda ambas as linhas, enfatizando a seleção genética via uso das informações genômicas. Contempla o estudo de QTLs baseado em análise de análise de ligação e de desequilíbrio de ligação (genética de associação - GWAS) e a seleção genômica ampla (GWS). A abordagem apresentada nesse documento baseia-se no livro publicado por Resende (2008).

2.1 Fundamentos da Análise de QTLs e da Seleção Genômica

O uso de marcadores genéticos moleculares para fins de seleção e melhoramento genético fundamenta-se na ligação gênica entre tais marcadores e os locos que governam as características quantitativas (QTLs) de interesse do melhoramento. Assim, os estudos de ligação entre marcador e QTL e também entre os próprios marcadores são essenciais no contexto da seleção genética empregando-se informações genômicas. É importante relatar que a definição de QTL refere-se apenas a uma associação estatística entre uma região do genoma e um caráter.

No contexto da genética clássica, a ligação entre fatores genéticos ou genes tem sido relatada desde 1906, e denota que genes ligados proximamente no cromossomo são herdados em conjunto. Em outras palavras, tais genes, em conjunto, não segregam de forma independente, não obedecendo a Segunda Lei de Mendel ou Lei da Segregação Independente. Quando os genes estão próximos no cromossomo ou grupo de ligação, essa é completa. Quando estão no mesmo grupo de ligação, porém com grande distância entre eles, a ligação é parcial.

A distância calculada entre dois genes é função da freqüência de recombinação entre eles e é fundamental na construção de mapas de ligação. Para que a ligação entre locos seja detectada e usada na seleção, é necessário que haja desequilíbrio de ligação na população ou família estudada. O desequilíbrio de ligação ou desequilíbrio de fase gamética é uma medida da dependência ou não entre alelos de dois ou mais locos. Em um grupo de indivíduos, se dois alelos de locos diferentes são encontrados juntos com freqüência maior do que aquela esperada com base no produto de suas freqüências, infere-se que tais alelos estão em desequilíbrio de ligação. Valores de desequilíbrio de ligação próximos de zero indicam equilíbrio ou independência (freqüência de recombinação igual a 0,5, ou seja, com valor máximo) entre os alelos de diferentes genes, e valores próximos de um, indicam desequilíbrio ou ligação entre alelos de diferentes genes.

O desequilíbrio de ligação (LD) entre marcadores e QTLs é crucial para a detecção de QTL, para a seleção auxiliada por marcadores e para a seleção genômica ampla. Especialmente relevante é a extensão desse desequilíbrio no cromossomo em uma população de seleção. Se um marcador e um QTL estão em equilíbrio na população, o conhecimento do genótipo do marcador em um indivíduo não apresenta

qualquer valor para a seleção. A permanência do desequilíbrio de ligação na população depende da distância entre os locos, ou seja, depende da taxa de recombinação entre os dois locos. Para locos intimamente ligados, qualquer LD que tenha sido criado permanecerá por muitas gerações. Mas, para locos fracamente (freqüência de recombinação maior que 0,1) ligados, o LD decrescerá rapidamente. Embora um marcador (loco m) e QTL (loco q) ligado a ele possam estar em equilíbrio de ligação na população, sempre existirá o desequilíbrio de ligação dentro de uma família ou cruzamento, mesmo para locos fracamente ligados. E esse desequilíbrio de ligação poderá se estender a uma grande distância, pois, para a produção da descendência de um indivíduo F1 heterozigoto, terá ocorrido apenas uma geração de recombinação.

Considere dois locos m e q ligados e quatro indivíduos heterozigotos para o marcador e com genótipos MQ/mq, Mq/mQ, MQ/mQ e Mq/mq. As famílias originárias dos dois primeiros indivíduos estarão em LD (pois, para locos ligados, gametas parentais são mais freqüentes que gametas recombinantes), porém em direções opostas, pois, a fase de ligação marcador-QTL difere entre os dois genitores. As famílias originárias dos dois últimos indivíduos não estarão em LD pois o QTL não está segregando nessas famílias. Quando ponderados entre famílias, os quatro tipos de desequilíbrio cancelarão, produzindo equilíbrio de ligação na população. Assim, o LD dentro de famílias é útil na análise de QTL desde que as diferentes fases de ligação sejam levadas em consideração.

De maneira genérica, em Genética de Populações, desequilíbrio refere-se à discrepância da freqüência conjunta de alelos em relação ao produto de suas freqüências individuais. O termo usualmente refere-se a alelos de diferentes locos em um mesmo gameta, mas pode referir-se também a pares de alelos do mesmo loco, caracterizando a falta de equilíbrio de Hardy-Weinberg.

O mapeamento de QTLs, a seleção auxiliada por marcadores (MAS) proposta por Lande & Thompson (1990) e a seleção genômica ampla (GWS) proposta por Meuwissen et al. (2001), são fundamentadas na ocorrência de desequilíbrio de ligação na população (ou cruzamento) estudada. Nesse caso, os alelos dos marcadores informam sobre a presença e efeitos dos locos que governam os caracteres quantitativos, fornecendo meios para estimação dos efeitos dos locos dos QTLs e para o seu eficiente uso na seleção genética. As causas do desequilíbrio de ligação nas populações são: mutação, migração, seleção e tamanho efetivo populacional reduzido (deriva genética devida à amostragem). Ou seja, todos os fatores que afetam o equilíbrio de Hardy-Weinberg nas populações afetam também o equilíbrio de ligação.

Atualmente, marcadores genéticos moleculares do tipo SNP (polimorfismo de um único nucleotídeo), os quais baseiam-se na detecção de polimorfismos resultantes da alteração de uma única base no genoma, têm sido usados. E para que uma variação seja considerada SNP, essa deve ocorrer em pelo menos 1 % da população. Os SNPs são a forma mais abundante de variação do DNA em genomas, e são preferidos em relação a outros marcadores genéticos devido à sua baixa taxa de mutação e facilidade de genotipagem, aliados ao baixo custo. Milhares de SNPs

podem ser usados para cobrir o genoma de um organismo com marcadores que não estão a mais de 1 cM (1 milhão de bases) um do outro no genoma inteiro. Os marcadores moleculares do tipo microssatélites também são usados. Tais marcadores são eficientes por serem co-dominantes, multi-alélicos, abundantes e apresentarem alta transferibilidade entre indivíduos e espécies.

Os marcadores SNPs apresentam natureza bialélica, conforme ilustrado a seguir:

Indivíduo 1: TCA*C*CGCG

Indivíduo 2: TCA*T*CGCG

Verifica-se polimorfismo de SNPs entre os dois indivíduos. Na seqüência especificada na fita simples de DNA ocorre troca de uma única base, caracterizando o referido polimorfismo. Mais de 1,5 milhão de SNPs foram identificados no genoma humano. Suas posições estão localizadas em um espaçamento médio de 2×10^{-3} cM (Hartl e Jones, 2002).

Marcadores DArT (Diversity Array Technology) são também bi-alélicos e adequados à GWS pois são abundantes tais quais os SNPs, e podem ser obtidos com alta velocidade e rendimento. No entanto, tais marcadores são dominantes e essa pode ser uma desvantagem em relação aos SNPs, que são codominantes. Entretanto, podem comportar-se de duas maneiras: dominante (presença vs ausência) ou codominante (2 doses vs 1 dose vs ausência).

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

Esses efeitos genéticos dos marcadores sobre fenótipos de caracteres quantitativos são somados e usados na predição de valores genéticos de indivíduos apenas genotipados, candidatos à seleção em programas de melhoramento genético. A predição e a seleção podem ser realizadas em fases muito juvenis de plantas e animais, acelerando assim o processo de melhoramento genético. Adicionalmente, a própria predição tende a ser mais acurada por considerar o real parentesco genético dos indivíduos em avaliação, em detrimento do parentesco médio esperado matematicamente (Resende, 2007). A GWS propicia uma forma de seleção precoce direta (SPD), pois, atua precocemente sobre genes expressos na idade adulta. Ao contrário a seleção precoce tradicional é indireta, pois, atua (via avaliação fenotípica)

sobre genes ativados na idade precoce, esperando que esses informem parcialmente sobre genes expressos na idade adulta. Assim, a SPD propiciada pela GWS é especialmente importante para o melhoramento de organismos perenes como animais, espécies florestais, fruteiras (e outras frutíferas), forrageiras, cana-de-açúcar, café, dentre outras.

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

A MAS surgiu basicamente na década de 1990. Os primeiros trabalhos relativos a organismos perenes foram os de Fernando e Grossman (1989), Lande e Thompson (1990), Goddard (1991) e Kennedy et al. (1992). A GWS é um produto do terceiro milênio. Após a proposição da GWS em 2001 o procedimento permaneceu discreto até 2007, quando vários trabalhos abordaram o método e sua acurácia no melhoramento animal e vegetal (Fernando et al., 2007; Goddard e Hayes, 2007; Meuwissen, 2007; Bernardo e Yu, 2007; Resende 2007). Outros trabalhos relatam que a GWS é o novo paradigma em genética quantitativa (Resende, 2008; Gianola et al., 2009), melhoramento de gado de leite (Hayes et al., 2009; Van Raden, 2008; VanRaden et al., 2009), de corte (Ferraz e Rezende, 2011), de aves (Gonzales-Recio et al., 2009), de plantas anuais (Heffner et al., 2009) e de espécies florestais (Resende et al. 2008; Grattapaglia e Resende, 2011).

Atualmente resultados práticos já existem para eucalipto (Resende et al., 2012), pinus (Resende Júnior et al., 2012), suínos (Rocha et al., 2012; Azevedo et al., 2012), milho (Fritsche Neto et al., 2012) e caju (Cavalcanti et al., 2012). Acredita-se que a GWS propiciará um impacto positivo nos métodos de seleção e nas estratégias de melhoramento de plantas e animais. No entanto, é preciso adquirir experiência prática com a GWS, visando inferir sobre sua efetividade.

2.2 Análise de Ligação (LA) e Análise de Desequilíbrio de Ligação (LDA)

A quantidade de material genético herdável de um indivíduo é finita e refere-se ao tamanho do genoma. Em humanos, o genoma é composto de cerca de 35 mil genes (Ewing e Green, 2000). Assim, um número finito de genes deve controlar cada um dos caracteres quantitativos e isso torna possível a avaliação de todos os locos associados ao controle genético de um caráter.

Existem basicamente três abordagens para a descoberta de um QTL: (i) abordagem de genes candidatos; (ii) abordagem de mapeamento via análise de ligação ou linkage analysis (LA); (iii) abordagem de mapeamento via análise de desequilíbrio de ligação ou *linkage disequilibrium analysis* (LDA). A estratégia de genes

candidatos considera que um gene envolvido na fisiologia do caráter abriga uma mutação causadora de variação no caráter. Esse gene é então seqüenciado em diferentes indivíduos e as variações encontradas nas seqüências de DNA são avaliadas em termos de associação com variações encontradas nos fenótipos do caráter (Anderson e Georges, 2004). Essa abordagem apresenta problemas tais quais o grande número possível de genes candidatos e a possibilidade de que a mutação causadora da variação esteja em um gene não tomado a priori como candidato.

As abordagens de mapeamento visam identificar regiões cromossômicas associadas a variações fenotípicas nos caracteres de interesse, e assumem que os genes não são conhecidos mas apenas marcados por genes de efeitos nulos. Baseiam-se então em associações entre alelos dos genes marcadores e variações nos caracteres quantitativos. Um marcador molecular de DNA é uma região física identificável no cromossomo cuja herança pode ser monitorada e que geralmente não apresenta função codificadora.

Um marcador é considerado informativo quando se pode determinar sem erro, qual alelo parental foi transmitido para a progênie. Assim, se um genitor genotipado é homozigoto para o marcador, este não será informativo em qualquer dos indivíduos da progênie, pois não será possível determinar qual alelo parental foi transmitido. Mesmo se ambos, genitor e progênie, são heterozigotos, o marcador pode ainda ser não informativo. Se somente um genitor é genotipado, e a progênie tem o mesmo genótipo que seu genitor, a progênie pode ter recebido determinado alelo do pai ou da mãe. A freqüência esperada de indivíduos para os quais a origem do alelo pode ser determinada será 1 – (p + q)/2, em que p e q são as freqüências dos dois alelos marcadores parentais. Assim, se somente dois alelos marcadores estão presentes na população, metade dos filhos terão o mesmo genótipo que o genitor. Para locos multi-alélicos como os microssatélites, (p + q) pode ser muito menor do que 1 (Weller, 2001).

A estratégia da análise de ligação (LA) considera apenas o desequilíbrio de ligação que existe dentro de famílias ou cruzamentos, que estende-se por dezenas de cM e é quebrado por recombinação após algumas poucas gerações. Essa abordagem usa um limitado número de marcadores por cromossomo e, então, devido à recombinação entre distantes marcador e QTL, a associação entre marcadores e QTLs permanecerá apenas dentro de famílias e por um limitado número de gerações. Essa estratégia conduz ao mapeamento de QTL em um grande intervalo de confiança no cromossomo, exceto se um enorme número de indivíduos por família for usado. A fórmula de Darvasi e Soller (1997) pode ilustrar isso. No caso de um mapa genético de alta densidade, o intervalo de confiança (IC) é dado por IC = 3000/(kns^2), em que k é o número de genitores informativos por indivíduo (1 para famílias de meios irmãos e 2 para famílias de irmãos germanos e populações *F2*), *n* é o número de indivíduos genotipados, *s* é o efeito de substituição alélica associado ao alelo favorável do QTL e 3.000 cM é o tamanho do genoma de gado bovino (nessa espécie cada cM contempla aproximadamente 8 genes).

Com base nessa expressão e considerando um QTL segregante com *s* igual a 0,5 desvios padrões residuais, em uma família de meios irmãos de 1,000 indivíduos, tem-

se que o IC a 95 % de probabilidade será de 12 cM. Os reflexos desse grande IC são: (i) se o objetivo for a adoção da abordagem de genes candidatos dentro desse intervalo, um grande número de genes deve ser seqüenciado e estudado (80 genes considerando um total de 20 mil genes em um genoma de 3.000 cM); (ii) se o objetivo for a MAS, a ligação entre marcador e QTL não é suficientemente próxima para garantir que a associação marcador-QTL persista através de toda uma população e, nesse caso, a fase de ligação marcador-QTL dentro de cada família deve ser estabelecida para aplicação da MAS (Hayes, 2008). Por exemplo, um indivíduo da população poderá apresentar o alelo M do marcador associado ao alelo favorável do QTL e outro indivíduo, da mesma população, mas de família diferente, poderá apresentar o alelo *m* do marcador associado a esse mesmo alelo favorável do QTL.

A abordagem LA baseia-se na associação entre alelos do marcador e classes fenotípicas do QTL e foi muito usada até recentemente devido ao fato de que o número de marcadores identificados nas várias espécies era baixo e o custo de genotipagem muito alto. Com o recente advento dos marcadores SNPs, os quais são em grande número e baratos, uma alta densidade de marcadores no genoma tornou-se possível e a marcação próxima dos próprios QTLs também. Nesse caso, a adoção da abordagem LDA tornou-se possível e vantajosa sobre a LA.

A estratégia LDA baseia-se no desequilíbrio de ligação entre marcador e QTL na população inteira e não apenas dentro de família como na LA. Para que isso ocorra, marcador e QTL devem estar em ligação muito próxima. E, nesse caso, a associação entre eles é uma propriedade da população inteira e persistirá por um grande número de gerações.

Meuwissen e Goddard (2000) revelaram que o intervalo de confiança poderia ser reduzido para 1 cM pela aplicação do mapeamento via LDA. Se o polimorfismo de um QTL é devido a uma mutação recente ou devido a uma recente introdução de uma outra população, então torna-se possível detectar LD em nível populacional entre QTL e genes marcadores proximamente ligados. Quanto mais perto o marcador do QTL, maior será o desequilíbrio de ligação. O intervalo de confiança pode ser reduzido ainda mais pela combinação das estratégias de análise LA e LDA e por uma análise multi-característica (Meuwissen e Goddard, 2004).

A análise de associação é usada no mapeamento fino e fundamenta-se no desequilíbrio de ligação em nível populacional. A associação pode ocorrer em duas situações: (i) devida ao efeito direto do gene em uma característica; (ii) devida ao desequilíbrio de ligação entre o marcador e o gene que controla a característica. No primeiro caso, o efeito do gene é medido diretamente e o marcador é funcional. No segundo caso, o teste de associação requer o desequilíbrio de ligacão entre o marcador e o QTL. Quando uma mutação ocorre no cromossomo, forma-se uma combinação haplotípica com os locos adjascentes no cromossomo. Na geração seguinte existe a tendência de que essa mutação ocorra no mesmo haplótipo original, a menos que ocorra recombinação. Isso caracteriza o desequilíbrio de ligação usado no mapeamento de associação.

3 Análise de QTL e da Expressão Gênica

3.1 Métodos de Análise de QTL

QTL (*quantitative trait loci*) são locos ou segmentos cromossômicos que governam as características quantitativas, mas essa definição refere-se apenas a uma associação estatística entre uma região do genoma e um caráter fenotípico. Marcadores genéticos em ligação próxima com QTL são usados para mapea-los e também para a seleção auxiliada por marcadores (MAS) em conjunto com informações fenotípicas. A disponibilidade de marcadores moleculares foi aumentada recentemente com o advento dos microssatélites, dos SNPs e dos DArTs e os genes a eles ligados podem ser mapeados em grupos de ligação.

Os procedimentos de mapeamento são baseados no desequilíbrio de ligação entre alelos de diferentes genes. Mapas de ligação entre marcadores polimórficos cobrindo todo o genoma são necessários no mapeamento de QTLs. Tais mapas estão agora disponíveis para um grande número de organismos e informações desses mapas juntamente com medidas fenotípicas obtidas de acordo com algum delineamento de cruzamento e experimental são usados para mapear e estimar efeitos de QTLs. O mapeamento de QTLs envolve a detecção, localização (determinação da posição) e estimação dos efeitos de QTLs.

Diferentes abordagens estatísticas são usadas no mapeamento de QTLs, dependendo da estrutura da população de mapeamento e do número (densidade) e tipo de marcadores usados. Com limitado número de marcadores por cromossomo e desequilíbrio de ligação apenas dentro de famílias ou cruzamentos, a estratégia da análise de ligação (LA) deve ser usada. Com grande número e alta densidade de marcadores no genoma torna-se possível a marcação mais próxima dos QTLs e a abordagem LDA (análise de desequilíbrio de ligação) deve ser usada. Nesse caso, a LDA tornou-se possível e é vantajosa sobre a LA.

Para a LA, em plantas anuais, cruzamentos entre linhagens endogâmicas são geralmente realizados e análises são conduzidas nas populações das gerações F_2, F_3, retrocruzamentos e de haplóides duplicados. Em plantas perenes tais quais espécies florestais, fruteiras, forrageiras e cana-de-açúcar, famílias de irmãos completos ou grandes famílias de meios irmãos, obtidas do cruzamento entre indivíduos heterozigotos, são usadas. Em humanos e em animais domésticos, além das referidas famílias, populações associadas a pedigrees complexos e multi-gerações são também usados nos estudos de QTLs. Em cada caso, as referidas populações são fenotipadas e pelo menos as populações e os genitores são genotipados, ou seja, a genotipagem envolve pelo menos duas gerações.

O mapeamento de QTL envolve quatro etapas: escolha da população de mapeamento; obtenção dos dados de marcadores em cada individuo; obtenção dos dados fenotípicos em cada individuo; aplicação de métodos estatísticos na analise simultânea dos dados fenotípicos e de marcadores. Indivíduos pertencentes a essas populações de mapeamento são genotipados para um número de marcadores moleculares distribuídos a intervalos regulares no genoma e avaliados para os

caracteres quantitativos de interesse. Se existirem diferenças significativas nas médias fenotípicas entre classes genotípicas de um marcador, pode-se inferir que existe um QTL ligado àquele marcador. A associação entre QTL e marcador pode ser avaliada usando um, dois ou vários marcadores simultaneamente.

Dentre os métodos gerais de análise e mapeamento de QTLs destacam-se: (i) a análise de marcas únicas (um marcador de cada vez), que é útil quando o objetivo é somente a detecção de QTL ligado ao marcador, mas não a estimação da posição e dos efeitos do QTL; (ii) o mapeamento por intervalo simples, proposto por Lander e Botstein (1989), que considera marcadores adjacentes e então propicia um aumento no poder de detecção e estimativas mais precisas da posição e efeitos dos QTLs; (iii) o mapeamento por intervalo composto apresentado por Zeng (1994), que considera vários marcadores simultaneamente e é uma abordagem ainda melhor quando múltiplos QTLs estão ligados no intervalo ou marcadores considerados; (iv) mapeamento por intervalos múltiplos, que considera vários QTLs simultaneamente e permite incluir os efeitos epistáticos no modelo.

Os métodos baseados em intervalo são superiores pois a análise de marcas simples apresenta duas grandes limitações: (i) o confundimento dos efeitos de um QTL com os de outros QTLs que influenciam o mesmo caráter; (ii) a não distinção entre um QTL de grande efeito mas em ligação distante com o marcador, de um QTL de pequeno efeito mas em ligação próxima com o marcador. Por essa abordagem, a localização do QTL em relação ao marcador não pode ser determinada pois a frequência de recombinação é confundida com o efeito genético. Os métodos de mapeamento por intervalo demandam mapas de ligação entre marcadores polimórficos cobrindo todo o genoma e permitem a verificação da presença de QTL em cada intervalo, determinado por dois marcadores flanqueadores. Para que um QTL se separe de dois marcadores flanqueadores são necessários dois eventos de recombinação, fato que é mais raro. Assim, o uso do intervalo conduz a melhores resultados.

Quanto aos métodos de estimação, a análise de marcas únicas pode ser feita usando métodos estatísticos comuns tais quais a estatística t de Student, regressão linear simples, análise de variância e máxima verossimilhança (LOD escore). Para a análise de QTL baseada em dois marcadores flanqueadores (mapeamento por intervalo), os principais métodos usados são o método de regressão proposto por Haley e Knott (1992) e o método de máxima verossimilhança proposto por Lander e Botstein (1989).

Se um marcador apresenta um efeito significativo sobre um QTL, a diferença entre as médias das classes dos genótipos marcadores para o referido caráter é um estimador viciado do efeito do QTL, devido à possível recombinação entre o marcador e o QTL. Weller (1986) mostrou que o método de máxima verossimilhança poderia ser usado para obter estimativas da posição e efeito do QTL não viciadas pela recombinação.

O mapeamento por intervalo composto usa vários marcadores simultaneamente e também ambos os métodos, máxima verossimilhança e regressão. A seleção de marcadores a serem incluídos na regressão é baseada nos procedimentos *stepwise*. Outro método é a máxima verossimilhança residual (REML) baseada em um modelo linear misto incorporando efeitos alélicos do QTL com distribuição normal e com uma matriz de covariância condicional aos dados observados dos marcadores. Métodos bayesianos são também usados.

As abordagens estatísticas para análise de QTL diferem em relação às suposições de efeitos fixos ou aleatórios de QTL. Alguns métodos assumem o QTL como efeito fixo e com número finito de alelos (geralmente 2). Outros o assumem como efeito aleatório com um infinito número de alelos. Os métodos estatísticos que tratam o QTL com número finito de alelos variam desde modelos simples de regressão (Knott et al., 1996) a abordagens Bayesianas. Os modelos estatísticos de efeitos fixos são misturas de distribuições, em que o número de densidades componentes é determinado pelo número de genótipos do QTL. As suposições relativas ao número de alelos segregantes tem um grande efeito na formulação do modelo estatístico (George et al., 2000). Modelos de efeitos aleatórios, baseados na simples premissa de que indivíduos com fenótipos parecidos provavelmente compartilham alelos idênticos por descendência, oferecem uma abordagem menos parametrizada para o mapeamento.

Weller (2001) relata a simulação de um genoma com 100 locos e o uso dos 20 com maiores efeitos em um programa de seleção assistida por marcadores. O ganho com seleção mostrou-se o dobro quando os efeitos de QTL foram tratados como aleatórios, em relação à situação em que foram tratados como fixos. Embora os modelos aleatórios assumam um número infinito de possíveis alelos do QTL, as estimativas das variâncias dos QTLs são robustas à desvios dessa suposição e estimativas fidedignas podem ser obtidas mesmo quando apenas dois alelos por QTL são simulados.

Antes da análise de QTL propriamente dita, uma análise criteriosa dos marcadores deve ser realizada. Assim, deve ser realizada uma análise de segregação de marcas, verificando se a proporção de segregação esperada (3:1 em F2 e 1:1 em retrocruzamentos, por exemplo) se concretiza. Nesse caso, verifica-se se existe distorção de segregação e, em caso positivo, esses marcadores devem ser descartados da subseqüente análise de QTL. Também, os dados fenotípicos devem ser analisados previamente quanto à normalidade. Assim, os seguintes passos devem ser adotados em um estudo de QTL: avaliação de uma população segregante para o caráter; análise de DNA por uma técnica de marcadores (microssatélites ou SNPs); análise de segregação de marcas; análise de QTL ou de co-segregação entre marcador e QTL. A análise de co-segregação permite o estabelecimento de grupos de ligação de acordo com a porcentagem de recombinação entre os vários locos.

Em um procedimento de mapeamento de QTL, inicialmente, análises de marcadores únicos são realizadas por meio de métodos estatísticos simples como a ANOVA, a ANOVA não paramétrica de Kruskal-Wallis, a estatística t de Student, a regressão linear simples, a máxima verossimilhança (LOD score). Estes

procedimentos permitem a detecção de associação entre os marcadores e o caráter de interesse, sem usar informação de mapa genético. Isto é feito para cada marcador, contrastando as observações fenotípicas entre as classes de cada marcador. Tais classes são tomadas como se fossem tratamentos a serem comparados. Posteriormente, o mapeamento por intervalo (Lander e Botstein, 1989), considerando dois marcadores, pode ser feito visando à seleção de marcadores a serem usados como potenciais cofatores em uma análise de regressão múltipla do tipo *stepwise*. Também, o mapeamento por intervalo composto pode ser efetuado quando múltiplos QTLs estão ligados ao intervalo ou marcador considerados.

Em geral, os procedimentos de mapeamento têm usado diretamente os dados de campo para análise. Tais dados, em conjunto com a informação molecular são usados nos *softwares* padrões para mapeamento de QTL tais quais o MapMaker-QTL (Lander e Botstein, 1989). Ou seja, não são rotineiramente usados valores genéticos preditos após a eliminação dos efeitos ambientais. Entretanto, é recomendável que o mapeamento seja baseado em valores genéticos preditos sob um modelo que contemple também os efeitos ambientais de escala global (locais, blocos), os efeitos ambientais de escala localizada (resíduo correlacionado ou espacial) e os efeitos de competição (se houverem). Também, em caso de experimentos envolvendo múltiplos locais, os efeitos da interação genótipo x ambiente devem também ser incluídos no modelo.

No entanto, o procedimento ideal refere-se à inclusão simultânea dos efeitos dos marcadores no modelo de predição dos valores genéticos, de forma que o mapeamento seja realizado simultaneamente à predição. A superioridade dessa abordagem foi comprovada por Moreau et al. (1999) no contexto da análise espacial de experimentos. Este procedimento é superior devido ao fato de que os valores ou efeitos genéticos são preditos com diferentes precisões e também podem ser correlacionados devido à predição. Essas diferentes precisões e a correlação não são levadas em consideração quando não se adota a análise simultânea.

O ajuste dos dados fenotípicos antes da análise de QTL, visando eliminar efeitos ambientais é desejável. No entanto, não devem ser usados valores genéticos preditos sob o modelo genético poligênico infinitesimal. Podem ser usados valores genotípicos totais preditos. O uso do modelo infinitesimal supõe a ausência de QTL de grande efeito, que é exatamente o que se procura com a análise de QTL. E a consequência principal do uso do modelo infinitesimal é o incorreto uso da informação referente à segregação mendeliana, por ocasião da composição da matriz de parentesco. Maiores detalhes são apresentados no capítulo sobre GWS.

Mas é relevante enfatizar a necessidade de correção para os efeitos ambientais antes ou durante a análise de QTL. Com a correção há uma redução na amplitude de variação da população de mapeamento e torna-se mais difícil a detecção de QTL. Mas os resultados são mais realistas. Sem a correção para os efeitos ambientais, muitas vezes esses são mapeados como se fossem QTLs.

3.2 Análise de QTL como Efeito Aleatório via Modelos Lineares Mistos

O tradicional mapeamento de QTL baseia-se em análise de ligação, sendo que existem duas estratégias principais de modelagem: (i) tratamento dos efeitos de QTL como fixos e designação das origens dos alelos a cada fundador; (ii) tratamento dos efeitos de QTL como aleatórios e cômputo das matrizes de covariância IBD, condicionais à informação de marcadores. Perez-Enciso e Varona (2000) demonstraram que ambas abordagens são os extremos de uma modelagem genérica de modelos mistos. A opção de QTL como efeito fixo é apropriada quando a origem dos alelos pode ser identificada e o seu número é pequeno, como no cruzamento entre linhagens endogâmicas. A abordagem de QTL como efeito aleatório é mais flexível e encaixa na metodologia de modelos mistos, a qual apresenta inúmeras vantagens.

Assumindo QTLs como efeitos aleatórios, a significância dos efeitos dos locos marcados pode ser testada por meio do REMLRT no contexto dos modelos lineares mistos. Um modelo incluindo o efeito do suposto QTL, os efeitos poligênicos residuais (g*), os efeitos ambientais identificáveis (b) e os efeitos ambientais residuais é da forma $y = Xb + Qq + Zg^* + e$, em que q é um vetor de efeitos genéticos associados ao QTL marcado, com distribuição $q \sim N(0, G_q\sigma_q^2)$, em que σ_q^2 é a variância genética do QTL marcado e G_q é a matriz de covariância para q, condicional à informação do marcador. Todos os efeitos aleatórios são assumidos como não correlacionados e com distribuição normal multivariada, conforme a seguir: $g \sim N(0, A\sigma_g^2)$, $q \sim N(0, G_q\sigma_q^2)$ e $e \sim N(0, R\sigma_e^2)$, em que σ_g^2 é a variância genética aditiva poligênica, σ_e^2 é a variância residual, R é uma matriz diagonal conhecida e A é a matriz dos numeradores do coeficiente de parentesco genético aditivo de Wright. X e Z são as matrizes de incidência para os respectivos efeitos especificados. Para indivíduos não endógamos, G representa a proporção de alelos idênticos por descendência no QTL marcado.

Quando se assume que nenhum QTL marcado está segregando na população, o modelo misto é da forma $y = Xb + Zg + e$, o qual é hierárquico ao anterior. Assim, a presença de um QTL em uma particular posição no cromossomo pode ser testada pelo REMLRT envolvendo estes dois modelos. Estes modelos podem ser estendidos pela incorporação de efeitos espaciais, competição e interação genótipo x ambiente. Esses modelos podem ser ajustados no software ASREML desde que a matriz G seja informada pelo usuário.

Outro eficiente método de análise de QTL foi apresentado por Gilmour (2007). É denominado mapeamento via regressão sob modelos mistos (MMRM) e é adequado para populações de retrocruzamento e F2. Relaciona-se ao mapeamento por intervalo e por intervalo composto, mas difere no sentido em que se testa a presença de QTL´s em cada grupo de ligação, antes de fazer a regressão. Para isso, o método MMRM inicialmente ajusta todos os marcadores como efeitos aleatórios com variância comum dentro de cada grupo de ligação. A significância dos efeitos dos marcadores é avaliada via REMLRT e, se existir um componente de variância significativo associado com um grupo de ligação, a análise de QTL via regressão prossegue.

3.3 Análise de QTL em Famílias de Irmãos Germanos

O mapeamento de QTL em famílias de irmãos germanos é comum em plantas perenes, animais e humanos. Um método que tem sido aplicado nessas espécies é a análise por intervalo simples por meio da regressão de pares de irmãos. Esse método foi apresentado por Fulker e Cardon (1994) e deriva do método de Haseman e Elston (1972). O método de Haseman e Elston é fundamentado na regressão linear dos quadrados das diferenças fenotípicas entre dois irmãos dentro de uma família, em função da proporção de genes idênticos por descendência (IBD) compartilhados entre eles, ou seja, entre pares de indivíduos que possuem o mesmo QTL marcado. Esse método tem a limitação de confundir o efeito genético do QTL com a taxa de recombinação entre o QTL e o marcador. O método de Fulker e Cardon foi desenvolvido visando isolar a variância do QTL da taxa de recombinação, bem como localizar o QTL no cromossomo. É uma extensão do método de Haseman e Elston e trata-se de procedimento de mapeamento por intervalo, apresentando maior poder. Tal procedimento utiliza dois marcadores flanqueando o QTL para estimar separadamente a posição e o efeito do QTL sobre o caráter.

O método da regressão de pares de irmãos baseia-se na similaridade entre indivíduos aparentados. O seguinte modelo pode ser especificado: $y = Xb + Zg + q + e$, em que y é o vetor de dados fenotípicos, b é o vetor de efeitos ambientais identificáveis (efeitos fixos), g é o vetor dos efeitos genéticos devidos aos poligenes, q é o efeito genético devido ao QTL e e é o efeito ambiental residual. X e Z são matrizes de incidência que associam b e g aos dados fenotípicos. Os efeitos g são considerados aleatórios e o efeito do QTL pode ser considerado como fixo ou aleatório.

O componente poligênico g é dependente do parentesco genético entre os indivíduos em avaliação e o componente do QTL depende da proporção de alelos idênticos por descendência (IBD) entre pares de indivíduos que possuem o mesmo QTL. O componente poligênico está associado a muitos genes de pequenos efeitos, e é assumido que a média (sobre os vários poligenes) da proporção de alelos IBD entre dois indivíduos é equivalente ao parentesco genético médio entre dois irmãos. No caso do QTL, a proporção IBD varia entre pares de irmãos e é estimada através dos genótipos observados nos locos dos marcadores. Logicamente, a proporção de alelos IBD do QTL não é observada diretamente. Os IBDs do QTL são avaliados em cada segregação em nível individual e não em nível médio, permitindo conhecer exatamente cada segregação.

A proporção π_m de alelos IBD entre irmãos em um loco marcador informativo pode ser 0; 0,5 ou 1, quando os indivíduos tem 0; 1 ou 2 alelos parentais em comum, respectivamente. De acordo com Haseman e Elston (1972), o cômputo de π_m pode ser dado por $\pi_m = f_2 + (1/2)f_1$, em que f_i é a probabilidade de que dois indivíduos possuam i (= 0, 1 ou 2) alelos IBD, ou seja, alelos que são cópia de um mesmo alelo ancestral. Essas probabilidades são dadas pela análise dos genótipos marcadores observados nos irmãos e em seus pais. Um par de irmãos completos pode possuir

zero, um ou dois alelos IBD com probabilidades (1/4), (1/2) e (1/4), respectivamente. Assim, nesse caso, $\pi_m = f_2 + (1/2)f_1 = (1/4) + (1/2)(1/2) = 1/2$.

Fulker e Cardon (1994) apresentaram uma expressão para o cômputo da média condicional da proporção de alelos IBD do QTL como função linear dos π_m nos dois marcadores de um intervalo. Essa expressão é dada por $\hat{\pi}_q = E[\pi_q \mid \pi_{m1}\pi_{m2}] = \alpha + \beta_1\pi_{m1} + \beta_2\pi_{m2}$, em que π_{m1} e π_{m2} são as proporções IBD para os dois marcadores. Os valores de β são dados pelo sistema de equações: $\begin{bmatrix} Cov(\pi_{m1},\pi_q) \\ Cov(\pi_{m2},\pi_q) \end{bmatrix} = \begin{bmatrix} Var(\pi_{m1}) & Cov(\pi_{m1},\pi_{m2}) \\ Cov(\pi_{m1},\pi_{m2}) & Var(\pi_{m2}) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$.

Define-se r_{12}, r_{1q} e r_{q2} como as taxas de recombinação entre os dois marcadores, entre marcador 1 e QTL e entre marcador 2 e QTL, respectivamente. Para irmãos germanos, têm-se as equivalências $Var(\pi_{mi}) = 1/8$, $Cov(\pi_i,\pi_j) = (1-2r_{ij})^2/8$ e $\pi_m = (1/2)$. Resolvendo-se o sistema matricial, obtêm-se os estimadores de β para famílias de irmãos completos:

$$\hat{\beta}_1 = [(1-2r_{1q})^2 - (1-2r_{q2})^2(1-2r_{12})^2]/[1-(1-2r_{12})^4$$
$$\hat{\beta}_2 = [(1-2r_{q2})^2 - (1-2r_{1q})^2(1-2r_{12})^2]/[1-(1-2r_{12})^4.$$

O componente α é dado por $\alpha = (1-\hat{\beta}_1 - \hat{\beta}_2)/2$. De posse das estimativas de α e β, obtém-se a proporção IBD ($\hat{\pi}_m$) para o QTL. Essa proporção depende essencialmente da fração de recombinação entre os locos. As freqüências de recombinação podem ser computadas a partir da freqüência gamética de cada genitor ou a partir da freqüência genotípica da progênie. Informações multilocos entre pares de marcas adjacentes são usadas na estimação.

O algoritmo proposto por Fulker e Cardon (1994) atua da seguinte forma: (i) para qualquer intervalo entre dois marcadores flanqueadores, divida o intervalo λ_{12} em N intervalos de λ_1 e λ_2 tal que $\lambda_{12} = \lambda_1 + \lambda_2$; (ii) converta os valores de λ em r_{ij}, usando para isso uma função de mapeamento como a de Haldane; (iii) estime $\hat{\pi}_q$ usando as expressões para α e β; (iv) regresse os quadrados das diferenças fenotípicas entre dois irmãos dentro de uma família nas N estimativas $\hat{\pi}_q$; (v) selecione o coeficiente de regressão $\beta_{\hat{\pi}_q}$ que corresponde à mínima soma de quadrados dos resíduos e calcule σ_q^2 (via $\beta_{\hat{\pi}_q}$, conforme relação apresentada abaixo), a estatística t para $\beta_{\hat{\pi}_q}$ e obtenha a localização do QTL. Cruz et al. (2009) apresentam detalhes desse método.

Segundo o modelo $y = Xb + Zg + q + e$, tem-se que $Var(y) = \sigma_g^2 + \sigma_q^2 + \sigma_e^2$ e a covariância entre pares de irmãos é dada por $Cov(Y_{ij}, Y_{ij'}) = (1/2)\sigma_g^2 + \pi_q\sigma_q^2$, onde π_q é substituído por $\hat{\pi}_q$. No caso, a variância aditiva contribuída por todos os locos equivale a $\sigma_a^2 = \sigma_g^2 + \sigma_q^2$.

A análise de famílias de irmãos germanos pode basear-se também em ANOVA (Lynch e Walsh, 1998), por meio do quadrado médio entre classes de genótipos dos marcadores e cômputo de uma estatística F (razão entre o referido quadrado médio e a variância residual) para cada loco marcador por vez. Dessa forma, é possível estimar o efeito de substituição alélica para cada genitor, ou seja, para o feminino e masculino, que representam duas populações distintas quando o cruzamento é completamente informativo (do tipo M_iM_j x M_kM_l, conforme Tabela 19). Pode-se também estimar a variância genética total associada com o marcador. A ANOVA terá duas fontes de informação: (i) entre genótipos marcadores (M_iM_k, M_iM_l, M_jM_k e M_jM_l, com 3 graus de liberdade para o caso de cruzamento completamente informativo); (ii) resíduo.

Tabela 19. Constituição genotípica dos genitores e nível de informatividade

Tipo de Cruzamento	Constituição Genotípica	Grau de Informação	Proporção de Segregação
Cruzamento entre F1 Divergentes	M_iM_j x M_kM_l	Toda a progênie é informativa para ambos os genitores	1:1:1:1
Retrocruzamento	M_iM_j x M_kM_k	A progênie é informativa somente para o genitor heterozigoto	1:1
Cruzamento entre F1 Idênticos (Geração de F2)	M_iM_j x M_iM_j	Somente indivíduos homozigotos da progênie são informativos	1:2:1

Quando várias famílias existem, pode-se também realizar uma ANOVA com efeito de marcador hierárquico dentro de cada família. A ANOVA terá três fontes de informação: (i) entre famílias (com f-1 graus de liberdade); (ii) entre genótipos marcadores (com 3f graus de liberdade para o caso de cruzamento completamente informativo); (iii) resíduo.

Contrastes de médias para os efeitos alélicos dentro de cada genitor da família de irmãos completos podem ser realizados. Para o genitor M_iM_j a diferença entre a média fenotípica dos indivíduos com o alelo M_i no loco marcador e a média fenotípica dos indivíduos com o alelo M_j fornece a seguinte quantidade $M_{if} - M_{jf} = (1-2r)(a_i - a_j)$, em ausência de dominância. As quantidades r, a_i e a_j referem-se à taxa de recombinação, efeito médio do alelo i e efeito médio do alelo j, respectivamente. Com ligação completa entre o loco do QTL (Q) e o loco do marcador (M) tem-se r = o e a quantidade $a_{ij} = (a_i - a_j)$ fornece o efeito médio de substituição gênica. Esse efeito refere-se à conseqüência média de se substituir o alelo Q_j por Q_i no heterozigoto Q_iQ_j (tornando-o Q_iQ_i) e no homozigoto Q_jQ_j (tornando-o Q_iQ_j). De maneira similar, pode-se obter a quantidade $a_{kl} = (a_k - a_l)$, que fornece o efeito médio de substituição gênica de Q_l por Q_k. Comparando-se a_{ij} com a_{kl}, pode-se inferir qual dos quatro alelos é mais favorável. A Tabela 20 ilustra essa questão. Verifica-se que o alelo mais favorável é M_i, seguido por M_l.

Tabela 20. Efeitos alélicos comparativos no cruzamento M_iM_j x M_kM_l, dados as médias fenotípicas.

Diferença Alélica	Média Fenotípica dos Alelos M_i e M_k	Média Fenotípica dos Alelos M_j e M_l	Efeito do Alelo no Genitor
M_i-M_j	17.7 (M_i)	12.03 (M_j)	5.67
M_k-M_l	14.5 (M_k)	15.64 (M_l)	-1.14

M_i: média de (M_iM_k + M_iM_l); M_j: média de (M_jM_k + M_jM_l); M_k: média de (M_iM_k + M_jM_k) ; M_l: média de (M_iM_l + M_jM_l).

3.4 Estimação da Herdabilidade via Parentesco Genômico

Conforme Lynch e Walsh (1998), o modelo para o valor fenotípico de um (j) dos membros do par i de irmãos completos é dado por $Y_{ij} = u + q_{ij} + e_{ij}$, em que q é o efeito aditivo do QTL e e é o efeito residual, o qual inclui efeito ambiental e poligênico residual. A diferença entre efeitos residuais dos dois indivíduos do par, $e_i = e_{i1} - e_{i2}$, é assumida com média zero e variância σ_e^2, e não correlacionada com $q_i = q_{i1} - q_{i2}$. Constata-se que a diferença entre os valores fenotípicos dos irmãos é desejável pois, cancelam os efeitos de ambiente comum que afetam os membros da família.

O quadrado da diferença entre os valores fenotípicos dos irmãos tem valor esperado dado por

$$E(Y_j) = E[(q_{i1} - q_{i2} + e_{i1} - e_{i2})^2]$$
$$= E[(q_{i1} - q_{i2})^2] + \sigma_e^2$$
$$= 2[\sigma_q^2 - \sigma(q_{i1}, q_{i2})] + \sigma_e^2$$

A expressão para a covariância equivale a $\sigma(q_{i1}, q_{i2})] = \sigma_q^2\pi_q$. A esperança de Y condicional à proporção de alelos IBD no QTL é dada por $E(Y_i|\pi_q) = \alpha + \beta\pi_q = (2\sigma_q^2 + \sigma_e^2) - (2\sigma_q^2)\pi_q$ em que a inclinação (β) da regressão tem sinal negativo. Genericamente (r diferente de zero), considerando a fração de recombinação entre marcador e QTL tem-se, conforme Haseman e Elston (1972): $\alpha = 2[1-2(1-r)r]\sigma_q^2 + \sigma_e^2$ e $\beta = 2(1-2r)^2\sigma_q^2$. Uma inclinação significativa propicia evidência de um QTL ligado ao marcador. E o poder do teste estatístico é dependente das magnitudes de r e de σ_q^2. Em ausência de dominância, porém com ligação incompleta, $E(\beta) = -2(1-2r)^2\sigma_q^2$.

Pelo método de Fulker e Cardon, o coeficiente de regressão $\beta_{\hat{\pi}_q}$ é relacionado à herdabilidade do loco e permite estimá-la. A regressão dos quadrados das diferenças fenotípicas entre dois irmãos dentro de uma família na estimativa da proporção IBD $\hat{\pi}_q$ obedece a seguinte equação: $(Y_{i1} - Y_{i2})^2|\hat{\pi}_q = \alpha + \beta\hat{\pi}_q$, em que Y_{i1} e Y_{i2} referem-se aos fenótipos dos indivíduos 1 e 2 da família i. O coeficiente de regressão β é proporcional à variância genética aditiva contribuída pelo loco (σ_q^2) e quando os genitores são não endógamos e a ligação é completa, em ausência de dominância, equivale a $\beta = -2\sigma_q^2$, ou seja $\sigma_q^2 = -\beta/2$.

Assim, a herdabilidade aditiva do loco é dada por $\hat{h}^2_{aq} = -\beta/(2\sigma^2_y)$, em que σ^2_y é a variância fenotípica individual da população (não apenas dentro da família de irmãos completos). Quando apenas a variância fenotípica individual dentro de família de irmãos completos (σ^2_{ydfic}) é computada, a herdabilidade do QTL dentro de família deve ser calculada como $\hat{h}^2_{aqd} = [(1/2)\sigma^2_q]/\sigma^2_{ydfic} = -\beta/(4\sigma^2_{ydfic})$ e a herdabilidade do QTL na população deve ser calculada como $\hat{h}^2_{aq} = \sigma^2_q/[\sigma^2_{ydfic} + (1/2)\sigma^2_q] = (-\beta/2)/(\sigma^2_{ydfic} - \beta/4)$. Se a variância fenotípica individual da população for computada, a herdabilidade do QTL dentro de família deve ser calculada por $\hat{h}^2_{aqd} = [(1/2)\sigma^2_q]/[\sigma^2_{ydfic} - (1/2)\sigma^2_q] = (-\beta/4)/(\sigma^2_y + \beta/4)$.

A herdabilidade de todo o caráter (sobre todos os locos) pode ser calculada por meio da regressão (β^*) dos quadrados das diferenças fenotípicas entre dois irmãos dentro de uma família na estimativa da proporção IBD ampla em todo o genoma. Nesse caso, $\beta^* = -2\sigma^2_g$ e $\hat{h}^2 = -\beta^*/(2\sigma^2_y)$. Maiores detalhes são apresentados por Visscher et al. (2006) e Odegard e Meuwissen (2012).

Para a estimação de parâmetros genéticos tais como a herdabilidade, são necessárias informações fenotípicas e de parentesco entre os indivíduos avaliados. As análises genéticas de dados moleculares fornecem informações sobre o parentesco entre os indivíduos. Resende (2008) apresenta estimadores para a herdabilidade nestas condições.

Definindo $Z_{ij} = \dfrac{(y_i - \bar{y})\ (y_j - \bar{y})}{Var\ (y)}$, como a similaridade fenotípica entre dois indivíduos na população, em que y_i e y_j referem-se a observações fenotípicas nos indivíduos i e j, $\bar{y}$ e $Var\ (y)$ são relativos à média e à variância do caráter y na população, tem-se: $Z_i = \dfrac{2r_{ij}\ \hat{\sigma}^2_g + e_{ij}}{\sigma^2_y} = 2r_{ij}\ h^2 + \dfrac{e_{ij}}{\sigma^2_y}$, em que: r_{ij}: coeficiente de parentesco de Malecot entre os indivíduos i e j; $h^2 = \dfrac{\sigma^2_g}{\sigma^2_y}$: herdabilidade individual no sentido restrito; $e'_{ij} = \dfrac{e_{ij}}{\sigma^2_y}$: resíduo devido aos efeitos ambientais; σ^2_g: variância genética aditiva; $\sigma^2_y = Var\ (y)$. O estimador da herdabilidade é dado por $\hat{h}^2 = \text{cov}\ (Z_{ij}, r_{ij})/[2\ Var(r_{ij})]$, pois $h^2 = \dfrac{\text{cov}\ (Z_{ij}, r_{ij})}{2\ Var\ (r_{ij})} = \dfrac{\text{cov}\ [(2\ r_{ij}\ h^2 + e_{ij}'),\ r_{ij}]}{2\ Var\ (r_{ij})} = \dfrac{2\ Var\ (r_{ij})\ h^2}{2\ Var\ (r_{ij})} = h^2$.

3.5 Funções de Mapeamento

Para a construção de um mapa de ligação é preciso que os marcadores ou genes sejam de herança simples. Os seguintes passos são adotados (Schuster e Cruz, 2004): (i) estimação da frequência de recombinação (distância) entre pares de marcadores; (ii) agrupamento dos marcadores em diferentes grupos de ligação; (iii) definição da ordem dos marcadores em cada grupo de ligação; (iv) estimação da frequência de recombinação multiponto entre marcadores adjacentes. Genericamente, a frequência de recombinação entre dois locos pode ser estimada como a razão entre o número de indivíduos com gametas recombinantes e o número total de indivíduos analisados. A frequência de recombinação expressa também a distância entre os locos. A partir da verificação da existência de ligação entre genes e do cálculo da distância entre eles, os mesmos podem ser ordenados e classificados em grupos de ligação.

A base do mapeamento é decorrente do fato de que a probabilidade de recombinação é maior para locos mais distantes do que para locos próximos. Por meio do conhecimento das frequências de recombinação entre diversos locos de um grupo de ligação, torna-se possível a estimação da ordem desses locos no grupo de ligação. Grupo de ligação é definido como um conjunto de marcadores genéticos que possuem menos de 50% de recombinação entre marcadores consecutivos (Schuster e Cruz, 2004). Além dos mapas genéticos, mapas físicos podem ser construídos por meio de técnicas citogenéticas, fragmentos de restrição e também pelo sequenciamento do genoma. As informações desses dois tipos de mapa são fundamentais para a clonagem de genes. Nesse caso, o mapeamento fino é necessário, visando a obtenção de mapas genéticos bastante saturados.

Para estimar a localização de um novo loco no genoma, é necessário assumir uma relação funcional entre fração de recombinação e distância genética entre pares de locos. Essa distância equivale ao número esperado de permutas que ocorre entre esses dois locos por ocasião da meiose. Uma vez que essas esperanças matemáticas são aditivas, essa definição propicia uma medida estatística aditiva de localização. A unidade de distância genética é o Morgan, que refere-se à distância em que se espera que ocorra uma permuta. A unidade de mapa de 1 cM equivale à freqüência de recombinação de 1%.

Quando dois locos são muito próximos, no máximo uma permuta pode ocorrer entre eles e então a distância genética equivale à fração de recombinação. Quando no máximo uma permuta pode ocorrer entre dois locos, tem-se o que é denominado interferência completa ou positiva. A interferência significa que uma permuta interfere na formação de qualquer permuta adicional. Por outro lado, se as permutas formam um processo Poisson ao longo do cromossomo, não há interferência, ou seja, a presença de uma permuta em um ponto qualquer não afeta a presenca de uma permuta em qualquer outro ponto do cromossomo. Nesse caso, tem-se permutas independentes e interferência nula.

Uma função que relaciona a distância genética no mapa com a frequênciia de recombinação é denominada função de mapa ou de mapeamento. As seguintes

funções de mapeamento foram propostas: (i) Haldane (1919) a qual assume interferência nula e é amplamente usada porque a independência condicional que ela assume conduz a maior simplicidade computacional (mas não é realística); (ii) Morgan (1928), que assume interferência completa e é uma aproximação realista para o caso de pequenas frações de recombinação; (iii) Kosambi (1944), que considera o nível de interferência.

As funções de mapeamento convertem frequências de recombinação dadas em unidades de mapa (cM) em distâncias entre genes. Visam tornar aditivas as distâncias entre pares de marcas. Sendo r a frequência de recombinação e D a distância entre locos tem-se os seguintes estimadores ou funções de mapeamento:

(a) Haldane

$D_H = -0{,}5 \ln(1 - 2r)$: distância de Haldane entre locos.

$r = (1 - e^{-2D_H})/2$: frequência de recombinação dada a distância de Haldane.

(b) Kosambi

$D_K = -0{,}25 \ln[(1 + 2r)/(1 - 2r)]$: distância de Kosambi entre locos.

$r = 0{,}5[(e^{4D_K} - 1)/(e^{4D_K} + 1)]$: frequência de recombinação dada a distância de Kosambi.

As funções de mapeamento de Kosambi e de Haldane são similares quando r apresenta valor próximo a zero. Elas diferem a medida em que r aumenta. Por exemplo, r = 0,30 corresponde às distâncias de mapa de 46 cM e 35 cM pelos métodos de Haldane e de Kosambi, respectivamente (Bernardo, 2002).

Para a formação dos grupos de ligação é necessário definir um limite máximo para a frequência de recombinação entre dois marcadores e também um limite mínimo para o LOD escore (logaritmo da razão de riscos), visando inferir que os dois marcadores estão ligados. Geralmente, esses limites tem sido adotados como r = 0,3 (ou 30 cM) e LOD = 3. Um LOD escore acima de 3 é geralmente usado como valor crítico, significando que a hipótese alternativa é 1000 vezes mais provável do que a hipótese nula (hipótese de independência entre locos). Esse critério parece muito severo. Entretanto, ele leva em consideração a probabilidade a priori de ligação. Conforme Norton (1955), existe uma probabilidade razoável (5% em seres humanos, em 23 pares de cromossomos) de que dois locos sejam ligados, devido ao número finito de cromossomos. O LOD escore para um valor particular θ de recombinação pode ser escrito como LOD-escore(θ) = (N − Nrec) log (1-θ) + Nrec log(θ) − N log(0.5), onde N é o número de indivíduos na progênie e Nrec é o número de recombinantes.

3.6 Análise da Expressão Gênica

Os estudos genômicos iniciaram-se com o mapeamento de QTLs por meio de varredura genômica de baixa densidade. Posteriormente, a seleção assistida por marcadores (MAS) foi proposta e implementada por meio de um modelo de herança mista combinando o componente poligênico com um componente devido a QTL de grande efeito (Fernando e Grossman, 1989; Lande e Thompson, 1990). Com a chegada dos marcadores SNP, tornou-se comum o mapeamento de associação, implementado via varredura genômica de alta densidade, possibilitando o mapeamento fino. Também, tornou-se possível a seleção genômica (GS) ou seleção genômica ampla (GWS), que é superior à MAS. A transcriptômica (ou transcrissômica, referente à expressão gênica) e a proteômica surgiram também como novas fontes de informação que podem ser usadas nos procedimentos de avaliação genética.

O termo Genética Genômica foi criado por Jansen e Nap (2001) para designar o estudo conjunto da variabilidade do transcriptoma e do polimorfismo de seqüências de DNA. Nessa linha, dois enfoques são empregados: (i) determinação da arquitetura genética do transcriptoma, em forma de análise de milhares de QTLs de expressão (eQTL), onde os fenótipos são níveis de cDNA (DNA complementar) associados a cada gene; (ii) uso de dados de expressão gênica para a localização de genes candidatos. Para que essa última abordagem tenha sucesso, é necessário que os níveis de expressão gênica estejam sob algum controle genético e que alguns dos níveis de expressão herdáveis estejam correlacionados com o caráter de interesse. Perez-Enciso et al. (2003) relatam a combinação das informações dos marcadores moleculares e das expressões gênicas para o mapeamento de características quantitativas.

Os dados de expressão referem-se à transcrição (níveis de RNA mensageiro). A tecnologia baseada em microarranjos é usada para determinar a expressão diferencial de genes, de todo o genoma, em amostras biológicas de tecidos específicos. Recentemente (Resende Jr., 2012), a tecnologia de sequenciamento em larga escala tem sido utilizada como uma alternativa ao uso de microarranjos. Esta técnica é conhecida como RNA-seq e baseia-se no sequenciamento de uma amostra de todos os transcritos de um individuo em determinada condição e em determinado tecido. A profundidade de leitura (*depth)* associada a cada transcrito é correlacionada com o nível de expressão do gene em questão. Maiores detalhes sobre estas abordagens são dados adiantes.

Os níveis de expressão gênica ou quantidades de RNAm detectados são então submetidos a análise de correlação com caracteres quantitativos em indivíduos de uma população segregante, visando à detecção de QTL. Como exemplo, diferenças na quantidade de RNAm produzida por plantas resistentes e suscetíveis a uma doença podem indicar que determinado RNAm está associado a um gene de resistência. O uso da quantidade de expressão gênica para a detecção de QTL é mais adequada para caracteres de resistência a estresses causados por fatores abióticos como seca e salinidade e, também, caracteres de resistência à doenças e pragas.

Na Genética Genômica, a associação entre nível de RNA mensageiro e polimorfismo de DNA, ao invés da associação entre fenótipo e polimorfismo de DNA, se justifica pela maior proximidade entre RNA e DNA do que entre fenótipo e DNA. Mas uma questão fundamental é como fazer a ligação entre a expressão de um QTL com o caráter fenotípico de interesse. Métodos diretos de análise da função e expressão gênica são também essenciais para determinar se dois marcadores muito próximos estão detectando o mesmo QTL ou dois QTLs muito próximos.

A combinação de dados genéticos e de expressão gênica sobre todo o genoma tem permitido entender a base genética da expressão gênica. Nesse caso, os níveis de RNAm são os dados fenotípicos, sujeitos a variações devidas à causas genéticas e ambientais. Neste caso, são identificados regiões do genoma que controlam o nível de expressão dos genes estudados. Basicamente, a regulação do nível de expressão e dividida em duas classes, *cis* e *trans*. Caso o polimorfismo associado ao nível de expressão diferencial esteja muito próximo ao gene do qual o mRNA foi transcrito, a regulação é do tipo *cis*. Do contrário, caso o marcador (e consequentemente o eQTL) esteja mapeado em uma posição diferente da posição do transcrito, o gene é regulado em *trans*. Este último tipo de regulação está normalmente associada a um fator de transcrição que altera (ou ativa/desativa) o nível de expressão do mRNA em questão (Resende Jr., 2012). Estudos têm demonstrado grande variação genética entre genótipos quanto à expressão gênica e estimativas significativas de herdabilidade têm sido obtidas. Em humanos, a herdabilidade dos níveis de expressão gênica é em média igual a 30 %. Isto é importante porque o poder estatístico para detectar variantes genéticos que afetam a expressão gênica depende da herdabilidade. Os genes são expressos em função de um estímulo ambiental.

Os dados de microarranjos (também referidos como slides ou lâminas) de DNA envolvem simultaneamente a expressão de milhares de genes em determinada idade do indivíduo e sob certas condições ambientais. Os procedimentos laboratoriais para a produção desse tipo de dados envolvem a extração de RNA mensageiro (mRNA), transcrição reversa para a obtenção do DNA complementar (cDNA), marcação fluorescente e hibridização do cDNA com sondas comerciais de DNA. A técnica de microarranjos propicia uma inferência sobre o nível de expressão gênica, via a abundância dos RNAs transcritos. Possibilita também, em alguns casos, a integração entre genética e fisiologia, via determinação de redes (*networks*) entre conjuntos de genes associados a características fisiológicas. Uma desvantagem do uso de microarranjos é a necessidade de conhecimento prévio das sequencias de DNA para desenvolvimento das sondas usadas na hibridização. Assim, caso um transcrito não seja previamente conhecido, não é possível construir uma sonda e assim a expressão desse gene não será detectada. No caso de RNA-seq, uma amostra de todos os transcritos é sequenciada, independente do conhecimento prévio da sequencia de cada gene(Resende Jr., 2012).

A análise de expressão gênica permite inferir sobre a função dos genes e possibilita a compreensão da expressão gênica diferencial entre tecidos, fases do desenvolvimento, em respostas a estresses ambientais, e entre genótipos distintos. A análise desse tipo de dados é tratada com detalhes na literatura (Kerr et al., 2000; Wolfinger et al.; 2001; Tempelman, 2005; Rosa et al., 2007; Ayroles e Gibson, 2006).

No caso de microarranjos, dois tipos de plataforma de arranjos podem ser usadas: (i) baseada em um sistema de duas cores, gerando duas amostras por arranjo (do tipo spotted cDNA); (ii) baseada em um sistema de cor (dye) única ou arranjo de canal único, gerando uma amostra por arranjo (do tipo Affymetrix). O sistema (i) demanda delineamentos (em loop ou circulares, parcelas subdivididas) e análises mais complexos, além de um alto nível de repetições técnicas. Por outro lado, o sistema (ii) permite múltiplas sondas por gene e apresenta uma tendência de se usar menor número de repetições. Também o delineamento é mais simples e não existe amostra de referência.

Duas abordagens principais têm sido empregadas na experimentação com arranjos de duas cores: (i) uma cor (verde = Cyanine 3 ou Cy3) reservada para a amostra de referência ou controle e outra cor (vermelha = Cyanine 5 ou Cy5) usada para avaliar os tratamentos; (ii) as duas cores são usadas para avaliar tratamentos de interesse. Na abordagem (i) a proporção Cy3/Cy5 entre as intensidades de fluorescência propicia uma medida de intensidade de expressão gênica. Essa abordagem é intuitiva e adequada em situações em que existe um grande número de tratamentos do mesmo fator com baixo número de repetições. A abordagem (ii) requer delineamentos mais refinados para evitar o confundimento entre fatores (lâminas e amostras de ácido nucléico). Os efeitos de corantes ou *dye* são pronunciados e torna-se essencial que cada amostra seja representada por repetições técnicas de ambos *dyes* em iguais proporções.

O delineamento em loop deve ser empregado quando o interesse é contrastar as contribuições de cada fator. O delineamento em parcela subdividida deve ser usado quando o interesse reside no efeito de um fator através de amostras que incluem efeitos de um outro fator de menor interesse. Para qualquer das duas abordagens (i e ii) o efeito de arranjo deve ser ajustado como aleatório, visando considerar o fato de que as duas medidas em um mesmo arranjo são correlacionadas. Isso ajusta para o efeito do ambiente comum de arranjo. O delineamento experimental guia a formulação do modelo linear apropriado para a análise. Cada lâmina ou arranjo é análogo a um bloco incompleto pois, contempla apenas dois dos vários tratamentos. Adicionalmente, cada lâmina contém os efeitos de dois corantes e o delineamento torna-se então do tipo linha e coluna com dimensão 2 x s, em que s é o número de slides ou arranjos.

No caso de arranjos de canal único o delineamento experimental é simplificado e não há necessidade de considerar os efeitos de arranjo e de dye, pois não há confundimento uma vez que cada amostra é hibridizada sobre um arranjo diferente e é medida independentemente. A amostra de referência ou controle é usada exatamente para corrigir os dados para os efeitos de lâmina. Nesse caso, o delineamento é do tipo blocos incompletos com tratamentos comuns. A comparação entre tratamentos é realizada de forma indireta, por meio da diferença entre contrastes de cada tratamento e a referência ou controle em cada lâmina. Por outro lado, nos delineamentos circulares, os efeitos comparativos de tratamentos são estimados por meio de combinações entre comparações diretas (entre tratamentos dentro de blocos ou lâminas) e de comparações indiretas (entre tratamentos entre blocos ou lâminas).

No caso do sequenciamento de cDNA (RNA-seq), os delineamentos experimentais tendem a ser mais simples. Em geral, fatores ajustados no modelo são os efeitos de diferentes canais (*lanes*), efeitos de diferentes corridas (*flow-cell*) e estes são ajustados em um delineamento em blocos ao acaso (Auer e Doerge, 2010). A expressão gênica é quantificada normalmente pela normalização da cobertura de leitura e do tamanho da sequencia. O método mais comum é a normalização em número de leitura por quilobase (Kb) por milhão de sequencias mapeadas à referencia (RPKM, do ingles – *Reads Per Kilobase of exon model per Million mapped reads* – Mortazavi *et al.* 2008) (Resende Jr., 2012).

Em transcriptômica é feita distinção entre repetição técnica e repetição biológica. Repetição técnica refere-se a repetir hibridização das mesmas amostras de RNA originadas de uma mesma fonte biológica comum. Assim, essas repetições não são totalmente independentes uma da outra e são usadas para validar a acurácia das medidas do nível de transcritos e para modelar efeitos residuais como a variação devida ao sequenciamento da mesma amostra em diferentes canais. Assim, não propiciam informação sobre o nível de variação na população. De maneira similar, sondas repetidas dentro de um arranjo são usadas para reduzir a necessidade de repetições técnicas por meio do aumento da confiança de medidas de abundância de transcritos para determinado gene alvo. Com os arranjos comerciais de alta qualidade, os erros técnicos são muito menores do que a variância biológica, de forma que geralmente não há motivos para usar mais que duas repetições por amostra.

Repetição biológica refere-se a repetir hibridização de amostras de RNA originadas de fontes biológicas independentes sob as mesmas condições ou tratamentos, tais quais amostras extraídas de diferentes indivíduos que receberam a mesma dose de um tratamento ou mesmo duas réplicas de um mesmo genótipo de uma planta. Essas réplicas objetivam propiciar informação sobre a variação biológica entre indivíduos (Ayroles e Gibson, 2006). Quanto ao número de repetições a se utilizar, Wolfinger et al. (2001) e Tempelman (2005) recomendam o uso de ao menos quatro repetições técnicas para cada repetição biológica visando detectar 80% dos genes expressos diferencialmente entre os grupos experimentais.

Os dados de intensidade de expressão em cada dye são inicialmente convertidos para escala logarítmica na base 2. A transformação log tem a vantagem de tornar os dados mais próximos a uma distribuição normal e mais simétricos. Com dados transformados na escala logarítmica, componentes de média associados a modelos lineares podem ser usados como procedimentos estatísticos adequados. Ou seja, não há necessidade de se usar outras estatísticas, como a mediana. Sem a aplicação da transformação logarítmica o uso da estatística mediana é recomendado, pois essa é robusta a dados discrepantes (outliers). Após limpeza (remoção de genes não expressos, arranjos com baixa intensidade, etc) dos dados, os mesmos necessitam ser normalizados visando remover efeitos globais de arranjos e corantes, os quais não refletem variação genética verdadeira dentro ou entre arranjos. Esses vieses resultam de fatores tal qual a variação da quantidade de DNA colocada entre arranjos. Métodos de normalização tal qual o LOWESS (regressão não-paramétrica robusta) podem ser usados. Esse método usa regressões locais para remover correlações gerais

entre intensidade e proporção de intensidade. Outro procedimento é a normalização de quantis ou quantílica, a qual realiza uma transformação não linear que produz cada arranjo com iguais médias, medianas e variâncias, por meio da obtenção da intensidade média de cada quantil através dos arranjos.

No entanto, tal normalização global pode remover artificialmente verdadeiras diferenças biológicas. Assim, modelos alternativos podem ser usados para remover os efeitos de lâminas e de corantes. Em outras palavras, a própria modelagem desses efeitos na análise estatística permite ajustar os dados para os mesmos. Wolfinger et al. (2001) propuseram uma modelagem em duas etapas: (i) o primeiro modelo ajusta os dados (transformados por log) para os efeitos globais (todos os genes simultaneamente) de lâmina ou arranjo (A), corantes ou dye (D) e sua interação (AD) por meio do modelo $Log_2(y) = u + A + D + AD + Resíduo(1)$; (ii) o segundo usa o Resíduo estimado pela modelagem anterior em um novo modelo de análise designado para genes específicos ou individuais. O primeiro modelo, designado para uma normalização global, expressa a intensidade de fluorescência como desvios da média geral e a segunda modelagem permite inferir se esses desvios diferem entre fatores (tratamentos, etc) do modelo e para genes individuais.

O modelo gene-específico de Wolfinger et al. (2001) é dado por $Residuo(1) = u + A + D + AD + T +$ erro, em que T é o fator de tratamentos e erro é um vetor de erros específico para cada gene. Esse modelo é ajustado separadamente para cada gene no arranjo e, portanto, considera componentes de variância específicos para cada gene. Os efeitos A e AD devem ser ajustados como aleatórios e os efeitos do fator D como efeitos fixos. Os efeitos do fator T devem ser tomados como fixos quando se referirem a comparação de diferentes níveis de estresse aos quais determinado genótipo é submetido e tomados como aleatórios quando se referirem a mais de cinco genótipos tomados de uma população. Testes de significância podem ser aplicados aos fatores de efeitos fixos (F, Wald) e aleatórios (LRT ou análise de deviance).

Uma alternativa é a realização da normalização simultaneamente ao ajuste de todos os demais fatores do modelo e também da avaliação de todos os efeitos de genes individuais, conforme Kerr et al. (2000), por meio do modelo:
$y_{ijkm} = u + A_i + D_j + AD_{ij} + G_m + AG_{im} + DG_{jm} + T_k + TG_{km} + e_{ijkm}$, em que y_{ijkm} é a variável abundância de transcrição na escala log_2 e e_{ijkm} é um resíduo comum a todos os genes. Os efeitos de genes (G) e suas interações devem ser considerados como aleatórios. A interação de maior interesse é TG_{km} que retrata o efeito do tratamento k sobre o nível de expressão do gene m. Modelos mais complexos, contemplando níveis de variação biológica (diferentes genótipos) podem também serem usados e permitem a estimação de componentes de variância e herdabilidade dos padrões de expressão gênica. Esse modelo é relevante porque considera todos os efeitos simultaneamente em uma única análise. No entanto, apresenta a desvantagem de considerar uma variância residual comum a todos os genes.

O uso do método de quadrados mínimos na análise de dados de microarranjos com todos os genes simultaneamente apresenta restrições, devido ao elevado número de genes em relação ao número de lâminas, ou seja, maior número de efeitos a estimar do que número de dados. Isso conduz a problemas de estimação para modelar

covariâncias entre níveis de expressão de vários genes, devido ao reduzido número de graus de liberdade. A alternativa a ser adotada refere-se ao uso dos estimadores do tipo shrinkage para os componentes de variância (Cui et al., 2005).

Alguns experimentos podem fazer uso de várias sondas dentro de um arranjo e um modelo ao nível de observações em cada sonda (S) pode ser ajustado para cada gene. Tal modelo pode ser da forma
$y_{ijkm} = u + A_i + D_j + AD_{ij} + S_m + T_k + TS_{km} + e_{ijkm}$, em que o efeito de sonda é aleatório assim como o termo da interação (TS_{km}).

Modelos desse tipo foram empregados por Drost et al. (2008) em eucalipto. Nesse gênero, marcadores genéticos têm sido gerados a partir de dados de expressão gênica. Assim, intensidade de expressão e detecção de polimorfismos de seqüência são obtidos simultaneamente. Duas classes de polimorfismo são obtidas: (i) polimorfismo em seqüências complementares a oligonucleotídeos de genes expressos (SFP-single feature polymorphisms); e (ii) marcadores de expressão gênica, GEM (gene expression markers). A distinção entre SFPs e GMEs a partir da análise de dados de microarranjos permite a rápida obtenção de marcadores SFPs para uso em estudos de associação e implementação da seleção genômica.

Nos testes de significância dos efeitos do modelo os p-valores necessitam ser ajustados quando múltiplos testes são realizados em um experimento, como no caso de milhares de genes testados simultaneamente. Nesse caso, por meio da correção de Bonferroni especifica-se o nível geral de significância desejado α e o divide pelo número n de testes a serem realizados. Tem-se então o nível de significância corrigido $\alpha^* = \alpha/n$ que é utilizado como limite de significância para cada um dos testes. Essa abordagem é conservativa e diminui o poder dos testes. Um critério mais apropriado para esse caso é a taxa de falsos positivos (FDR) definida como a proporção esperada de falsos positivos dentre todos os testes significativos (Rosa et al., 2007).

Os estudos de expressão gênica e a estimação dos efeitos dos marcadores SNPs ou DArTs (no contexto da seleção genômica ampla) possibilitam a caracterização ou determinação das assinaturas moleculares ou genéticas dos caracteres. Isso refere-se à determinação de todo o conjunto de genes que afeta determinada característica fenotípica.

4 Genética de associação (GWAS)

A genética de associação visa determinar os efeitos dos genes (QTL) sobre os valores genéticos dos indivíduos em uma população. Para esse fim, usa como meio as associações entre marcas moleculares e fenótipos. As seguintes associações poderiam ser estudadas:
QTLs e Valores Genéticos: são desconhecidos e são o alvo da GWAS;
Marcas e Fenótipos: são conhecidos e são os meios da GWAS;
Marcas e Valores Genéticos: GWS.
QTLs e Fenótipos: MAS
Marcas e QTLs: mapeamento.
Fenótipos e Valores Genéticos: BLUP tradicional.

As importâncias das associações são apresentadas abaixo:

Genótipos	Valor Fenotípico	Valor Genético
Marcas	+	++
QTL	++	+++

As associações assinaladas com + e ++ podem incorrer no erro de mapear marca como gene e/ou efeito ambiental como efeito genético. Para atingir a associação assinalada como +++ devem ser realizadas as seguintes análises de transformação de marca em QTL e de valor fenotípico em valor genético:

	QTL	Valor Genético
Marcas	p-valor muito baixo (10^{-5}): LDA	-
Valor Fenotípico	-	Segreg Mendeliana Desregressada: Análise de Pedigree

4.1 Coeficientes e Medidas de Desequilíbrio de Ligação

A definição de desequilíbrio de ligação refere-se à associação não aleatória de alelos de diferentes locos. Considere um loco com alelos A e a e outro loco com alelos B e b. O desequilíbrio gamético é dado por D = prob(AB) prob(ab) – prob(Ab) prob(aB), em que prob denota probabilidade ou freqüência dos respectivos haplótipos. Assim, o desequilíbrio existe (D diferente de zero) quando os gametas em associação e repulsão diferem em freqüência. Valores de D positivos revelam que os gametas em associação estão em excesso. Valores de D negativos revelam que os gametas em repulsão estão em excesso. Após t gerações de cruzamentos ao acaso, $D_t = D_o (1 - r)^t$ e, portanto, $t = (\log D_t) / [\log D_o (1 - r)]$ fornece o número de gerações para se atingir o equilíbrio, em que D_o é o desequilíbrio inicial e r é a taxa de recombinação.

Considere as seguintes freqüências alélicas: $p(A) = p_1$; $p(a) = p_2$; $p(B) = q_1$; $p(b) = q_2$. Tem-se então as seguintes igualdades $D = prob(AB)\ prob(ab) - prob(Ab)\ prob(aB) = P_{11} P_{22} - P_{12} P_{21} = p_1q_1\ p_2q_2 - p_1q_2\ p_2q_1 = P_{11} - p_1q_1 = P_{22} - p_2q_2 = p_1q_2 - P_{12} = p_2q_1 - P_{21}$. Assim, os valores máximos e mínimos de desequilíbrio são dados por $Dmax = \min(Ab, aB) = \min(p_1q_2, p_2q_1)$ e $Dmin = \max(AB, ab) = \max(-p_1q_1, -p_2q_2)$.

Como exercício, considere o seguinte: Dois locos com dois alelos estão segregando na população e são fornecidas as seguintes informações: prob (AB) = 0,35; p (A) = 0,7 e p (b) = 0,4. Essa população encontra-se em equilíbrio gamético? Com base nas informações tem-se: p (B) = 1 – 0,4 = 0,6 e a probabilidade esperada de AB é P (AB) = p (A) p(B) = 0,7 x 0,6 = 0,42. Assim D = prob (AB) - p (A) p (B) = $P_{11} - p_1q_1$ = 0,35 – 0,42 = -0,07. Assim, a população encontra-se em desequilíbrio de ligação e existe um excesso de gametas em repulsão. Supondo locos ligados com taxa de permuta de 2 %, o número de gerações para que o desequilíbrio caia à metade (D_t/D_o = 0,5) será dado por $D_t/D_o = (1 - r)^t$ = 0,5. Assim, 0,5 = $(1 - r)^t$ e 0,5 = $(1 - 0,02)^t$ e, portanto, resolvendo para t obtém-se t = 34,31 gerações.

A estatística de desequilíbrio de ligação apresentada por Hill (1981) e usada acima, D = prob(AB) prob (ab) – prob (Ab) prob (aB), é muito dependente das freqüências de alelos individuais e portanto não é útil para comparação do LD entre múltiplos pares de locos envolvendo diferentes pontos ao longo do genoma. A estatística r^2 desenvolvida por Hill e Robertson (1968) é mais adequada, pois é menos dependente das freqüências alélicas. Tal estatística é dada por $r^2 = D^2$ / [prob(A) prob (a) prob (B) prob (b)]. Os valores de r^2 variam de zero (pares de locos com nenhum desequilíbrio entre eles) a 1 (pares de locos com completo LD). Considerando o exemplo acima, têm-se as seguintes freqüências observadas dos haplótipos: P (AB) = 0,35; P (ab) = 0,05; P (aB) = 0,25; p(Ab) = 0,35. Assim, D = P (AB) P(ab) – P (Ab) P(aB) = -0,07 e D^2 = 0,0049. O valor de r^2 é então dado por $r^2 = D^2$ / [prob(A) prob (a) prob (B) prob (b)] = 0,0049 / [0,7 0,3 0,6 0,4] = 0,0972. Esse nível de desequilíbrio é considerado baixo. Valores moderados de r^2 são da ordem de 0,2 ou mais (Hayes et al., 2006).

Outra medida de LD é a estatística D`= módulo (D) / Dmax, proposta por Lewontin (1964), a qual refere-se ao próprio D padronizado pelo D máximo. O Dmáx é dado por Dmax = min (p_1q_2 , p_2q_1) se D > 0 e Dmax = min (p_1q_1 , p_2q_2) se D < 0. Essa medida de LD não é muito precisa pois pode ser inflacionada quando estimada a partir de amostras pequenas ou em situação de baixas freqüências alélicas (McRae et al., 2002). Outra característica de D´ refere-se à sua incapacidade de predição da densidade de marcadores necessária para uma completa varredura do genoma usando LD.

A estatística r^2 é então preferida. O significado genético de r^2 entre um marcador e um QTL não observado é que ele mede a proporção da variação causada por alelos do QTL que é explicada pelos marcadores. Assim, o decréscimo de r^2 com o aumento da distância indica quantos marcadores e fenótipos são necessários para a acurada predição no contexto da seleção genômica ampla e da detecção de QTL usando LD em nível populacional. Os tamanhos amostrais devem aumentar em uma proporção dada por $1/r^2$ para detectar um QTL não observado, em comparação com a amostragem necessária para avaliar o próprio QTL (Pritchard e Przeworski, 2001).

As medidas de desequilíbrio apresentadas referem-se a locos com dois alelos, ou seja, marcadores bi-alélicos. Isto é adequado para marcadores do tipo SNPs, embora possam ser estendidos também para marcadores multi-alélicos como os microssatélites. No entanto, um estimador de desequilíbrio de ligação multi-aléllico

foi proposto por Zhao et al. (2005), por meio da estatística x^{2*} dada por $\chi^{2*} = [1/(m-1)]\sum_{i=1}^{k}\sum_{j=1}^{n}\{D_{ij}^2/[p(a_i)p(b_j)]\}$, em que $D_{ij}^2 = p(a_ib_j) - p(a_i)p(b_j)$ e $p(a_i)$ e $p(b_j)$ são as frequências dos alelos i e j dos marcadores a e b, respectivamente. Por sua vez, $p(a_ib_j)$ refere-se à frequência do haplótipo (a_ib_j). A quantidade m refere-se ao mínimo do número de alelos nos marcadores a e b. A estatística x^{2*} é uma generalização de r^2 e para marcadores bi-alélicos $x^{2*} = r^2$. As simulações realizadas por Zhao et al. (2005) mostraram que x^{2*} é o melhor preditor da proporção da variação causada por alelos do QTL que é explicada pelos marcadores.

A estatística r^2 desenvolvida por Hill e Robertson (1968), dada por $r^2 = D^2 / [\text{prob(A) prob(a) prob(B) prob(b)}]$, tem como esperança ou valor esperado a expressão de Sved (1971), dada por $E(r^2) = 1/(4\ Ne\ r + 1)$. Essa expressão é dada em função da taxa de recombinação r em Morgans. Assim, com base no tamanho efetivo populacional (Ne) e na taxa de recombinação, pode-se inferir sobre o r^2. Inferências sobre o r^2 são importantes no cômputo da acurácia da MAS e da GWS.

Em espécies exogâmicas domesticadas (animais e plantas perenes preferencialmente alógamas) o reduzido tamanho efetivo populacional é a principal causa de desequilíbrio de ligação. Nesse caso, o valor esperado desse desequilíbrio em um dado segmento cromossômico de tamanho S (em Morgans) pode ser calculado pela seguinte expressão $E(r^2) = 1/(4\ Ne\ S + 1)$. Pela equação de Sved, verifica-se que o desequilíbrio de ligação reduz-se rapidamente com o aumento da distância entre os genes, ou seja, com o aumento do tamanho do segmento considerado. Essa redução é tanto maior quanto maior for o tamanho efetivo populacional (Tabela 21).

Tabela 21. Valores esperados ($E(r^2)$) do desequilíbrio de ligação entre dois locos, em função do tamanho efetivo populacional (Ne) e do comprimento (L) do segmento cromossômico entre os dois locos.

Ne	S (Morgan)	S (CentiMorgan)	$E(r^2)$	Ne	S (Morgan)	S (CentiMorgan)	$E(r^2)$
10	0.005	0.5	0.83	100	0.005	0.5	0.33
10	**0.01**	**1**	**0.71**	**100**	**0.01**	**1**	**0.20**
10	0.02	2	0.56	100	0.02	2	0.11
10	0.03	3	0.45	100	0.03	3	0.08
10	0.04	4	0.38	100	0.04	4	0.06
10	0.05	5	0.33	100	0.05	5	0.05
20	0.005	0.5	0.71	200	0.005	0.5	0.20
20	**0.01**	**1**	**0.56**	**200**	**0.01**	**1**	**0.11**
20	0.02	2	0.38	200	0.02	2	0.06
20	0.03	3	0.29	200	0.03	3	0.04
20	0.04	4	0.24	200	0.04	4	0.03
20	0.05	5	0.20	200	0.05	5	0.02
30	0.005	0.5	0.63	500	0.005	0.5	0.09
30	**0.01**	**1**	**0.45**	**500**	**0.01**	**1**	**0.05**
30	0.02	2	0.29	500	0.02	2	0.02
30	0.03	3	0.22	500	0.03	3	0.02
30	0.04	4	0.17	500	0.04	4	0.01
30	0.05	5	0.14	500	0.05	5	0.01
50	0.005	0.5	0.50	1000	0.005	0.5	0.05
50	**0.01**	**1**	**0.33**	**1000**	**0.01**	**1**	**0.02**
50	0.02	2	0.20	1000	0.02	2	0.01
50	0.03	3	0.14	1000	0.03	3	0.01
50	0.04	4	0.11	1000	0.04	4	0.01
50	0.05	5	0.09	1000	0.05	5	0.00

Verifica-se pela Tabela 21, para os tamanhos efetivos praticados no melhoramento de plantas perenes (30 a 100), que os desequilíbrios de ligação adequados (maiores ou iguais a 0,2) para a seleção de QTLs são obtidos com marcadores espaçados de 1 a 3 cM. O r^2_{mq} ou E(r^2) é uma média ponderada do r^2 de cada par marcador-QTL. O r^2 é o quadrado da correlação (r) entre alelos ou genótipos presentes no loco marcador e no loco do QTL (Tabela 22).

Tabela 22. Cálculo do desequilíbrio de ligação entre marcador e QTL.

Indivíduo	N. Alelos Loco Marcador (W_a)	N. Alelos Loco QTL (W_b)
1	0	0
2	2	1
3	1	1
4	1	0
5	2	1
Correlação r	r = 0.76	r^2= 0.58

O r^2 tem então três interpretações: (i) desvio da frequência observada de haplótipos em relação à esperada segundo segregação independente ($D = \Pr ob(ab) - \Pr ob(a)\Pr ob(b)$); (ii) quadrado da correlação (r) entre alelos (Tabela 22); (ii) proporção da variação no QTL explicada pelo marcador. As provas dessas três interpretações e equivalências são apresentadas a seguir.

O coeficiente de correlação entre duas variáveis ou alelos nos locos a e b é 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 W dos marcadores o valor de r pode ser dado por $r_{(a,b)} = \frac{Cov(W_{ia}, W_{ib})}{[Var(W_{ia})]^{1/2}[Var(W_{ib})]^{1/2}}$. Definem-se as quantidades $D = \Pr ob(ab) - \Pr ob(a)\Pr ob(b)$, 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.

A relação entre efeitos genéticos do marcador e do QTL pode ser melhor entendida segundo os modelos a seguir: modelo para fenótipo via efeito genético do QTL (g_{QTL}): $y = u + g_{QTL} + e$; modelo para fenótipo via efeito genético do marcador (g_m): $y = u + g_{QTL} + e = u + Wg_m + e$. A quantidade g_m é um coeficiente de regressão dado por $g_m = Cov(y, W)/Var(W) = Cov(g_{QTL}, W)/Var(W)$
$= r[Var(g_{QTL})/Var(W)]^{1/2} = r\{Var(g_{QTL})/[2p(1-p)]\}^{1/2}$.

A quantidade da variação no QTL explicada pelo marcador é dada por $Var(Wg_m) = g_m^2 Var(W) = r^2[Var(g_{QTL})/Var(W)]Var(W) = r^2 Var(g_{QTL})$. Assim, surge o conceito de r^2 como a proporção da variação do QTL explicada pelo marcador.

A extensão do desequilíbrio de ligação depende de recombinações recentes e também antigas bem como do Ne atual e passado. Populações domesticadas de plantas e animais apresentam menor Ne atual do que Ne passado. Em humanos ocorre o contrário, devido ao grande aumento populacional na era atual. Hayes et al. (2003) relatam que o desequilíbrio de ligação em segmentos cromossômicos curtos (em distâncias pequenas) depende do tamanho efetivo antigo da população, muitas gerações atrás. Por outro lado, o desequilíbrio a longa distância depende da recente história da população. Considerando que mudanças lineares nos tamanhos das populações são realizadas, tem-se que a medida de desequilíbrio r^2 é reflexo do Ne a $1/(2S)$ gerações atrás. Dessa forma, o valor esperado de r^2 quando o Ne é alterado através das gerações é dado por $E(r^2) = 1/(4\ Ne_t\ S + 1)$, em que $t = 1/(2S)$. Em humanos, o Ne equivale a aproximadamente 10.000 (Kruglyak, 1999). Em animais domésticos e plantas perenes, o Ne pode ser baixo, da ordem de 100. Assim, o desequilíbrio de ligação (LD) deveria ser menor nos humanos. No entanto, no passado, o Ne da população humana foi baixo. Assim, a longas distâncias entre marcadores, os valores de r^2 em humanos são menores do que nas espécies domésticas de plantas e animais. E a curtas distâncias entre marcadores os valores de r^2 são mais similares em humanos e em espécies animais domesticadas. Moderado LD (r^2 maior ou igual a 0,2) em humanos estende a menos que 5 kb ou 0,005 cM. Em gado bovino, moderado LD estende-se até 100 kb. Entretanto, valores muito altos (r^2 maior ou igual a 0,8) de LD estendem-se apenas a distâncias muito curtas tanto em humanos quanto em gado bovino (Tenesa et al., 2007).

Em gado bovino leiteiro, populações holandesas e australianas apresentam declínio em LD similar, pois são populações relacionadas por origem e apresentam história e Ne semelhantes. Por outro lado, a população bovina norueguesa vermelha (Ne igual a 400) apresenta mais rápido declínio em LD do que o gado bovino leiteiro holandês (Ne global igual a 150). Os diferentes Ne justificam esse comportamento diferenciado do LD nas duas populações (Zenger et al., 2007). Considerações importantes sobre cálculos de tamanho efetivo populacional são apresentadas por Vencovsky e Crossa (1999).

4.2 Métodos de Análise de QTL via LDA

Durante muito tempo, os estudos de mapeamento basearam-se na análise de ligação associada aos dados de *pedigree*. Recentemente, métodos baseados em desequilíbrio de ligação associados a indivíduos não aparentados têm sido recomendados como ferramentas poderosas para produzir estimativas refinadas da localização de genes. Tais métodos são baseados nas seguintes premissas. Quando um novo alelo é introduzido na população, seja por mutação ou migração, este passa a existir na população conjuntamente com um grupo de alelos marcadores. O comprimento desse haplótipo é reduzido ao longo das gerações devido a eventos de recombinação e, depois de muitas gerações, somente os marcadores na vizinhança imediata do loco do novo alelo provavelmente permanecerão no mesmo segmento cromossômico. Se esse alelo influencia determinado caráter, uma correlação de alta magnitude entre o caráter e o alelo marcador deverá indicar que o loco que codifica o caráter situa-se muito próximo ao marcador.

O mapeamento via LDA visa aumentar a precisão da estimativa da posição do QTL, pois, em algumas situações, o número de meioses associadas ao *pedigree* genotipado não é suficiente para que a LA seja precisa. Os métodos LDA propiciam um mapeamento fino, o qual fundamenta-se na quantificação do desequilíbrio de ligação em fase gamética presente através das famílias em uma população alógama. Nesse caso, a fase de ligação não varia entre famílias e nem entre gerações. A base do método refere-se ao fato de que quando uma população é pequena, os fundadores terão um pequeno número de diferentes haplótipos e, com locos intimamente ligados, não haverá tempo suficiente para a recombinação quebrar a associação entre marcadores e a mutação que afeta o QTL (Perez-Enciso et al., 2003). A mutação funcional é referida como nucleotídeo de característica quantitativa (QTN).

Tal mapeamento é também denominado mapeamento de associação, o qual tornou possível com o advento dos marcadores SNPs e DArTs, que permitem uma alta densidade de marcadores no genoma. Os marcadores SNP são codominantes e bialélicos, embora raramente (menos que 1%) sejam encontrados SNPs trialélicos ou tetralélicos, pois, a plausibilidade de ocorrência de uma segunda mutação na mesma posição do nucleotídeo é muito pequena. Os marcadores DArT podem ser dominantes ou codominantes. A estratégia de associação caráter-marcador em nível populacional baseia-se em pequenos blocos gênicos em desequilíbrio de ligação e, portanto, a resolução é muito grande (menores distâncias entre genes). Embora a resolução seja maior, a detecção de QTLs e a precisão do mapeamento demandam um número muito grande de marcadores. O mapeamento de associação opera na população em geral e não especificamente em uma população de mapeamento. A associação entre marcador e QTL depende da freqüência de recombinação entre eles. Para encontrar um marcador razoavelmente próximo a um QTL é necessário uma baixa freqüência de recombinação. Quanto maior o desequilíbrio de ligação, mais próximo o marcador estará do gene e esse LD ou associação serão válidos mesmo para indivíduos geneticamente mais distantes.

Duas abordagens podem ser usadas na genética ou mapeamento de associação: varredura genômica e genes candidatos. Nessa última abordagem, marcadores são usados apenas dentro de genes candidatos individuais. Para a genética de associação, a população de mapeamento deve ser grande e com alto grau de desequilíbrio de ligação. O mapeamento via LDA baseia-se em varredura genômica usando mapa de marcadores de alta densidade, com um marcador a cada 0,5 cM a 2 cM. O sucesso do método depende da extensão do desequilíbrio de ligação na população. Uma vez que os marcadores podem não estar em completo LD com os QTLs, tanto as associações entre marcadores e QTLs na população quanto a co-segregação de marcadores e QTLs dentro de famílias podem ser usados simultaneamente na detecção de QTL, via o método LDA-LA, o qual combina as propriedades dos marcadores LD (em desequilíbrio de ligação) e LE (em desequilíbrio de ligação), respectivamente.

O mapeamento baseado em LDA é conduzido por meio do cálculo das probabilidades de que os haplótipos compartilhados pelos indivíduos sejam idênticos por descendência de um ancestral comum, condicional aos dados de marcadores. A correta determinação das fases de ligação e dos genótipos do QTL é necessária no

mapeamento fino. Assim, uma pura análise LDA pode resultar em um alto número de falsos positivos, ou seja, falsa inferência de associação em ausência de ligação. Em função disso, métodos (LA-LDA) que incorporam simultaneamente as informações de LD populacional e de ligação dentro de famílias são indicados, visando mitigar os efeitos da associação espúria entre marcadores e QTLs (Meuwissen e Goddard, 2004).

A seleção auxiliada por marcadores moleculares (MAS) e a seleção genômica ampla (GWS) serão tanto mais efetivas quanto mais próximos estiverem os marcadores dos QTLs. Dado o pequeno espaçamento entre genes no cromossomo, o mapeamento de QTLs com precisão é uma tarefa cruel. Em média, um segmento cromossômico de 10 cM pode conter cerca de 200 genes. Assim, uma alta densidade de marcadores genotipados aumenta a resolução do mapeamento de QTLs. Mas se o objetivo for encontrar o próprio gene que afeta o caráter, o intervalo de confiança para a localização do QTL ainda será amplo, mesmo para QTL de grande efeito e com grande tamanho amostral (Weller, 2001). Estratégias de mapeamento baseadas em LDA são relatadas a seguir.

4.3 Mapeamento genômico amplo via regressão em marcas únicas

A GWAS (Genome Wide Association Studies) procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses visando detectar efeitos com significância estatística. O seguinte modelo de regressão em marcas simples pode ser empregado visando à associação entre marcador e QTL em uma população panmítica (Resende, 2008): $y = Ju + Wm_i + e$, em que y é o vetor de observações fenotípicas, J é um vetor com valores 1, u é o escalar referente à média geral, m_i é o efeito fixo de um dos alelos do marcador bialélico e e refere-se ao vetor de resíduos aleatórios. W é a matriz de incidência para m_i. Esse modelo assume que o marcador afetará o caráter apenas se ele estiver em LD com o suposto QTL. Outros efeitos fixos e aleatórios podem ser incorporados nesse modelo. Como exemplo, considere a avaliação de 12 indivíduos para um caráter e para um marcador do tipo SNP. Os dados referentes aos genótipos e fenótipos dos indivíduos são apresentados a seguir.

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

A matriz de incidência W associa os números de cada alelo do SNP aos fenótipos. É suficiente ajustar o efeito de apenas um dos alelos. Assim, a matriz W terá apenas uma coluna para o efeito de um dos alelos do SNP, por exemplo o A. Essa coluna contém o número de cópias do alelo A que os indivíduos possuem.

Portanto, contém os valores 0, 1 ou 2 para um indivíduo diplóide. O número de linhas dessa matriz é igual ao número de indivíduos.

A matriz J inclui uma coluna para a média geral. As matrizes J e W (número de alelos A), apresentadas na forma transposta são dadas por $J'_{(12x1)}=[1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$ e $W'_{(12x1)}=[1\ 2\ 1\ 2\ 1\ 0\ 2\ 1\ 0\ 2\ 1\ 2]$. As equações de quadrados mínimos para a estimação dos efeitos da média geral e do SNP equivalem a:

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

A hipótese da nulidade, ou seja, de que o marcador não apresenta qualquer efeito sobre o caráter, pode ser avaliada pelo teste F. A hipótese nula é rejeitada se $F > F(a,v_1,v_2)$, em que F é a estatística de Snedecor calculada dos dados, *a* é o nível de significância e v_1 e v_2 são os graus de liberdade associado à distribuição F tabelada. A hipótese alternativa é de que o marcador afeta o caráter, ou seja, devido ao fato de que marcador e QTL encontram-se em desequilíbrio de ligação. O valor da estatística F é calculado via $F=\dfrac{QM\ \text{Re}\,gressão}{\hat{\sigma}_e^2}=\dfrac{\hat{m}W'y+\hat{u}\ J'y-(1/n)\ (J'y)^2}{(y'y-\hat{m}W'y-\hat{u}J'y)/(n-2)}$.

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

Um modelo mais completo é da forma:

$y=Xb+Ts+Wm_i+Zg+e$, que b, s e g são vetores de efeitos fixos de natureza ambiental, de covariável de efeitos fixos referente à estrutura de população e de efeitos aditivos poligênicos (aleatórios), respectivamente, com matrizes de incidência X, T e Z.

Os valores da covariável associada a s podem ser tomados como os autovetores decorrentes da decomposição espectral da matriz de parentesco genômico G. Maiores detalhes no capítulo 6.

4.4 Poder estatístico e significância na associação e detecção de QTL

O poder do teste de associação marcador-QTL depende dos seguintes fatores (Pritchard e Przeworski, 2001; Meuwissen et al., 2002; Hayes et al., 2006; Fernando et al., 2004; Macleod et al., 2007):

(i) Do r^2 (medida estatística do desequilíbrio de ligação) entre marcador e QTL. O significado genético de r^2 entre um marcador e um QTL não observado é que ele mede a proporção da variação causada por alelos do QTL que é explicada pelos marcadores. Os tamanhos amostrais devem aumentar em uma proporção dada por $1/r^2$ para detectar um QTL não observado, em comparação com a amostragem necessária para avaliar o próprio QTL.

(ii) Da proporção da variação fenotípica explicada pelo QTL, ou seja, do coeficiente de determinação do efeito do QTL ($h_q^2 = \sigma_q^2 / \sigma_{total}^2$).

(iii) Do número *n* de indivíduos avaliados.

(iv) Do nível de significância especificado.

(v) Da freqüência *p* do alelo raro do marcador, a qual determina o número mínimo de observações necessárias para estimar um efeito alélico. Se *p* é menor do que 0,1, o poder torna-se sensível a essa freqüência alélica.

O **poder** de um teste refere-se à probabilidade de se rejeitar H_o, quando H_o é falsa, ou seja, à capacidade de detectar um QTL na população, quando ele realmente existe. O poder de um teste de detecção de QTL em função de diferentes níveis de r^2 entre o QTL e o marcador pode ser calculado pela fórmula de Luo (1998). Para conseguir um poder maior ou igual a 80 %, visando à detecção de um QTL com h_q^2 igual a 0,05 com base em 1.000 observações fenotípicas, é necessário um r^2 de pelo menos 0,2. Esse resultado considerou a frequência do alelo raro maior do que 0,2.

Macleod et al. (2007) relataram que o poder de detecção de QTL com h_q^2 igual a 5 % e 365 indivíduos genotipados foi de 37 % ($p<0,001$). Verificaram também uma forte correlação entre os valores de F associados a SNPs significativos e seus r^2 com o QTL. A correlação entre as estatísticas F de Snedecor e D' foram praticamente zero.

Ao fazer uma inferência, o pesquisador incorre no erro tipo I, quando rejeita uma hipótese H_o que é verdadeira e incorre no erro tipo II, quando aceita uma hipótese H_o que é falsa. A probabilidade de cometer um erro tipo I é designada por α e o maior valor de α para H_o verdadeira é denominado **nível de significância** de um teste estatístico, ou seja, a significância de um teste é a probabilidade máxima que se admite correr o risco de cometer um erro tipo I.

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

A técnica do teste de permutação foi proposta por Churchill e Doerge (1994) para contornar a questão de múltiplos testes nos experimentos de mapeamento de QTL. Essa técnica é apropriada para estabelecer os adequados níveis de significância. Hoggart et al. (2008) derivaram uma aproximação explícita para o erro tipo I a qual evita a necessidade de procedimentos de permutação. Outra alternativa para evitar falsos positivos é monitorar esse número em relação ao número de resultados positivos, conforme Fernando et al. (2004). O pesquisador pode estabelecer um nível de significância associado a uma proporção aceitável de falsos positivos.

A taxa de descobertas falsas (FDR) é definida como a proporção esperada de QTLs detectados que são falsos positivos. A FDR pode ser calculada como FDR = m Pmax/n, em que Pmax é o maior Pvalor de QTL que excede o nível de significância, *n* é o número de QTLs que excedem o nível de significância e *m* é o número de marcadores testados (Weller, 2001). Com 10 mil SNPs testados, nível de significância (Pvalor) de 0,001 e 80 SNPs declarados como significativos, a FDR = 10.000 x 0,001/80 = 0,125. Essa magnitude (12,5 %) de taxa de falsa descoberta pode ser considerada aceitável.

Uma alternativa para diminuir a taxa de falsos positivos é a adoção de modelo com inclusão do vetor de efeitos poligênicos, o qual contempla a matriz de parentesco e permite correção para estrutura de população. Macleod et al. (2007) relatam um aumento no número de falsos positivos (erros tipo 1) quando os efeitos poligênicos não são incluídos no modelo. Nesse caso, o uso dos próprios marcadores é indicado para inferir sobre a matriz de parentesco, conforme Hayes et al. (2007). Para um dado loco marcador, a similaridade genética Sxy entre dois indivíduos x e y, é calculada da seguinte forma:

(a) Sxy = 1, quando o genótipo x = ii (ambos alelos no loco são idênticos) e o genótipo y = ii, ou quando x = ij e y = ij;

(b) Sxy = 0,5, quando o genótipo x = ii e o genótipo y = ij, ou vice-versa;

(c) Sxy = 0,25, quando o genótipo x = ij e o genótipo y = ik;

(d) Sxy = 0, quando os dois indivíduos não têm alelos comuns no loco.

A similaridade resultante do acaso é dada por $Sa = \sum_{i=1}^{g} p_i^2$, em que *p* é a freqüência do alelo na população e *g* é o número de alelos no loco. O parentesco entre os

indivíduos x e y no loco é então calculado como r = (Sxy – Sa)/(1 – Sa). O parentesco médio entre os indivíduos é então computado como a média de *r* sobre todos os locos. Com grande número de marcadores, a matriz de parentesco derivada de marcadores pode capturar os efeitos da segregação mendeliana.

Para a estimação de intervalos de confiança em estudo de associação genômica ampla, métodos baseados em validação cruzada podem ser usados. Nesse caso, o conjunto de dados é dividido em duas partes e o estudo de associação é realizado três vezes, uma vez em cada metade dos dados e uma vez no conjunto total de dados. O intervalo a 95 % de confiança associado à posição do QTL é dado pela posição do SNP mais significativo na análise com os dados completos ± 1,96 s, em que *s* é o erro padrão do QTL e é dado por $s = (\frac{1}{4n}\sum_{i=1}^{n} x_{1i} - x_{21})^{1/2}$ para *n* pares de SNPs com efeitos significativos. Os componentes $x_{1i}\ e\ x_{21}$ são as posições do SNP mais significativo em cada uma das metades do dado completo, para a i-ésima posição mais significativa do QTL no conjunto total de dados. Isto é válido quando a análise de cada metade dos dados confirma um SNP declarado como significativo na análise com os dados completos.

4.5 Mapeamento genômico amplo via modelos mistos com haplótipos

Haplótipos são determinadas combinações de múltiplos marcadores ligados e podem ser considerados como alelos de um "supraloco". Podem ser usados em lugar de marcas simples nos estudos de associação genômica ampla. Apresentam a vantagem de poder estar em maior desequilíbrio de ligação com os QTLs. Quando isso acontece, o r^2 é maior e, portanto, o poder do experimento é aumentado. A proporção da variância do QTL explicada pelos marcadores pode ser calculada da seguinte forma (Hayes et al., 2006): Sendo q_1 e q_2 as freqüências dos dois alelos do QTL, os marcadores podem ser classificados em *n* haplótipos, com freqüência p_i para o i-ésimo haplótipo. Isto pode ser representado em uma tabela de contingência:

	Haplótipos			
	1	i	n	**Totais**
Alelo 1 do QTL	$p_1q_1 - D_1$	$p_iq_1 - D_i$	$p_nq_1 - D_n$	Q_1
Alelo 2 do QTL	$p_1q_2 + D_1$	$p_iq_2 + D_i$	$P_nq_2 + D_n$	Q_2
Totais	p_1	p_i	p_n	1

Para um haplótipo *i* representado nos dados, o desequilíbrio de ligação é calculado por $D_i = p_i(q_1) - p_iq_1$, em que $p_i(q_1)$ é a proporção de haplótipos *i* no conjunto de dados, que carregam o alelo 1 do QTL (observado dos dados), p_i é a proporção de haplótipos *i* e q_1 é a freqüência do alelo 1 do QTL. A proporção da variância do QTL explicada pelos haplótipos e corrigida para os efeitos de amostragem pode ser calculada por $r^2(h,q) = \frac{1}{q_1q_2}\sum_{i=1}^{n} D_i^2 / p_i$. Assim, r^2 depende do LD, da freqüência do haplótipo e das freqüências dos alelos do QTL. Valores de r^2 podem ser obtidos via simulação de diferentes freqüências q_1 e q_2 e tamanhos de genoma e

haplótipos. Quanto maior o tamanho efetivo populacional, menor proporção da variação genética será explicada pelos haplótipos.

O seguinte modelo linear misto geral é usado para estimar os efeitos de haplótipos:$y = J'u + Wh + Zg^* + e$, em que y é o vetor de observações fenotípicas, u é o escalar da media (de efeito fixo), J é um vetor de uns, h é o vetor dos efeitos aleatórios de haplótipos (intervalos), g^* é o vetor de efeitos poligênicos (aleatório) e e refere-se ao vetor de resíduos aleatórios. W e Z são as matrizes de incidência para h e g^*. Os efeitos de haplótipos devem ser tratados preferencialmente como aleatórios porque eles são em grande número e alguns deles ocorrem em um número limitado de vezes (nesse caso, esses haplótipos com pequeno número de observações devem ser penalizados pelo efeito de shrinkage).

A dimensão de h é igual ao número de intervalos multiplicado por 4 (número de haplótipos possíveis para cada intervalo entre duas marcas). A matriz de incidência W contém os valores 0, 1 e 2 para o número de alelos (do suposto QTL) ou haplótipos do tipo h_i em um indivíduo diplóide. Detalhes algébricos desse modelo são apresentados por Resende (2008). A variação genética aditiva $\sigma^2_{g^*}$ e a dos haplótipos σ^2_h podem ser estimadas por REML sobre os dados fenotípicos e pela própria variação entre os haplótipos ou variância dos segmentos cromossômicos. A significância dos efeitos de haplótipos é avaliada via teste da razão de verossimilhança. Para o mapeamento, o ajuste do modelo descrito enfatiza a estimação do componente de variância σ^2_h e o teste de sua significância via LRT. Não há interesse especificamente nos efeitos BLUP de h, os quais são enfatizados e utilizados na MAS.

4.6 Mapeamento genômico amplo via abordagem IBD-LD

No mapeamento via IBD-LD, o efeito do suposto QTL é incluído no modelo e não o efeito do marcador ou do haplótipo. A informação dos haplótipos é usada para inferir sobre a probabilidade de que dois indivíduos possuem o mesmo alelo do QTL em uma suposta posição. A ocorrência de LD revela que existem pequenos segmentos de cromossomo na população os quais descendem de um mesmo ancestral comum. Esses cromossomos são então idênticos por descendência (IBD) e carregam idênticos haplótipos marcadores e também alelos do QTL. Indivíduos com esses segmentos cromossômicos IBD terão seus fenótipos correlacionados.

Nesse caso, o modelo a ser ajustado é o seguinte: $y = Xb + \sum_j Q_j q_j + Zg^* + e$, em que:

q_j : vetor que contém duas incógnitas para cada indivíduo em cada loco (um efeito do QTL no cromossomo maternal e outro no paternal); g*: vetor aleatório de efeitos poligênicos, excluindo q; Q_j : matriz de incidência para os alelos do QTL no segmento cromossômico j; Z : matriz de incidência para g*; b e e_i : vetor de efeitos fixos e erro aleatório, respectivamente; X : matriz de incidência para b. Esse modelo, que inclui ambos, os efeitos do QTL e poligênico infinitesimal, é denominado modelo misto de herança (Fernando et al., 1994). Detalhes algébricos desse modelo são apresentados por Resende (2008). O mapeamento dos QTLs é realizado com base

na estimação das variâncias $\sigma^2_{g^*}$ e σ^2_q. Essas são estimadas por REML. O mapeamento prossegue então propondo uma suposta posição para o QTL em intervalos ao longo do cromossomo. Em cada ponto, a variância do QTL é estimada e a verossimilhança dos dados, dada a posição do QTL e a variância poligênica, é calculada e verificada quanto ao seu máximo. Assim, a presença de um QTL em uma particular posição no cromossomo pode ser testada pelo LRT, comparando a verossimilhança de dois modelos, um com a inclusão e outro sem a inclusão do QTL.

4.7 Mapeamento genômico amplo via abordagem LDA-LA

O modelo a ser ajustado nesse caso difere do modelo apresentado no tópico anterior apenas na forma de construção da matriz IBD. A combinação das informações de ligação e de LD é interessante, visando minimizar os efeitos de associação espúria. Isso produz o método LDA-LA de mapeamento, o qual é poderoso para filtrar picos espúrios de verossimilhança obtidos nas análises isoladas LDA e LA. Nesse método, a matriz IBD é composta de duas partes: uma submatriz (bloco [a]) que descreve os coeficientes IBD entre haplótipos dos indivíduos fundadores e fornece informação sobre LD; uma submatriz (bloco [b]) que descreve a transmissão dos alelos do QTL dos indivíduos fundadores para as gerações atuais dos indivíduos genotipados e fornece informação sobre a ligação (LA). Meuwissen et al. (2002) descrevem a obtenção da matriz IBD para o método LDA-LA associado a um delineamento de progênies de meios irmãos. De posse da matriz IBD, um modelo de componente de variância similar ao descrito no tópico anterior pode ser ajustado.

4.8 Mapeamento genômico amplo via abordagem GWS

Embora a GWS atue sobre todos os genes de um caráter quantitativo, os marcadores com os efeitos estimados maiores podem ser considerados como supostamente ligados a QTLs. Assim, apesar da GWS não ser um processo de descoberta de genes, a mesma pode ser usada para o mapeamento de QTL.

4.9 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) 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, 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 (GWAS – SE) 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 (SE) de maneira similar ao método da GWS. É baseado em regressão aleatória para a predição de efeitos latentes de QTL. 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), nem sempre declarados como IBD, dada a genealogia usada.

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 $\sigma^2_{gi} = 2p_i(1-p_i)m_i^2$, 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 $\sigma^2_g = \sum_i 2p_i(1-p_i)m_i^2$.

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 $(\hat{e} = y - X\hat{b} - Z\hat{g})$, após o ajuste para g sob modelo poligênico infinitesimal (ajuste ou eliminação dos efeitos de família ou variação entre pedigrees ou estrutura de população). 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 $g \sim N(0, A\sigma^2_g)$.

4.10 Captura da h² em humanos e imperfeito LD entre SNPs e variantes causais

Visscher et al. (2010) abordam os resultados da GWAS referente ao caráter altura em humanos. A h² 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 r² 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 h² 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 r² pode atingir é fortemente determinado pelas frequências alélicas nos dois locos (Wray, 2005). Quanto mais diferentes as frequências alélicas, menor o valor de r². Assim, como a maioria dos SNP genotipados são comuns, se os variantes são raros, r² será baixo e, então a variação σ^2_{mi} associada aos SNP é substancialmente menor que a variação σ^2_{gi} no QTL (Visscher et al., 2010). As expressões $r^2 = \sigma^2_{mi} / \sigma^2_{gi}$ e $\sigma^2_{mi} = r^2\sigma^2_{gi}$ ilustram essa questão.

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 alto 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 (via RR-BLUP) 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 da 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) e também o método regressão via componentes independentes (ICR) (Azevedo et al., 2012).

4.11 GWAS via BayesCpi e BayesDpi

Os métodos BayesCπ e BayesDπ (descritos por Habier et al., 2011; Resende et al., 2011) 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. São vantajosos em relação à análise de regressão marcas únicas devido ao fato de considerar simultaneamente todas as marcas.

No método BayesC uma variância comum é especificada para todos os locos. O método BayesDπ mantém variâncias específicas para cada loco. Adicionalmente, π é tratada como uma incógnita com distribuição *a priori* uniforme (0,1) produzindo o métodos BayesCπ e BayesDπ. 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 $g_j = \sum_{i=1}^{n} m_i w_{ij} \delta_i$, 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, que devidamente especificada transforma-se em uma Uniforme (0,1) (Legarra et al., 2011).

As quantidades w_{ij} são elementos do 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 i, e m_i é definido como elementos do 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.

5 Seleção Auxiliada por Marcadores Moleculares (MAS)

5.1 Tipos de seleção via marcadores genéticos

Existem quatro tipos de seleção que empregam marcadores moleculares: (i) seleção auxiliada por genes conhecidos (GAS), baseada em mutação funcional e genes com efeitos conhecidos, ou seja, os marcadores são os próprios genes; (ii) seleção auxiliada por marcadores em equilíbrio de ligação com QTLs na população (LE-MAS), mas em desequilíbrio de ligação dentro de famílias e cruzamentos; (iii) seleção auxiliada por marcadores em desequilíbrio de ligação com QTLs em nível populacional (LD-MAS); (iv) seleção genômica (GS) ou seleção genômica ampla (GWS), baseada em milhares de marcadores em desequilíbrio de ligação populacional com todos os QTLs de um caráter poligênico. Na GWS, não há necessidade de uso das informações fenotípicas na população de seleção e nem do conhecimento e detecção de QTLs individuais baseados em significância estatística com arbitrários níveis de significância. São usados fenótipos apenas na população de descoberta ou de estimação dos efeitos dos vários locos, via marcadores. A GWS é mais um tipo de seleção auxiliada pelo fenótipo (PAS) do que um tipo de MAS, pois os fenótipos são mais usados como auxílio em uma seleção baseada essencialmente em genótipos marcadores, cujos efeitos foram estimados previamente em uma amostra da população de seleção.

Os tipos LD-MAS e GWS tendem a ser mais eficientes. A LE-MAS, aplicada em nível populacional, requer uma genotipagem muito intensa e procedimentos estatísticos complexos, conforme proposto por Wang, Fernando e Grossman (1998). Em populações com equilíbrio de ligação entre marcadores e QTL, a informação usada na seleção advém da co-segregação entre marcadores e QTL dentro de cada família na população de seleção. Assim, a co-ancestria condicional à informação de marcadores precisa ser computada dentro de cada família para um dado segmento no genoma. Nesse caso, a acurácia da seleção usando marcadores depende principalmente da proporção da variação dentro de família que é devida ao QTL.

A LE-MAS requer grande quantidade de genotipagem e também avaliações fenotípicas em todos os candidatos à seleção, pois nem todos os locos que controlam o caráter são amostrados pela genotipagem. Devem, então, ser amostrados pela fenotipagem. Requer também que a fase de ligação entre marcadores e QTL seja re-estimada em cada geração. Isto torna a LE-MAS mais onerosa do que a seleção tradicional baseada em fenótipos. O presente capítulo aborda os vários tipos de seleção via marcadores baseando-se nos textos publicados por Resende (2007; 2008) e Resende et al. (2008; 2010; 2011).

5.2 Seleção em genes de efeitos conhecidos ou marcadores diretos (GAS)

No contexto dos modelos mistos, a inclusão de genes de efeitos conhecidos na avaliação genética pode ser feita segundo o seguinte modelo, conforme Kennedy et al. (1992): $y = Xb + Zg* + Qq + e$, em que: b : vetor de efeitos fixos; g*: vetor aleatório de efeitos poligênicos, excluindo *q*, ou seja, corrigidos para *q;* q : vetor de efeitos genotípicos (fixos) dos genes conhecidos; referem-se aos efeitos dos genótipos observáveis em um único loco; e : vetor erros aleatórios;X : matriz de incidência para *b*; Z : matriz de incidência para *g**; Q : matriz de incidência para os efeitos do gene conhecido.

Esse modelo tem a seguinte estrutura de médias e variâncias:

$E(y) = Xb + E(Qq)$; $Var(y) = Var(Qq) + ZVar(g*)Z' + Var(e)$. Assumindo os genótipos do QTL como de efeitos fixos, tem-se $E(Qq) = Qq$ e $Var(Qq) = 0$. Assim, $Var(y) = ZVar(g*)Z' + Var(e)$.

Assim, as equações de modelo misto são dadas por:

$$\begin{bmatrix} X'X & X'Z & X'Q \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'Q \\ Q'X & Q'Z & Q'Q \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}* \\ \hat{q} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Q'y \end{bmatrix}$$, em que: $\lambda_1 = \frac{\sigma_e^2}{\sigma_{g*}^2}$; A : matriz de parentesco genético aditivo; σ_{g*}^2: variância aditiva poligênica, ajustada para os efeitos dos genes conhecidos; σ_e^2: variância residual. A seleção dos indivíduos é baseada em $\hat{g} = \hat{g}* + \hat{q}$.

O presente modelo considera que o marcador é o próprio gene que afeta o caráter. Se o marcador não é o próprio gene, mas encontra-se ligado a ele, a recombinação entre eles pode conduzir ao fato de que um mesmo alelo do marcador carregue diferentes alelos do QTL. Nesse caso, o efeito do gene ou QTL pode ser considerado como aleatório e a probabilidade de identidade por descendência dos alelos do QTL pode ser calculada a partir dos genótipos marcadores, desde que a freqüência de recombinação entre marcador e QTL seja conhecida. Isto produz o modelo de Fernando e Grossman (1989), abordado no tópico seguinte.

5.3 MAS via marcadores em equilíbrio de ligação (LE-MAS)

A MAS surgiu basicamente na década de 1990. Os primeiros trabalhos foram o de Fernando e Grossman (1989), Lande e Thompson (1990), Goddard (1992). Com marcadores LE, as probabilidades de identidade por descendência associadas ao QTL, derivadas com base nos genótipos marcadores, serão afetadas pela taxa de recombinação entre marcador e QTL e pela extensão do desequilíbrio de ligação entre eles. Como o LD existe apenas dentro de família, os efeitos de marcadores e a fase de ligação marcador-QTL deve ser determinada separadamente para cada família.

Para a seleção em populações em equilíbrio, Fernando e Grossman (1989)

desenvolveram um procedimento BLUP para a seleção auxiliada por marcadores, o qual se baseia no desequilíbrio de ligação dentro de famílias. O método pode ser usado para a predição dos valores genéticos de todos os indivíduos da população, incluindo os efeitos de QTL via ligação com marcadores genéticos, desde que todos os indivíduos sejam genotipados e a herdabilidade e a freqüência de recombinação entre o QTL e marcador sejam conhecidas. É adequado para qualquer estrutura de população. Goddard (1992) ampliou este modelo para considerar múltiplos QTLs e múltiplos marcadores. Enfatizou o caso em que existe no máximo um QTL segregante, localizado entre dois marcadores (mapeamento por intervalo), obtendo um procedimento que utiliza também a matriz de parentesco do QTL associado aos marcadores. O mapeamento por intervalo permite que a informação dos valores dos segmentos cromossômicos não se perca tão rapidamente de uma geração para outra e, em teoria, esta abordagem maximiza o ganho com a MAS em qualquer programa de melhoramento. Esse modelo é descrito por Resende (2002; 2008).

Nos vários estudos realizados em espécies florestais, exceto em poucos casos, os QTLs individuais para crescimento, qualidade da madeira, adaptação e reprodução não explicaram mais que 5% a 10% da variação fenotípica. Esses resultados sugerem que os caracteres de importância comercial são de herança poligênica e, então, QTLs de grande efeito provavelmente não serão detectados. Assim, o uso da LE-MAS tenderá a ser pouco efetiva nessas espécies.

5.4 MAS via marcadores em desequilíbrio de ligação (LD-MAS)

A maioria dos projetos de pesquisa com QTL mudaram o seu curso para o mapeamento fino baseado em marcadores LD ou diretamente nas mutações causadoras da variação nos QTLs. No caso dos marcadores LD, os mesmos propiciam informações sobre os QTLs em toda a população (através de todas as famílias) e então, a abordagem não difere muito em eficiência, do uso de marcadores diretos (mutação). Pela abordagem LD, a inclusão de informações dos marcadores ou dos haplótipos nos esquemas de avaliação genética pode ser realizada por meio do modelo de QTL aleatório de Fernando e Grosmann (1989). Nesse caso, as covariâncias baseadas em probabilidades de identidade por descendência (IBD) podem ser obtidas além do *pedigree*, via LD e similaridade entre haplótipos ou marcadores. Meuwissen e Goddard (2001) propuseram o uso das informações via LDA e LA para calcular a matriz de covariância via IBD. Lee e van der Werf (2005) mostraram que com alta densidade de marcadores, o valor do uso da informação de ligação e do *pedigree* é reduzido e pouco acrescenta em relação ao uso apenas da informação de LD.

5.5 LD-MAS via Análise de Marcas Únicas

Um modelo misto para a estimação dos efeitos de marcadores individuais é dado por $y = Ju + Wm + e$, em que y é o vetor de observações fenotípicas, J é um vetor com valores 1, u é o escalar referente à média geral, m é o efeito fixo do marcador e e refere-se ao vetor de resíduos aleatórios. W é a matriz de incidência para m. Para marcadores bi-alélicos e modelo de ação gênica aditiva, m é um escalar m_i.

A seleção genética via o marcador é realizada por meio do valor genético predito, dado por $\hat{v} = W\hat{m}$. W, no caso, é uma matriz que associa os genótipos marcadores aos efeitos dos alelos marcadores. Os elementos de W são iguais a zero se o genótipo é aa, 1 se o genótipo é Aa e 2 se o genótipo é AA. A acurácia desse método de seleção é baixa pois, um só marcador explica uma pequena proporção da variação genética do caráter. Essa acurácia pode ser aumentada por meio da inclusão de outros efeitos fixos e aleatórios no modelo, especialmente os efeitos poligênicos não contemplados pelo marcador. Outra forma de aumentar a acurácia é por meio do uso de múltiplos marcadores.

Com a inclusão dos efeitos poligênicos, o seguinte modelo linear misto geral é usado $y = Ju + Wm + Zg^* + e$, em que y é o vetor de observações fenotípicas, u é o escalar da média (de efeito fixo), m é o vetor dos efeitos de alelos do marcador, a^* é o vetor de efeitos poligênicos (aleatório) e e refere-se ao vetor de resíduos aleatórios. W e Z são as matrizes de incidência para m e g^*. Sob esse modelo, a seleção é praticada com base no ordenamento por $\hat{v} = W\hat{m} + \hat{g}^*$.

Como exemplo, considere a avaliação de 12 indivíduos para um caráter e para um marcador do tipo SNP. Os dados referentes aos genótipos e fenótipos dos indivíduos são apresentados na Tabela a seguir, que apresenta também o *pedigree* de mais três indivíduos que não foram avaliados fenotipicamente mas, apenas por seus genótipos.

Indivíduo	Pai	Mãe	Fenótipo	Primeiro Alelo do SNP1	Segundo Alelo do SNP1
1	-	-	9,87	A	A
2	-	-	14,48	A	A
3	-	-	8,91	A	a
4	-	-	14,64	A	A
5	-	-	9,55	A	a
6	-	-	7,96	a	a
7	-	-	16,07	A	A
8	-	-	14,01	A	a
9	-	-	7,96	a	a
10	-	-	21,17	A	A
11	-	-	10,19	A	a
12	-	-	9,23	A	A
13	1	2	-	a	A
14	1	3	-	A	A
15	4	5	-	a	a

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

A matriz J inclui uma coluna para a média geral. As matrizes 1 e W (número de alelos A) são dadas por $J'_{(12x1)} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$ e $W'_{(12x1)} = [1\ 2\ 1\ 2\ 1\ 0\ 2\ 1\ 0\ 2\ 1\ 2]$.

A matriz Z equivale a

Indiv	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

A matriz A equivale a

Indiv	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0.5	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0.5	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.5
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.5
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0.5	0.5	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0.5	0	0.5	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0.5	0.5	0	0	0	0	0	0	0	0	0	1

Assim, as equações de modelo misto são dadas por:

$$\begin{bmatrix} J'J & J'W & J'Z \\ W'J & W'W & W'Z \\ Z'J & Z'W & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m} \\ \hat{g}* \end{bmatrix} = \begin{bmatrix} J'y \\ W'y \\ Z'y \end{bmatrix}$$, em que: $\lambda_1 = \frac{\sigma_e^2}{\sigma_{g*}^2}$; σ_{g*}^2: variância aditiva dos QTLs não associados aos segmentos cromossômicos marcados; σ_e^2: variância residual. Resolvendo-se o sistema matricial, obtém-se o seguinte vetor de soluções:

Efeitos	Soluções
Média Geral	7.2713
Efeito do Alelo A do SNP	3.7856
Efeito Genético Poligênico Indiv. 1	-0.2374
Efeito Genético Poligênico Indiv. 2	-0.0725
Efeito Genético Poligênico Indiv. 3	-0.4294
Efeito Genético Poligênico Indiv. 4	-0.0405
Efeito Genético Poligênico Indiv. 5	-0.3014
Efeito Genético Poligênico Indiv. 6	0.1377
Efeito Genético Poligênico Indiv. 7	0.2455
Efeito Genético Poligênico Indiv. 8	0.5906
Efeito Genético Poligênico Indiv. 9	0.1377
Efeito Genético Poligênico Indiv. 10	1.2655
Efeito Genético Poligênico Indiv. 11	-0.1734
Efeito Genético Poligênico Indiv. 12	-1.1225
Efeito Genético Poligênico Indiv. 13	-0.1549
Efeito Genético Poligênico Indiv. 14	-0.3334
Efeito Genético Poligênico Indiv. 15	-0.1709

Os valores genéticos totais dos indivíduos com avaliações fenotípicas e genotípicas são dados por $\hat{v} = \hat{u} + W\hat{m} + \hat{g}^*$. Para os indivíduos com avaliação genotípica apenas, os valores genéticos totais são dados por $\hat{v} = \hat{u} + W^*\hat{m} + \hat{g}^*$, em que W* é a matriz de incidência molecular para os últimos três indivíduos, a qual difere de W e contém os valores 1, 2 e zero, respectivamente. Os valores genéticos totais dos indivíduos são apresentados a seguir.

Valores Genéticos Totais	Predições
Indivíduo 1	10.820
Indivíduo 2	14.770
Indivíduo 3	10.628
Indivíduo 4	14.802
Indivíduo 5	10.756
Indivíduo 6	7.409
Indivíduo 7	15.088
Indivíduo 8	11.648
Indivíduo 9	7.409
Indivíduo 10	16.108
Indivíduo 11	10.884
Indivíduo 12	13.720
Indivíduo 13	10.902
Indivíduo 14	14.509
Indivíduo 15	7.100

5.6 LD-MAS via Análise de Múltiplos Marcadores e Regressão de Cumeeira

A idéia básica da seleção auxiliada por marcadores é explorar as dependências estatísticas (desequilíbrio de ligação) existentes na distribuição conjunta dos genótipos dos marcadores e do QTL. O desequilíbrio de ligação entre marcadores e QTL pode ser usado com dois objetivos: (i) inferir sobre a localização genômica e efeitos do QTL que afetam um caráter; (ii) obter predições do mérito genético dos candidatos à seleção em um programa de melhoramento genético. Esse segundo objetivo não necessariamente requer o mapeamento de QTLs.

O uso de múltiplos marcadores, advindos de estudos de associação genômica ampla, na predição de valores genéticos, deve considerar que alguns marcadores podem estar detectando o mesmo QTL. Isto porque vários deles podem estar em desequilíbrio de ligação com um QTL de grande efeito. Uma maneira de considerar isso é por meio do uso da regressão múltipla ajustando todos os marcadores simultaneamente, segundo o modelo $y = u + \sum_{i=1}^{p} W_i m_i + e$, em que p é o número de marcadores significativos detectados no estudo de associação genômica ampla.

Pelo método de quadrados mínimos ou regressão, todos os marcadores são testados um por um quanto a sua significância estatística. Outros procedimentos do tipo *stepwise* podem também ser utilizados. Então, os efeitos dos marcadores estatisticamente significativos são estimados simultaneamente. Devido à falta de número suficiente de graus de liberdade, nem todos os marcadores podem ser

testados simultaneamente. Antes de aplicar a regressão múltipla na LD-MAS, é necessário definir quantos marcadores devem ser usados. Isso deve ser definido com base na quantidade de variação genética que é explicada pelo conjunto de marcadores a serem usados na seleção. A vantagem da MAS é proporcional à quantidade de variação genética percentual explicada pelos marcadores. Quanto maior o nível de significância (maior a probabilidade de erro tipo I) adotado nos estudos genômicos, maior número de QTLs são detectados mas, maior também é o número de resultados falsos positivos. Além disso, vários SNPs estarão muito próximos entre eles no genoma e, conseqüentemente, estarão detectando o mesmo QTL.

Hayes et al. (2006) apresentam um método para estimar o verdadeiro número de QTLs controlando um caráter, por meio da correção do número de SNPs significativos levando em conta a ocorrência de falsos positivos e a redundância de alguns SNPs em detectar os mesmos QTLs. Usando esse método, os autores verificaram que o número de QTLs para alguns caracteres em bovinos leiteiros atingiram um platô em 145 a 188 locos. Assim, para capturar toda a variação genética, seriam necessários marcadores flanqueando entre 145 e 188 QTLs. No entanto, em caracteres quantitativos, vários QTL são de pequeno efeito e alguns são de grande efeito (Hayes e Goddard, 2001). Dessa forma, nem todos os QTLs precisarão ser considerados na MAS, pois apenas uma fração deles já explicará a maioria da variação genética. Em bovinos de leite e suínos, 10 % a 20 % dos QTLs explicaram 50 % da variação genética de um caráter quantitativo (Hayes e Goddard, 2001; 2003).

Outro aspecto relacionado à aplicação da regressão múltipla na LD-MAS refere-se ao fato da superestimação dos efeitos dos marcadores-QTLs quando tais efeitos são tratados como fixos (Weller et al., 2005). E se esses efeitos são superestimados, a vantagem potencial da MAS não se concretiza (Whittaker et al., 2000). Nesse contexto, a acurácia da MAS pode ser aumentada por meio de estimadores do tipo *shrinkage*. Os métodos (viciados ou não) que minimizam o erro quadrático médio de estimação conduzem a estimadores/preditores do tipo *shrinkage*. Genericamente, um estimador do tipo *shrinkage* tem a forma de um escalar (variando entre zero e um) multiplicado por um vetor de médias estimadas por quadrados mínimos ou por máxima verossimilhança. A regressão ou *shrinkage* penaliza a estimativa de acordo com o número de observações usadas para estimá-la. Quanto menor o número, mais a estimativa é regressada em direção à média geral. Uma forma de promover o *shrinkage* é tratar os efeitos como aleatórios.

O estimador $\hat{m} = (W'W + \lambda I)^{-1} W' y$ promove *shrinkage*. Quando λ não é conhecido, a escolha arbitrária do mesmo leva ao método de regressão de cumeeira ou "*ridge regression*" (RR), conforme Whittaker et al. (2000) que relataram um aumento de 7 % na eficiência da MAS por meio da RR. No caso, se o parâmetro de regressão for $\lambda = \sigma_e^2 / \sigma_{qtl}^2$, tem-se o BLUP para o efeito do QTL. Whittaker et al. (2000) relatam que o ajuste de muitos marcadores no modelo de regressão produz séria colinearidade, causando instáveis estimativas via quadrados mínimos e pobre predição do escore molecular. Então, sugerem o uso da regressão de cumeeira. Esse procedimento regressa as estimativas de quadrados mínimos em direção a zero, melhora a condição da matriz dos coeficientes das equações de quadrados mínimos e reduz o erro quadrático médio de estimação.

Weller et al. (2005) sugerem um eficiente método de máxima verossimilhança para estimação dos efeitos de QTL, em que as estimativas de quadrados mínimos são regressadas de acordo com uma assumida distribuição dos efeitos do QTL. Também Meuwissen et al. (2001) e Gianola et al. (2003) sugerem abordagens similares, porém, baseadas em princípios bayesianos, em que distribuições a priori para os efeitos de QTL são usadas. Segundo Gianola et al. (2003), o método RR faz mais sentido se visto de uma perspectiva Bayesiana. A regressão ridge é equivalente à adoção de uma priori normal para o vetor de regressão centrado em zero e com estrutura de covariância a priori igual à matriz identidade vezes um escalar, que é a variância da distribuição a priori. Detalhes sobre a estimação Bayesiana são apresentados no tópico sobre GWS.

Além das técnicas do índice de seleção e do BLUP, da regressão múltipla e da regressão de cumeeira dos fenótipos sobre os genótipos marcadores, outras técnicas foram propostas para a MAS. Gianola et al. (2003) propuseram a modelagem das associações fenótipo-marcadores de forma hierárquica via modelos multiníveis incluindo efeitos cromossômicos, covariância espacial de efeitos de marcadores dentro de cromossomos e heterogeneidade de famílias. Segundo os autores, existem problemas estatísticos com o índice de seleção de Lande e Thompson (1990), uma vez que a matriz de covariância dos escores moleculares é singular e leva a um infinito número de soluções. Outra dificuldade do método de quadrados mínimos existe quando o número de marcadores é quase da mesma ordem que o número de indivíduos. Nesse caso, alguma técnica de redução dimensional, como por exemplo, a decomposição por valor singular, deve ser usada. Gianola et al. (2003) defendem uma abordagem que trata todos os efeitos como alelatórios. Isso propicia flexibilidade para acomodar novos efeitos no modelo. O estimador de regressão ridge de Whittaker et al. (2000) implica que todos os efeitos de marcadores são independentes. Entretanto, existe evidência de co-expressão de genes pelo menos no mesmo cromossomo. Isso indica que não se verifica a suposição de que os segmentos marcados dentro de cromossomo têm efeitos independentes. QTLs adjacentes podem ser mais dependentes do que QTLs distantes. Assim, alguma estrutura de covariância espacial ao longo do cromossomo pode ser necessária.

Outro fato é que alguns cromossomos podem ter mais QTLs que outros, conduzindo a variação entre cromossomos. Essa heterogeneidade pode ser acomodada pela introdução de efeitos cromossômicos no modelo, com marcadores em diferentes cromossomos tendo distintas distribuições (Gianola et al., 2003). Procedimentos semi-paramétricos foram também apresentados por Gianola et al. (2006), os quais permitem estimar interações entre milhares de marcadores. Os métodos incluem regressão kernel, a qual regressa os efeitos de marcadores de acordo com um parâmetro de alisamento imbutido nas equações de modelo misto.

5.7 LD-MAS via Análise de IBD

Fernando e Grossman (1989) assumiram efeito aleatório de QTL com variância conhecida e desenvolveram uma abordagem para a predição de valores genéticos de todos os indivíduos de uma população, em que todos os candidatos à seleção participam da construção da matriz IBD. O método inclui os efeitos de QTL via ligação com marcadores genéticos e é adequado quando a herdabilidade e a freqüência de recombinação entre marcadores e QTLs são também conhecidos. O método é válido para qualquer estrutura de população diplóide e é uma extensão da modelo tradicional de análise de ligação dentro de família, por meio do uso da informação de todo o *pedigree*.

De posse de estimativas das variâncias do QTL e da variância genética aditiva, tem-se o seguinte modelo de avaliação genética: $y = Xb + Zg^* + Qq + e$ em que: q é o vetor de efeitos genéticos aditivos gaméticos, que contém duas incógnitas para cada indivíduo em cada loco (um efeito do alelo do QTL no cromossomo maternal e outro no paternal); g^* é o vetor aleatório de efeitos poligênicos, excluindo q; b e e são vetores de efeitos fixos e de erros aleatórios, respectivamente; Q é a matriz de incidência para os efeitos gaméticos do QTL; Z é a matriz de incidência para g^*; X é a matriz de incidência para b. A seleção dos indivíduos é baseada em $\hat{g} = \hat{g}^* + \hat{q}$.

5.8 Número de Locos a ser Usado na LD-MAS

A razão pela qual uma limitada fração da variação genética é explicada pelos QTLs identificados refere-se aos baixos níveis de significância adotados na detecção, visando evitar a detecção de muitos falsos positivos, quando se avalia muitas posições para a presença de QTL. Relaxados níveis de significância, da ordem de 20 % a 40 %, têm maximizado o ganho genético com a MAS (Hospital et al., 1997). Esses níveis de significância conduzem ao uso na seleção de um maior número de locos, detectados como significativos. Bernardo e Yu (2007) obtiveram resultados similares.

6 Seleção genômica ampla (GWS)

6.1 Fundamentos da *Genome Wide Selection* (GWS)

A seleção genética tem sido praticada pelo procedimento BLUP (em suas versões frequentista e bayesiana) usando dados fenotípicos avaliados a campo. Uma primeira proposição realizada para aumentar a eficiência desse procedimento baseado em dados fenotípicos foi descrita por Lande & Thompson (1990), por meio da seleção auxiliada por marcadores (MAS) moleculares. A MAS utiliza simultaneamente dados fenotípicos e dados de marcadores moleculares em ligação gênica próxima com alguns locos controladores de características quantitativas (QTL). Em geral, os dados de marcadores são utilizados como covariáveis (efeitos fixos) na explicação dos valores fenotípicos dos indivíduos em avaliação ou como efeitos aleatórios incorporados no modelo para o fenótipo (Fernando e Grossman, 1989). Esses marcadores são eleitos ou não como determinantes dos efeitos de QTLs após modelagem estatística sujeita a erros do tipo II (probabilidade de aceitar uma hipótese falsa, ou seja, tomar como verdadeira uma hipótese falsa de ausência de efeitos).

A seleção baseada na MAS apresenta as seguintes características: requer o estabelecimento (análise de ligação) de associações marcadores-QTLs para cada família em avaliação, ou seja, essas associações apresentam utilidade para seleção apenas dentro de cada família mapeada; para ser útil precisa explicar grande parte da variação genética de uma característica quantitativa, que é governada por muitos locos de pequenos efeitos. Isto não tem sido observado na prática, exatamente em função da natureza poligênica e alta influência ambiental nos caracteres quantitativos, fato que conduz à detecção apenas de um pequeno número de QTLs de grandes efeitos, os quais não explicam suficientemente toda a variação genética; só apresenta superioridade considerável em relação à seleção baseada em dados fenotípicos, quando o tamanho de família avaliado e genotipado é muito grande (da ordem de 500 ou mais). Em função desses aspectos, a implementação da MAS tem sido limitada e os ganhos em eficiência muito reduzidos (Dekkers, 2004).

O atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações de DNA na seleção, de forma a permitir alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos. Visando a esses objetivos, Meuwissen et al. (2001) propuseram um novo método de seleção denominado seleção genômica (GS) ou seleção genômica ampla (*genome wide selection* – GWS) ou seleção genômica total (whole *genome selection* – WGS), a qual pode ser aplicada em todas as famílias em avaliação nos programas de melhoramento genético, apresenta alta acurácia seletiva para a seleção baseada exclusivamente em marcadores (após terem seus efeitos genéticos estimados a partir de dados fenotípicos em uma amostra da população de seleção) e não exige prévio conhecimento das posições (mapa) dos QTLs, não estando sujeita aos erros tipo II associados à seleção de marcadores ligados a QTLs.

Esse método permaneceu discreto por cerca de seis anos, devido ao fato dos marcadores moleculares disponíveis à época serem caros e restritos. Recentemente, com o desenvolvimento e baixo custo dos marcadores tipo SNP (*single nucleotide polymorphism*), o método tornou-se atrativo (Meuwissen, 2007; Goddard & Hayes, 2007; Fernando et al., 2007; Resende 2007; Bernardo e Yu, 2007). A GWS permite a predição de valores genéticos genômicos e é excelente para caracteres de baixa herdabilidade, ao contrário da MAS, que não é útil para caracteres de baixa herdabilidade.

A analise de QTL baseia-se na detecção, mapeamento e uso de QTLs na seleção (MAS). Ou seja, enfatiza a determinação do número, posição e efeitos dos QTLs marcados. A GWS é definida como a seleção simultânea para centenas ou milhares de marcadores, os quais cobrem o genoma de uma maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Esses marcadores em desequilíbrio de ligação com os QTLs, tanto de grandes quanto de pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo. O número de SNPs é de tal magnitude que a probabilidade de se encontrar um QTL em desequilíbrio de ligação com pelo menos um marcador é muito alta. Este aspecto é importante uma vez que somente os marcadores em desequilíbrio de ligação com os QTLs serão úteis na determinação dos fenótipos e na explicação da variação genética. Os efeitos dos marcadores são estimados em uma amostra de indivíduos pertencentes a várias famílias. Assim, o impacto de determinadas famílias específicas (com específicos padrões de desequilíbrio de ligação) nas estimativas dos efeitos dos marcadores será minimizado. É importante enfatizar que os marcadores terão seus efeitos genéticos estimados a partir de uma amostra de pelo menos 1.000 indivíduos genotipados e fenotipados, ou seja, com base em pelo menos 1.000 repetições experimentais de cada loco. Assim, embora a herdabilidade de cada marcador efetivo (aquele que identifica um dos poligenes com precisão) seja muito baixa, com 1.000 repetições essa herdabilidade se torna alta. Em outras palavras, o efeito de ambiente será minimizado por meio do uso de um número de repetições muito alto. Essa é a mesma filosofia da avaliação e seleção de características quantitativas com base em fenótipos em experimentos de campo, implantados com grande número de repetições.

A GWS é ampla porque atua em todo o genoma, capturando todos os genes que afetam um caráter quantitativo. E isso sem a necessidade prévia de identificar os marcadores com efeitos significativos e de mapear QTLs, como no caso da MAS. Valores genéticos genômicos associados a cada marcador ou alelo são usados para fornecer o valor genético genômico global de cada indivíduo. Há uma diferença básica na predição de valores genéticos tradicionais e na predição de valores genéticos genômicos. Nos primeiros, informações fenotípicas são utilizadas visando a inferências sobre os efeitos dos genótipos dos indivíduos e, nos últimos, informações genotípicas (genótipos para os alelos marcadores) são usadas visando a inferências sobre os valores fenotípicos futuros (ou valores genéticos genômicos preditos) dos indivíduos. Em outras palavras, os métodos tradicionais usam o fenótipo para inferir sobre o efeito do genótipo e a GWS usa o genótipo, com efeito genético pré-estimado em uma amostra da população, para inferir sobre o fenótipo a ser expresso nos candidatos à seleção.

Os efeitos dos marcadores não serão, necessariamente, os mesmos em diferentes estudos e ambientes. Na GWS, os efeitos genéticos dos marcadores são estimados e usados na seleção para cada população de melhoramento e em um determinado ambiente. Modelos de estimação incluindo a interação genótipos x ambientes podem também ser usados, visando verificar a possibilidade de se obter estimativas válidas para um conjunto de ambientes. Mas, isso dependerá da magnitude da interação envolvendo os vários ambientes.

A GWS pode basear-se no uso de: (i) apenas dos marcadores; (ii) de haplótipos ou intervalos definidos por dois marcadores; (iii) haplótipos definidos por mais de dois marcadores, incluindo a covariância entre haplótipos devida à ligação. Segundo Callus et al. (2008), para caracteres de baixa herdabilidade (10%) não existem diferenças significativas entre essas três abordagens. Solberg et al. (2006) mostraram que é possível praticar a GWS eficientemente com o uso apenas dos marcadores, ou seja, com a predição direta dos efeitos dos marcadores. Relatam também que isso é vantajoso porque não há necessidade de estimar as fases de ligação entre os marcadores, as quais são estimadas com algum erro.

Cada par contíguo de marcadores define um haplótipo ou intervalo. Existem apenas dois alelos para cada marcador, pois os SNPs têm diferenças em um único par de bases. Dessa forma, para cada par de marcadores existem quatro haplótipos possíveis. A freqüência de cada haplótipo depende da freqüência dos alelos em cada marcador e da distância entre marcadores ou eventos de recombinação. Assim, um número suficiente de indivíduos devem ser genotipados de forma que todos os haplótipos estejam representados nos indivíduos com avaliações fenotípicas (Schaeffer, 2006).

Não apenas marcadores SNPs podem ser usados na GWS. Marcadores microssatélites também se prestam a esse fim. Solberg et al. (2006) relatam que o uso de SNPs requer quatro a cinco vezes maior densidade de marcadores do que o uso de microssatélites. Isto se deve à natureza bi-alélica (bi-nucleotídica) dos SNPs e multi-alélica dos microssatélites. Tais marcadores são eficientes por serem co-dominantes, multi-alélicos, abundantes e apresentarem alta transferibilidade entre indivíduos e espécies. Outra classe de marcadores que se adequa bem à GWS são os DArT (*Diversity Array Technology*), a qual permite amostrar amplamente o genoma sem a necessidade de conhecimento prévio das sequências de DNA.

A GWS fundamenta-se nos marcadores genéticos moleculares do tipo SNP (polimorfismo de um único nucleotídeo), o qual se baseia na detecção de polimorfismo resultante da alteração de um único par de base no genoma. E para que uma variação seja considerada SNP, essa deve ocorrer em pelo menos 1 % da população. Os SNPs são a forma mais abundante de variação do DNA em genomas e são preferidos em relação a outros marcadores genéticos devido à sua baixa taxa de mutação e facilidade de genotipagem. Milhares de SNPs podem ser usados para cobrir o genoma de um organismo com marcadores que não estão a mais de 1 cM um do outro no genoma inteiro. A GWS atua mais proximamente aos QTNs (nucleotídeos de características quantitativas) ou sobre marcadores fortemente ligados a esses. Os QTNs são polimorfismos funcionais, causadores diretos da

variação quantitativa observada. A análise de SNPs permite a detecção de polimorfismos funcionais ou polimorfismos em forte desequilíbrio de ligação com os QTNs. Tecnologias para genotipagem de milhares de SNPs em microarranjos estão disponíveis atualmente. Microarranjos são sistemas de arranjos de DNA que utilizam lâminas de vidro e sondas fluorescentes e permitem depositar milhares de seqüências de DNA. Nessa técnica são utilizados nucleotídeos marcados capazes de emitir fluorescência ao invés de radioatividade.

O desenvolvimento conceitual da GWS coincide com a tecnologia associada aos SNPs, a qual é acurada e relativamente barata. A GWS usa associações de um grande número de marcadores SNPs em todo o genoma com os fenótipos, capitalizando no desequilíbrio de ligação entre os marcadores e QTLs proximamente ligados. As predições derivadas de dados fenotípicos e de genótipos SNPs em alta densidade em uma geração são então usadas para obtenção dos valores genéticos genômicos (VGG) dos indivíduos de qualquer geração subseqüente, tendo por base os seus próprios genótipos marcadores.

Quando o desequilíbrio de ligação entre marcadores não é completo, as freqüências alélicas conjuntas envolvendo dois locos podem mudar substancialmente através das gerações, conduzindo a mudanças nos haplótipos. Assim, os efeitos dos marcadores necessitarão ser re-estimados para manter a acurácia da GWS em várias gerações (Dekkers, 2007). Com desequilíbrio de ligação completo os efeitos estimados permanecem constantes através das famílias e gerações em um mesmo ambiente.

6.2 Acurácia da GWS

A acurácia ($r_{q\hat{q}}$) da seleção GWS depende da proporção (r^2_{mq}) da variação genética explicada pelos marcadores e da acurácia ($r_{m\hat{m}}$) da predição dos efeitos dos marcadores ou haplótipos que estão em desequilíbrio de ligação com os QTLs, segundo a expressão $r_{q\hat{q}} = (r^2_{m\hat{m}} r^2_{mq})^{1/2}$. O parâmetro r^2_{mq} depende da densidade de marcadores e da extensão e padrão do desequilíbrio de ligação que existe na população. Por sua vez, o parâmetro $r_{m\hat{m}}$ depende da quantidade e precisão dos dados disponíveis para estimar os efeitos dos marcadores, além da eficiência da estratégia e dos métodos estatísticos usados na predição.

Resende (2008) e Resende et al. (2008) apresentaram uma abordagem para cômputo da acurácia esperada com a GWS, a qual foi empregada por Grattapaglia e Resende (2011). A acurácia esperada é dada por $r_{q\hat{q}} = (r^2_{mq} r^2_{m\hat{m}})^{1/2} = \sqrt{r^2_{mq}(Nh^2_m)/[1+(N-1)h^2_m]}$. Com ajuste dos efeitos poligênicos residuais no modelo de predição tem-se $h^2_m = (r^2_{mq} h^2 / n_Q)/(r^2_{mq} h^2 + (1-h^2))$. Sem ajuste dos efeitos poligênicos residuais tem-se $h^2_m = (r^2_{mq} h^2 / n_Q)$, em que h^2 é a herdabilidade individual no sentido restrito do caráter, h^2_m é a herdabilidade individual de um loco, N é o número de indivíduos genotipados e fenotipados, r^2_{mq} é a proporção da variação genética explicada pelos marcadores (magnitude do

desequilíbrio de ligação), $r^2_{m\hat{m}}$ é a confiabilidade da estimativas dos efeitos das marcas e n_Q é o número de genes (quando conhecido) controlando o caráter ou o número de segmentos cromossômicos independentes (quando o número de genes é desconhecido), os quais não sofrem recombinação dentro deles.

A magnitude do desequilíbrio de ligação é quantificada por $r^2_{mq} = E(r^2) = \frac{1}{4N_e S + 1}$ (Sved, 1971) ou $r^2_{mq} = E(r^2) = \frac{1}{4N_e S + 2}$ (Tenesa et al., 2007). O valor esperado da estatística r^2, que mede a magnitude do desequilíbrio de ligação, depende do tamanho efetivo populacional (Ne) e da freqüência de recombinação (função da distância S entre locos).

Para Ne = 10 e distância entre marcas de 1 cM, o valor esperado de r^2 é 0,71. Para essa mesma distância entre locos e Ne de 20 e 30, os valores esperados de r^2 são 0,56 e 0,45, respectivamente. Com o dobro de marcadores e espaçamento de 0.5 cM entre marcadores, os valores esperados de r^2 são: (i): Ne = 10; r^2 = 0.83; (ii) Ne = 20; r^2 = 0.71; (iii) Ne = 30; r^2 = 0.63. Em eucalipto (tamanho do genoma igual a 1.300 cM), com Ne igual a 20, 1.300 marcadores espaçados a 1 cM conduziriam a um r^2 de 0.56. Com o dobro de marcadores (2.600) e espaçamento de 0.5 cM entre marcadores, o valor esperado de r^2 é 0.71. Portanto, 2.600 marcadores seria um número mínimo de marcadores para implementação da GWS em eucalipto. Nessa situação, com N igual a 1.000 indivíduos genotipados e fenotipados, n_Q igual a 100 locos e h^2 de 30%, aplicando-se a fórmula da acurácia ter-se-ia uma acurácia de 70 %, valor esse muito interessante do ponto de vista prático, para a seleção precoce em plântulas.

Daetwyler et al. (2008) assumiram variância residual $\sigma^2_e = 1$ e $r^2_{mq} = 1$, obtendo $r_{m\hat{m}} = \sqrt{(Nh^2/n_Q)/[1+(Nh^2/n_Q)]} = \sqrt{(\omega h^2)/[1+(\omega h^2)]}$ mostrando a importância da quantidade $\omega = N/n_Q$, a qual é equivale ao número de indivíduos N usados para estimar o efeito de cada loco na população de estimação. Resende (2008) obteve uma expressão mais geral, não assumindo $\sigma^2_e = 1$ e $r^2_{mq} = 1$, ou seja, mantendo esses dois elementos na fórmula e assumindo $\sigma^2_y = 1$ (distribuição normal padrão para os fenótipos, em que σ^2_y é a variância fenotípica). Sem ajuste dos efeitos poligênicos residuais no modelo de predição tem-se a expressão (Resende, 2008):

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

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

.

Goddard et al. (2011) assumiram variância residual $\sigma^2_e = 1$ e obtiveram a expressão $r_{q\hat{q}} = (r^2_{mq} r^2_{m\hat{m}})^{1/2} = \sqrt{r^2_{mq}(Nr^2_{mq}h^2/n_Q)/[1+Nr^2_{mq}h^2/n_Q]}$, que é praticamente igual à expressão de Resende (2008). É igual desde que se assuma N = (N − 1). Goddard et al. (2011) consideram em lugar de n_Q o número efetivo de segmentos cromossômicos (M_e) segregando na população, ou seja, o numero de blocos de DNA que não sofrem recombinação dentro deles e que devem ser marcados adequadamente. O M_e depende do Ne da população e do tamanho do genoma da espécie. Detalhes sobre esse tema e sobre o cálculo de Me podem ser vistos no tópico 6.26.

Em resumo, a acurácia da GWS depende de cinco fatores: (i) da herdabilidade do caráter; (ii) do número de locos controlando o caráter e da distribuição de seus efeitos; (iii) do número de indivíduos na população de descoberta; (iv) do tamanho efetivo populacional; (v) do espaçamento entre marcadores, o qual depende do seu número e do tamanho do genoma. Os dois primeiros fatores não estão sobre o controle do melhorista. Os três últimos podem ser modificados pelo melhorista visando aumentar a acurácia da GWS.

Valores de acurácia esperada para várias situações foram tabelados por Resende (2008). Na Tabela 23 são apresentados resultados da acurácia seletiva da GWS para um caráter controlado por 100 locos e com herdabilidade individual no sentido restrito igual a 0.30. Verifica-se que, para uma população de eucalipto com tamanho efetivo 20 (r^2_{mq} = 0.7), a acurácia seletiva esperada com a GWS é de 0.79, para um tamanho amostral de N = 4000 indivíduos. Esse valor supera a acurácia máxima (0.70) para a seleção de indivíduos em testes de família, pelo BLUP tradicional na idade adulta, para um caráter com herdabilidade de 20%. Isto atesta o grande potencial da GWS.

Tabela 23. Aumento da acurácia da GWS em função do aumento do tamanho da população de estimação. Caráter controlado por 100 locos e com herdabilidade individual no sentido restrito igual a 0.30.

Número de Indivíduos	r^2_{mq} = 0,1	r^2_{mq} = 0,3	r^2_{mq} = 0,5	r^2_{mq} = 0,7	r^2_{mq} = 0,9
100	0,06	0,18	0,27	0,36	0,44
200	0,09	0,24	0,36	0,47	0,57
500	0,13	0,33	0,48	0,61	0,72
1000	0,17	0,40	0,57	**0,70**	0,81
2000	0,21	0,46	0,62	0,76	0,87
4000	0,25	0,50	0,66	**0,79**	0,91
8000	0,28	0,52	0,68	0,81	0,93

***Acurácia máxima para a seleção de indivíduos pelo BLUP tradicional na idade adulta = 0.70**

Ganhos adicionais podem ser conseguidos por unidade de tempo, conforme a Tabela 24. Verifica-se que ganho da ordem de 126% pode ser conseguido com a redução, de 4 para 2 anos (ou seja, 50%) do tempo necessário para completar um ciclo de seleção.

Tabela 24. Eficiência da GWS por unidade de tempo.

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

Detalhes das Expressões da Acurácia

Partindo da expressão de Resende (2008), $r_{g\hat{g}} = \sqrt{\frac{r_{mq}^2 (N r_{mq}^2 h^2 / n_Q)}{1 + (N-1) r_{mq}^2 h^2 / n_Q}}$, e assumindo $\frac{(N-1)}{N} = 1$ para N grande tem-se:

$r_{g\hat{g}} = \sqrt{\frac{r_{mq}^2 (N r_{mq}^2 h^2 / n_Q)}{1 + N r_{mq}^2 h^2 / n_Q}}$, expressão idêntica à de Goddard et al. (2011).

Rearranjando essa expressão tem-se:

$r_{g\hat{g}} = \sqrt{\frac{r_{mq}^2}{1 + \frac{1}{N r_{mq}^2 h^2 / n_Q}}}$, donde se verifica que o valor máximo atingido pela acurácia é igual à raiz quadrada de r_{mq}^2.

Assim, torna-se imperativo aumentar r_{mq}^2 para se aumentar a acurácia. E r_{mq}^2 é dada por $r_{mq}^2 = \frac{n_m}{n_m + M_e} = \frac{1}{1 + \frac{M_e}{n_m}}$, donde se verifica que o seu aumento só pode ser conseguido com o aumento do número n_m de marcadores, visto que M_e é fixo para determinada espécie e Ne da população, conforme mostrado mais adiante.

Rearranjando $r_{g\hat{g}} = \sqrt{\frac{r_{mq}^2}{1 + \frac{1}{N r_{mq}^2 h^2 / n_Q}}}$, tem-se $r_{g\hat{g}} = \sqrt{1/(\frac{1}{r_{mq}^2} + \frac{n_Q}{N (r_{mq}^2)^2 h^2})}$, donde se verifica que a acurácia é diretamente proporcional a N, h^2 e r^2_{mq} e inversamente proporcional a n_Q.

Se $r^2_{mq} = 1$, $r_{g\hat{g}} = \sqrt{1/(1 + \frac{1}{N h^2 / n_Q})}$ ou $r_{g\hat{g}} = \sqrt{1/(1 + \frac{n_Q}{N h^2})}$, expressão equivalente à de Daetwyler et al. (2008). Com base nessa expressão, o número de QTL ou genes pode ser estimado por $n_Q = \frac{1 - \hat{r}^2_{g\hat{g}}}{\hat{r}^2_{g\hat{g}}} N h^2$, em que $\hat{r}^2_{g\hat{g}}$ é a estimativa do quadrado da acurácia obtida com base na GWS aplicada sobre dados experimentais Daetwyler et al. (2010).

Se $r^2_{mq} = 1$ e se $h^2 = 1$: $r_{g\hat{g}} = \sqrt{1/(1 + \frac{n_Q}{N})}$ e a acurácia depende apenas de um tamanho amostral N de indivíduos suficiente para estimar efeitos de n_Q genes. Se $r^2_{mq} < 1$ e se $h^2 = 1$, tem-se $r_{g\hat{g}} = \sqrt{1/(\frac{1}{r_{mq}^2} + \frac{n_Q}{N (r_{mq}^2)^2})}$ e a acurácia depende também da maximização de r^2_{mq}, além de um N adequado.

Substituindo $r_{mq}^2 = \frac{n_m}{n_m + M_e}$ na expressão da acurácia e rearranjando chega-se a $r_{g\hat{g}} = \sqrt{1/[1 + \frac{M_e}{n_m} + \frac{n_Q}{N h^2} \frac{(M_e + n_m)^2}{n_m^2}]}$, donde se verifica (em $\frac{M_e}{n_m}$) a importância de n_m em explicar M_e (número de marcadores em explicar o número de segmentos

cromossômicos) e a importância de Nh^2 em explicar o número de genes (em $\frac{n_Q}{Nh^2}$). Verifica-se também que M_e é inversamente proporcional à acurácia.

Considerando $M_e = 4N_e L$ conforme Sved (1971) e $n_Q = \frac{2N_e L}{Ln(2Ne)}$ conforme Hayes et al. (2009) tem-se a expressão final $r_{g\hat{g}} = \sqrt{1/[1 + \frac{4Ne\ L}{n_m} + \frac{2Ne\ L}{Nh^2[Ln(2Ne)]}\frac{(4Ne\ L + n_m)^2}{n_m^2}]}$, a qual depende de cinco fatores: de maneira inversamente proporcional a Ne e L (tamanho total do genoma em Morgans), e, diretamente proporcional a N, h^2 e n_m.

Sved (1971) considera $M_e = 4N_e L$, e, portanto,

$$r_{mq}^2 = \frac{n_m}{n_m + M_e} = \frac{1}{1 + \frac{4\ Ne\ L}{n_m}}$$

, em que $4N_e L$ é o número total de segmentos cromossômicos (a serem marcados) e L/n_m = S. Hayes et al. (2009) consideram $M_e = 2N_e L$ como o número efetivo (de mesmo tamanho, ponderados pelos comprimentos) de segmentos cromossômicos (corrigidos, contendo genes). Esses mesmos autores consideram $n_Q = M_e\ Var(W) = M_e[2p(1-p)] = \frac{M_e}{Ln(2N_e)} = \frac{2N_e L}{Ln(2N_e)}$ como o número provável de QTL ou segmentos corrigidos para mesmo tamanho e frequência, em que $Var(W) = 2p(1-p) = \frac{1}{Ln(2Ne)}$ é a variância da variável indicadora W dos marcadores.

Quando $Var(W) = 1$, se $n_Q = 2NeL$, tem-se:

$$r_{g\hat{g}} = \sqrt{1/[1 + \frac{2\,n_Q}{n_m} + \frac{n_Q}{Nh^2}\frac{(2\,n_Q + n_m)^2}{n_m^2}]}$$

ou $r_{g\hat{g}} = \sqrt{1/[1 + \frac{4Ne\ L}{n_m} + \frac{2Ne\ L}{Nh^2}\frac{(4Ne\ L + n_m)^2}{n_m^2}]}$.

Essa expressão é conservadora e leva a menores acurácias estimadas.

Se $n_Q = 2NeL = Me_{max}$ e $(n_Q + n_m)/n_m$ tender a um tem:se: $r_{g\hat{g}} = \sqrt{1/(1 + \frac{n_Q}{n_m} + \frac{n_Q}{Nh^2})}$ ou $r_{g\hat{g}} = \sqrt{1/(1 + \frac{2Ne\ L}{n_m} + \frac{2Ne\ L}{Nh^2})}$. Essa expressão é mais simples para cômputo rápido.

Goddard (2011), acrescenta na expressão de Me uma divisão por $Ln(Ne\ L/k)$, em que k é o número de cromossomos, de forma que tem-se $Me = 2N_e L/[Ln(Ne\ L/k)$. Quanto maior o tamanho L/k do cromossomo, melhor (mais marcadores no cromossomo ajudando a capturar o mesmo QTL). Nesse caso, tem-se: $r_{mq}^2 = \frac{m}{m + 2N_e L/[Ln(Ne\ L/k)]}$.

A fração Ln (Ne L / k) advem do fato de se considerar o LD entre a marca alvo e todos os marcadores dentro de cromossomos e não apenas o vizinho mais próximo e o alvo. Outra forma de estimar Me é via o r² de Hill e Robertson: $Me = \frac{1}{\bar{r}^2_{pl}}$ em que $\bar{r}^2_{pl}$ é o r² médio para todos os pares de locos, o qual relaciona-se com a variancia dos coeficientes de parentesco (matriz A), por meio de Var (A) = $\bar{r}^2_{pl} = \frac{1}{Me}$.

Goddard (2011) usa a formula com Me e não n_Q. Assim,

$r_{g\hat{g}} = \sqrt{1/[1 + \frac{Me}{n_m} + \frac{n_Q}{Nh^2}\frac{(Me + n_m)^2}{n_m^2}]}$ equivale a $r_{g\hat{g}} = \sqrt{1/[1 + \frac{Me}{n_m} + \frac{Me}{Nh^2}\frac{(Me + n_m)^2}{n_m^2}]}$.

Em conclusão, recomenda-se a expressão $r_{g\hat{g}} = \sqrt{1/[1 + \frac{2NeL}{n_m} + \frac{2NeL}{Nh^2}\frac{(2NeL + n_m)^2}{n_m^2}]}$, obtida pela derivação de Resende (2008) e considerando Me = 2NeL.

Um resumo é apresentado a seguir.

Resumo das Expressões para a Acurácia

Autor	r_{gg}	r^2_{mq}	Me	n_Q	σ^2_e
Resende (2008)	$r_{g\hat{g}} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2/n_Q)}{1+(N-1)r^2_{mq}h^2/n_Q}}$	$r^2_{mq} = \frac{1}{1+4N_eL/n_m} = \frac{n_m}{n_m+M_e}$	$M_e = 4N_eL$	$n_Q = n_m[2p(1-p)]$ ou n_Q suposto conforme a herdabilidade	$\sigma^2_e = (1-h^2)\sigma^2_y$ $\sigma^2_y = 1$
Daetwyler et al. (2008)	$r_{g\hat{g}} = \sqrt{\frac{Nh^2/n_Q}{1+Nh^2/n_Q}} = \sqrt{1/(1+\frac{n_Q}{Nh^2})}$	$r^2_{mq} = 1$	-	n_Q suposto	$\sigma^2_e = \sigma^2_y$ $\sigma^2_e = 1$
Goddard (2009)	$r_{g\hat{g}} = \sqrt{\frac{Nh^2/n_Q}{1+Nh^2/n_Q}} = \sqrt{1/(1+\frac{n_Q}{Nh^2})}$	$r^2_{mq} = 1$	$M_e = \frac{2NeL}{Ln(4NeL)}$	$n_Q = \frac{M_e}{Ln(2Ne)}$	$\sigma^2_e = \sigma^2_y$ $\sigma^2_e = 1$
Hayes et al. (2009)	$r_{g\hat{g}} = \sqrt{\frac{Nh^2/n_Q}{1+Nh^2/n_Q}} = \sqrt{1/(1+\frac{n_Q}{Nh^2})}$	$r^2_{mq} = 1$	$M_e = 2NeL$	$n_Q = \frac{M_e}{Ln(2Ne)}$	$\sigma^2_e = \sigma^2_y$ $\sigma^2_e = 1$
Goddard et al. (2011)	$r_{g\hat{g}} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2/n_Q)}{1+Nr^2_{mq}h^2/n_Q}}$	$r^2_{mq} = \frac{n_m}{n_m+M_e}$	$M_e = \frac{2NeL}{Ln(NeL/c)}$	$n_Q = M_e$	$\sigma^2_e = \sigma^2_y$ $\sigma^2_e = 1$

Fórmulas Alternativas da Derivação de Resende (2008)

1	$r_{g\hat{g}} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2/n_Q)}{1+(N-1)r^2_{mq}h^2/n_Q}}$
2	$r_{g\hat{g}} = \sqrt{1/[1 + \frac{M_e}{n_m} + \frac{n_Q}{Nh^2}\frac{(M_e + n_m)^2}{n_m^2}]}$
3	$r_{g\hat{g}} = \sqrt{1/[1 + \frac{4NeL}{n_m} + \frac{2NeL}{Nh^2[Ln(2Ne)]}\frac{(4NeL + n_m)^2}{n_m^2}]}$
4	$r_{g\hat{g}} = \sqrt{1/[1 + \frac{2NeL}{n_m} + \frac{2NeL}{Nh^2}\frac{(2NeL + n_m)^2}{n_m^2}]}$

Casos especiais da derivação de Resende (2008)

Casos especiais da derivação de Resende (2008)	Se r^2_{mq} = 1 (dados de sequência ou n_m alto): $r_{gg\hat{}} = \sqrt{1/(1+\frac{n_Q}{Nh^2})}$ Nh^2 = 10 n_Q para obter acurácia de 90% e N = 50 n_Q para h^2 de 0.2: 200 locos: N = 10.000	Se h^2 = 1 (dados de valores genéticos com acurácia 100%) $r_{gg\hat{}} = \sqrt{1/(\frac{1}{r^2_{mq}}+\frac{n_Q}{N(r^2_{mq})^2})}$	Se r^2_{mq} = 1 e se h^2 = 1: $r_{gg\hat{}} = \sqrt{1/(1+\frac{n_Q}{N})}$: necessidade de grande número de dados para conhecer os efeitos de cada loco e alelo; N = 10 n_Q para obter acurácia de 90%. Se 200 locos: N = 2000.

* n_Q pode ser visto como n_Q = Ne L

A seguir ilustra-se o aumento de r^2_{mq} em função do aumento de n_m em Eucalipto, com base na expressão de Hayes et al. (2009).

Comportamento de r^2_{mq} em função de Ne, n_m e Me em Eucalipto.

Ne	Me	n_m	r^2_{mq}	Ne	Me	n_m	r^2_{mq}
10	86.8	3000	0.85	20	141.0	3000	0.74
10	86.8	5000	0.91	20	141.0	5000	0.83
10	86.8	10000	0.95	20	141.0	10000	0.91
10	86.8	15000	0.97	20	141.0	15000	0.94
10	86.8	20000	0.97	20	141.0	20000	0.95
10	86.8	30000	0.98	20	141.0	30000	0.97
30	190.5	3000	0.66	50	282.3	3000	0.54
30	190.5	5000	0.76	50	282.3	5000	0.66
30	190.5	10000	0.87	50	282.3	10000	0.79
30	190.5	15000	0.91	50	282.3	15000	0.85
30	190.5	20000	0.93	50	282.3	20000	0.88
30	190.5	30000	0.95	50	282.3	30000	0.92
100	490.7	3000	0.37	200	867.9	3000	0.22
100	490.7	5000	0.49	200	867.9	5000	0.32
100	490.7	10000	0.66	200	867.9	10000	0.49
100	490.7	15000	0.74	200	867.9	15000	0.59
100	490.7	20000	0.79	200	867.9	20000	0.66
100	490.7	30000	0.85	200	867.9	30000	0.74

A seguir ilustra-se o aumento da acurácia em função do aumento de N e n_m em Eucalipto, com base na expressão de Resende (2008).

Comportamento de r_{gg} em função de N e número de marcas em Eucalipto. Ne 100.

h^2_a	N	Nmarcas	r_{gg}	h^2_a	N	Nmarcas	r_{gg}
0.1	1000	10000	0.28	0.1	1000	20000	0.33
0.2	1000	10000	0.37	0.2	1000	20000	0.44
0.3	1000	10000	0.43	0.3	1000	20000	0.51
0.4	1000	10000	0.48	0.4	1000	20000	0.56
0.5	1000	10000	0.51	0.5	1000	20000	0.60
0.6	1000	10000	0.54	0.6	1000	20000	0.63

0.1	2000	10000	0.37	0.1	2000	20000	0.44
0.2	2000	10000	0.48	0.2	2000	20000	0.56
0.3	2000	10000	0.54	0.3	2000	20000	0.63
0.4	2000	10000	0.58	0.4	2000	20000	0.67
0.5	2000	10000	0.61	0.5	2000	20000	0.70
0.6	2000	10000	0.64	0.6	2000	20000	0.72
0.1	10000	10000	0.61	0.1	10000	20000	0.70
0.2	10000	10000	0.69	0.2	10000	20000	0.78
0.3	10000	10000	0.73	0.3	10000	20000	0.81
0.4	10000	10000	0.74	0.4	10000	20000	0.83
0.5	10000	10000	0.76	0.5	10000	20000	0.84
0.6	10000	10000	0.76	0.6	10000	20000	0.85

Tabelas válidas para eucalipto (L = 13 Morgans), pinus (L = 15 Morgans) e café (L = 14 Morgans).

A seguir ilustra-se os valores de acurácia de quatro métodos de seleção em Eucalipto.

Acurácias seletivas dos métodos de seleção em Eucalipto.

Herdabilidade	Massal	Blup Individual	GWS1	GWS2
0.1	0.32	0.67	0.44	0.70
0.2	0.45	0.72	0.56	0.78
0.3	0.55	0.76	0.63	0.81
0.4	0.63	0.78	0.67	0.83
0.5	0.71	0.81	0.70	0.84
0.6	0.77	0.84	0.72	0.85

GWS1: Seleção Genômica Ampla usando 2.000 indivíduos genotipados para 20.000 marcas; GWS2: Seleção Genômica Ampla usando 10.000 indivíduos genotipados para 20.000 marcas.

A seguir ilustra-se os valores de ganho com seleção de quatro métodos de seleção em Eucalipto.

Ganhos genéticos (em unidades de desvio padrão genético aditivo por unidade de tempo (ano) associados aos métodos de seleção em Eucalipto.

Herdabilidade	Massal	Entre e Dentro*	Blup Individual	GWS1	GWS2
0.1	0.277	0.494	0.592	1.55	2.46
0.2	0.392	0.547	0.635	1.96	2.73
0.3	0.481	0.583	0.663	2.19	2.85
0.4	0.555	0.614	0.688	2.35	2.91
0.5	0.620	0.642	0.713	2.46	2.95
0.6	0.680	0.670	0.740	2.54	2.98

GWS1: Seleção Genômica Ampla usando 2.000 indivíduos genotipados para 20.000 marcas;
GWS2: Seleção Genômica Ampla usando 10.000 indivíduos genotipados para 20.000 marcas.
* de Famílias.

6.3 Populações de Estimação, Validação e Seleção

Na prática da seleção genômica ampla, três populações podem ser definidas: população de estimação, validação e seleção. Essas podem: (i) ser fisicamente distintas (3 populações diferentes); (ii) exercer duas funções ao mesmo tempo (uma só população usada para estimação e validação); (iii) exercer três funções ao mesmo tempo (uma só população usada para estimação, validação e seleção). Em geral, as estratégias (i) e (ii) são as mais usadas, embora a (iii) seja também muito usada no método G-BLUP (ver tópico 6.12). A Figura 1 ilustra a estratégia (ii).

População de Estimação. Também denominada população de descoberta, de treinamento ou de referência. Esse conjunto de dados contempla um grande número de marcadores avaliados em um número moderado de indivíduos (1.000 a 10.000, dependendo da acurácia desejada, conforme relatado no tópico anterior), os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Equações de predição (regressão múltipla aleatória) de valores genéticos genômicos são obtidas para cada caráter de interesse. Essas equações associam a cada marcador ou intervalo o seu efeito predito no caráter de interesse. Nessa população são descobertos, via marcadores, os marcadores que explicam os locos que controlam os caracteres, bem como são estimados os seus efeitos.

População de Validação. Quando fisicamente disjunta da população de estimação, esse conjunto de dados é menor do que aquele da população de descoberta e contempla indivíduos avaliados para os marcadores SNPs e para os vários caracteres de interesse. As equações de predição de valores genéticos genômicos são testadas para verificar suas acurácias nessa amostra independente. Para computar essa acurácia, os valores genéticos genômicos são preditos (usando os efeitos estimados na população de estimação) e submetidos a análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos marcadores, os erros dos valores genéticos genômicos e dos valores fenotípicos são independentes e a correlação entre esses valores é predominantemente de natureza genética e equivale à capacidade preditiva ($r_{y\hat{y}}$) da GWS em estimar os fenótipos, sendo dada pela própria acurácia seletiva ($r_{q\hat{q}}$) multiplicada pela raiz quadrada da herdabilidade individual (h), ou seja, $r_{y\hat{y}} = r_{q\hat{q}}h$, conforme demonstrado no tópico 6.5. Assim, para estimação da própria acurácia deve-se obter $\hat{r}_{q\hat{q}} = \hat{r}_{y\hat{y}} / \hat{h}$. Isso é válido quando são usados os valores fenotípicos brutos para cômputo da correlação. Quando são usados valores genotípicos preditos com base nos fenótipos em vez dos valores fenotípicos brutos, a herdabilidade deve ser substituída pela confiabilidade. De maneira geral adota-se a estratégia (ii), segundo um esquema Jacknife de validação cruzada. Segundo Meuwissen (2007), quando dezenas a centenas de milhares de haplótipos são estimados, existe o risco de superparametrização, ou seja, erros nos dados serem explicados pelos efeitos de marcadores. A validação cruzada é então de grande importância para contornar esse problema.

População de Seleção. Esse conjunto de dados contempla apenas os marcadores avaliados nos candidatos à seleção. Essa população não necessita ter os seus fenótipos avaliados. As equações de predição derivadas na população de descoberta são então usadas na predição dos valores genéticos genômicos (VGG) ou fenótipos futuros dos candidatos à seleção. Mas a acurácia seletiva associada refere-se àquela calculada na população de validação. Na Figura 2 é Ilustrada a aplicação da seleção genômica ampla no melhoramento genético de suínos (Goddard e Hayes, 2009).

A seguinte estratégia e sequência de análise envolvendo as populações de estimação e validação podem ser indicadas: compute a predição dos valores genéticos genômicos (VGG) usando todos os marcadores e calcule a correlação $r_{VGG,y}$ entre VGG e y, em que $\hat{r}_{y\hat{y}} = r_{VGG,y}$; ordene os marcadores por maiores módulos dos efeitos estimados dos marcadores; crie arquivos com subconjuntos dos marcadores com maiores módulos dos efeitos estimados dos marcadores (100, 250, 500, 1000, 1500, 2000, ...); analise todos esses arquivos e compute as correlações $r_{VGG,y}$ e escolha o arquivo ótimo que maximiza a $r_{VGG,y}$; faça a validação nesse arquivo ótimo com k = 2 no processo Jacknife descrito a seguir; faça a validação nos outros arquivos menores que o ótimo e em um maior que o ótimo para ver tendências (usar k = 2); compute os valores de $\hat{r}_{q\hat{q}} = \hat{r}_{y\hat{y}} / \hat{h}$ nas validações realizadas.

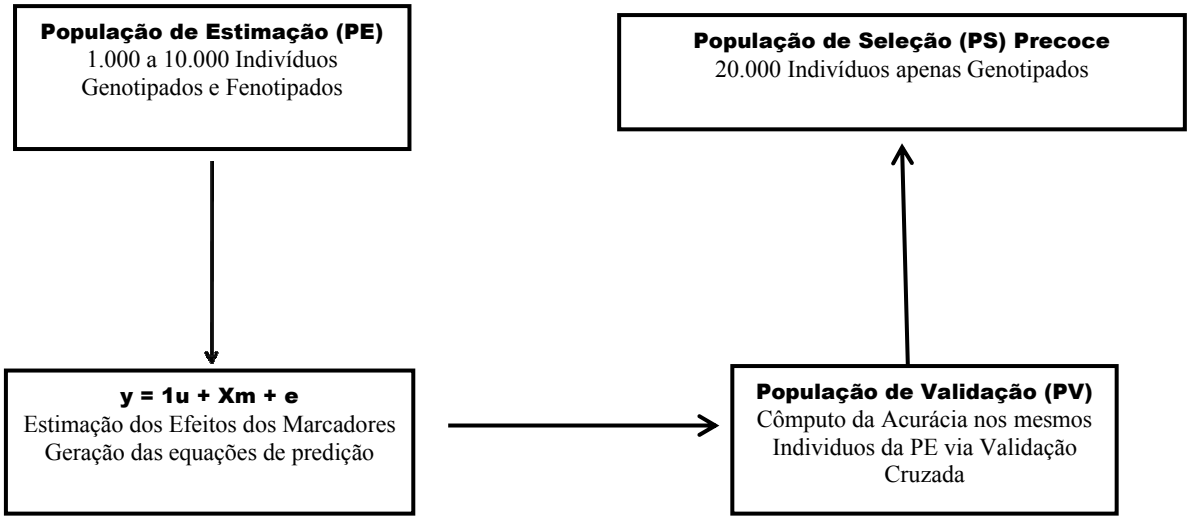


Figura 1 – Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético (Resende et al., 2010).

6.4 População de Validação e Jacknife

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

grupo de tamanho k foi removido. Com $k = 1$, $N = g$ e $(g - 1)k = g - 1 = N-1$, de forma que $\hat{\theta}_i$ refere-se à amostra em que foi omitida a observação i (Resende, 2008). Validações com k = 1 e k =2 tendem a conduzir aos mesmos valores de acurácia na população de validação. Assim, não há necessidade de usar k = 1, sendo que valores maiores são também suficientes para a validação cruzada.

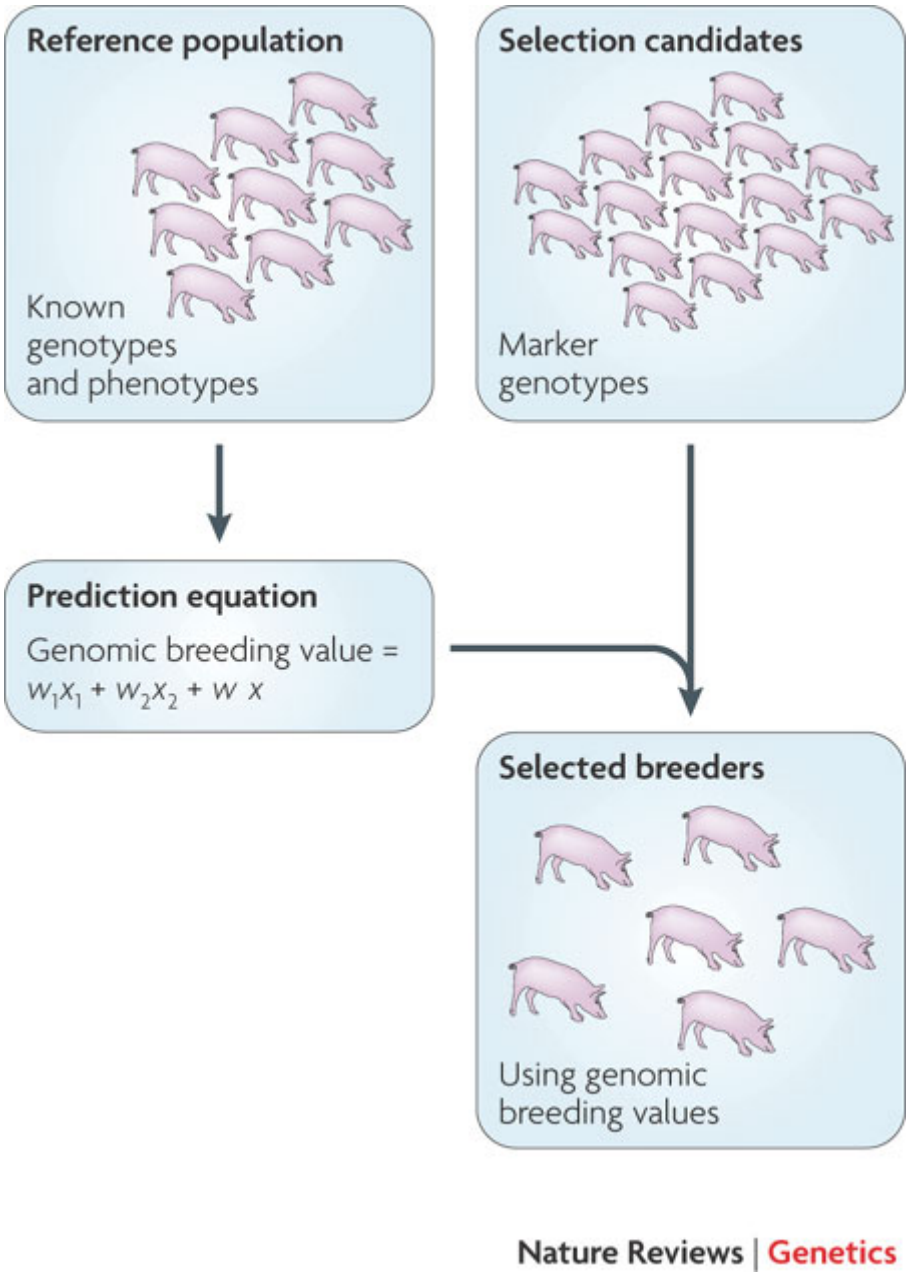


Figura 2 – Ilustração da aplicação da seleção genômica ampla no melhoramento genético de suínos (Goddard e Hayes, 2009).

6.5 Correlação e Regressão entre Valores Genéticos Preditos e Fenótipos na População de Validação

Os coeficientes de correlação e regressão envolvendo valores observados e preditos são medidas práticas da capacidade dos métodos predizerem de forma acurada e não viesada, respectivamente. A correlação fornece a capacidade preditiva, a qual equivale ao produto da acurácia pela raiz quadrada da herdabilidade. O coeficiente de regressão equivale algebricamente a 1. Coeficientes de regressão abaixo de 1 indicam que os valores genéticos são superestimados e apresentam variabilidade além da esperada e acima de 1 indicam que os valores genéticos estimados apresentam variabilidade aquém da esperada. Não vício é importante quando a seleção envolve indivíduos de muitas gerações usando efeitos dos marcadores estimados em uma só geração. Coeficientes de regressão próximos de 1 indicam que as avaliações são não viesadas e são efetivas em predizer as reais magnitudes das diferenças entre os indivíduos em avaliação. A seguir são apresentadas algumas definições paramétricas importantes envolvendo os valores fenotípicos corrigidos (y_c) e os valores genéticos genômicos preditos na população de validação ($\hat{g}_V$).

A. Covariância

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

B. Variâncias

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

C. Correlação

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

D. Regressão de y_c em $\hat{g}_V$

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

E. Acurácia

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

F. Confiabilidade

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

O erro padrão da estimativa da acurácia pode ser computado por $s(r_{\hat{g}g}) = [(1 - r^2_{\hat{g}g})/(N-2)]^{1/2}$. O coeficiente de regressão tem valor esperado igual a 1 e nessa situação indica que a predição foi não viesada. Assim sendo, pode-se também usar o coeficiente de regressão para estimar a herdabilidade (h^2_c) a ser empregada.

Vários valores de herdabilidade são avaliados e aquele que fornecer uma regressão igual a 1 deve ser escolhido como melhor estimativa. Se a regressão der resultado menor que 1 o valor de herdabilidade avaliado foi de alta magnitude e deve ser diminuído até a convergência para 1. Se a regressão der resultado maior que 1 o valor de herdabilidade avaliado foi de pequena magnitude e deve ser aumentado até a convergência para 1.

6.6 Métodos estatísticos na seleção genômica ampla

No contexto da seleção assistida por marcadores e da predição genômica, o método de quadrados mínimos (LS) apresenta sérias deficiências. Segundo Gianola et al. (2003), o índice de seleção (calculado como regressão envolvendo escores moleculares) apresentado por Lande e Thompson (1990) para a MAS falha quando formulado em uma maneira vetorial. Isto porque a matriz de covariância dos escores moleculares é singular uma vez que a distribuição dos valores ajustados da regressão é definida somente no espaço p-dimensional (número de covariáveis) e não no espaço n-dimensional (número de indivíduos com escores moleculares). Então, o índice de seleção conduz a um infinito número de soluções.

Outra dificuldade que surge é quando o número de marcadores iguala ou supera o número de indivíduos genotipados. Nessa situação, a colinearidade das variáveis preditoras causa problemas de identificação paramétrica e algum método de redução dimensional deve ser usado, como por exemplo a decomposição por valor singular. Outro problema é a própria inadmissibilidade (não propiciam mínimo erro quadrático médio) dos estimadores LS, resultado esse que desmorona a estimação por LS e por GLS (quadrados mínimos generalizados). Assim, o método LS não é recomendado na MAS e na GWS. Na GWS, devido ao número de marcadores maior do que o número de indivíduos, existe uma escassez de graus de liberdade para estimar os efeitos de todos os marcadores. Uma solução para contornar essa questão é usar o método da regressão ridge (RR de Whittaker et al., 2000) ou assumir os efeitos de marcadores como aleatórios ao invés de fixos. O ajuste de efeitos aleatórios não consome graus de liberdade, e então, os efeitos de todos os marcadores podem ser estimados simultaneamente. E isto conduz ao procedimento RR-BLUP, relatado a seguir.

O método LS é ineficiente devido a: impossibilidade de estimar todos os efeitos simultaneamente, pois o número de efeitos a estimar é maior do que o número de dados; estimando um efeito de cada vez e verificando a sua significância, conduz a superestimativas dos efeitos significativos; a acurácia do método é baixa; somente QTLs de grande efeito serão detectados e usados e, conseqüentemente, nem toda a variação genética será capturada pelos marcadores. O método LS assume distribuição *a priori* para os QTLs, com variância infinitamente grande, fato que é incompatível com a conhecida variância genética total. O RR-BLUP assume os efeitos de QTL com distribuição normal com variância constante através dos segmentos cromossômicos. A distribuição dos efeitos de QTL é conhecida em poucos caracteres e espécies. Em gado bovino leiteiro, Goddard & Hayes (2007) relatam a presença de 150 QTLs para o caráter produção de leite e estimaram a distribuição de seus efeitos como aproximadamente exponencial.

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 (Tabela 25) podem ser divididos em três grandes classes: regressão explícita, implícita e 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 de redes neurais, RKHS (*Reproducing Kernel Hilbert Spaces*, que é um método semi-paramétrico (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 componentes independentes, quadrados mínimos parciais e de componentes principais.

Tabela 25. 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
		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 Componentes Independentes	

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 25). 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.

Hayes et al. (2009) avaliaram a efetividade prática da seleção genômica em gado de leite nos Estados Unidos, Austrália e Nova Zelândia. Concluíram que o método BLUP mostrou-se aproximadamente igual a outros métodos mais complexos, em termos de acurácia. Adicionalmente, o método BLUP é vantajoso porque a única informação *a priori* necessária é uma estimativa da variância genética aditiva do caráter. Os autores relataram também a importância da inclusão do efeito poligênico no modelo de avaliação genética, como forma de capturar e selecionar QTLs de baixa freqüência não capturados pelos marcadores. Habier et al. (2007) compararam os métodos de quadrados mínimos (denominado por eles como regressão fixa ou FR-LS), BLUP (denominado por eles como regressão aleatória ou RR-BLUP) e Bayes B, em termos de acurácia seletiva na seleção ao longo prazo, após várias gerações depois da predição dos efeitos genéticos dos marcadores. Nessa situação, a acurácia tende a diminuir devido à modificação das relações de parentesco (em relação ao parentesco na geração de estimação dos efeitos genômicos) mas, há

um componente persistente da acurácia devido ao LD. Os resultados mostraram que o decréscimo na acurácia devido à modificação das relações de parentesco é maior no método RR-BLUP. Inicialmente, os métodos RR-BLUP e Bayes B apresentaram acurácia similar. Mas, após 11 gerações, o método Bayes B superou o RR-BLUP.

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 pode não ser 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). Outro método eficiente nessa classe é o ICR (Azevedo et al., 2012).

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 podem necessitar serem 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 texto contempla os métodos BayesA, BayesB, Fast BayesB, BayesCπ, BLASSO, IBLASSO, RR-BLUP, RR-BLUP-Het, MCMC-BLUP, PLSR, PCR, ICR 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.

6.7 Método RR-BLUP

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

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

significância para marcas individuais. Isto contrasta com o método da MAS proposto por Lande e Thompson (1990). Uma comparação entre as três proposições pode ser vista na Tabela 26.

Tabela 26. Comparação entre as três proposições de seleção auxiliada por marcadores.

Autores	Método	População	Número de Marcadores (m)	Teste de Significância	Extensão para o Enfoque Bayesiano
Lande e Thompson (1990)	MAS – Indice de Seleção Reg. Mult.	Dentro de família ou cruzamento	Muito menor que tamanho do cruzamento (N): m << N	Sim	Não
Whittaker et al. (2000)	MAS – Ridge Regression	Dentro de família ou cruzamento	Maior ou igual ao tamanho do cruzamento (N): m >= N	Não	Não
Meuwissen et al. (2001)	GWS – RR-BLUP	Toda a População	Muito maior que tamanho da população de estimação (N): m >> N	Não	Sim

Verifica-se pela Tabela 26, que a inovação de Meuwissen et al. (2001) não foi em termos de metodologia estatística mas, em termos conceituais enfatizando o uso do conceito de desequilíbrio de ligação em nível populacional e não apenas dentro de família e o não uso de testes de significância para marcas. E o maior mérito foi a demonstração, via simulação, do fato de que a GWS pode realmente funcionar na prática.

A GWS enfatiza também o não uso de significância estatística para a seleção de marcas. Esse ponto distingue a GWS da GWAS (Genome Wide Association Studies), a qual procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses visando detectar efeitos com significância estatística. A GWAS sofre com a alta taxa de falsos negativos devido ao uso de pontos de corte muito rigorosos visando evitar a ocorrência de falsos positivos. A GWS equivale à GWAS aplicada sobre todos os locos simultaneamente e baseando-se em estimação e predição em vez de teste de hipótese. Dessa forma consegue explicar parte muito maior da variabilidade genética e evitar a chamada herdabilidade faltante ou perdida (missing heritability), típica dos estudos de análise de ligação e de associação.

A distinção entre regressão fixa, regressão ridge e regressão aleatória, em um modelo usando somente fenótipos, está associada ao parâmetro de penalização λ^*, o qual é dado por $\lambda^* = (1-h^2)/h^2$. Valores pequenos de λ^* já são suficientes para reduzir o impacto da multicolinearidade presente entre as covariáveis na matriz W′W, que é aproximadamente singular. Valor de λ^* igual a zero (valor de h^2 igual a 1) caracteriza a regressão fixa. Valores de λ^* pequenos (0,01 a 1) caracterizam a regressão ridge e valores altos de λ^* (maiores que 0,1) caracterizam a regressão aleatória.

A predição via RR-BLUP é descrita a seguir com base em Resende (2007; 2008). O seguinte modelo linear misto geral é ajustado para estimar os efeitos dos marcadores: $y = Xb + Wm + e$, em que y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, m é o vetor dos efeitos aleatórios de marcadores e e refere-se ao

vetor de resíduos aleatórios. X e W são as matrizes de incidência para *b* e *m*. 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 diplóide. Outra forma equivalente de codificar é usar os valores -1, 0 e 1. As equações de modelo misto genômicas para a predição de *m* via o método RR-BLUP equivalem a:

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

O valor genético genômico global do indivíduo *j* é dado por $VGG = \hat{y}_j = \sum_i w_{ij}\hat{m}_i$, em que W_i equivale a 0, 1 ou 2 para os genótipos mm, Mm e MM, respectivamente, para o marcador bialélico e codominante i (SNP). O componente w_{ij} refere-se ao elemento i da linha j da matriz W, referente ao indivíduo j.

Modelos com efeitos de dominância (d) podem também serem ajustados. Esses são da forma *y = Xb + Wm + Td + e*. Nesse caso, os elementos de W 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 MM, Mm e mm, respectivamente. Valores de W e T codificados dessa forma são independentes e apresentam média zero e variância 1. Se os elementos de W 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.

As equações de predição apresentadas acima assumem *a priori* que todos os locos explicam iguais quantidades da variação genética. Assim, a variação genética explicada por cada loco é dada por σ_g^2/n_Q, em que σ_g^2 é a variação genética total e n_Q é o número de locos (quando cada loco está perfeitamente marcado por uma só marca). A variação genética σ_g^2 pode ser estimada por REML sobre os dados fenotípicos da maneira tradicional ou pela própria variação entre os marcadores ou segmentos cromossômicos de QTL, conforme descrito adiante. A quantidade n_Q é dada por $n_Q = 2\sum_i^n p_i(1-p_i)$.

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

A GWS melhora a acurácia da estimativa $\hat{g}_d$, referente aos efeitos da segregação mendeliana dentro de famílias e é o método que explora adequadamente a segregação de amostragem mendeliana que ocorre por ocasião da formação de gametas. Uma vez que a GWS avalia diretamente o DNA associado (via marcadores) a cada loco de todo o caráter poligênico, avalia diretamente cada segregação em nível individual e não em nível médio. Avaliando diretamente o genótipo dos filhos, permite conhecer cada segregação. Conforme Goddard & Hayes (2007), sob o modelo infinitesimal com grande número de locos de pequeno efeito, o BLUP genômico prediz os valores genéticos de maneira mais acurada do que o BLUP tradicional baseado em *pedigree* e dados fenotípicos. A GWS enfatiza mais o termo referente à segregação mendeliana $\hat{a}_d$, dando mais peso a esse componente do que o faz o BLUP tradicional. Isso leva à seleção de menos indivíduos aparentados do que o faz o BLUP, reduzindo assim o incremento da endogamia na população.

A matriz de parentesco realizada G pode ser também computada à parte e incorporada nas equações de modelo misto do BLUP tradicional, conforme o modelo (iii) descrito a seguir. Nesse caso, ela é dada por $G = (W^*W^{*\prime})/[2\sum_i^n p_i(1-p_i)]$ (para SNPs), em que p_i é a freqüência de um dos alelos do loco i e W* refere-se à matriz W corrigida para suas médias em cada loco ($2p_i$). Para garantir G como uma matriz positiva definida pode-se obter $G^P = G + 10^{-6} I$, em que I é uma matriz identidade. O coeficiente de endogamia genômico para o indivíduo i é dado por $G_{ii} - 1$. Outra forma de obter G é via $G = W^* D W^{*\prime}$, em que D é diagonal com D_{ii} dado por $D_{ii} = 1/\{n[2p_i(1-p_i)]\}$, em que n é o número de marcadores.

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

A predição de valores genéticos genômicos via BLUP pode ser computada via 3 métodos equivalentes:

(i) Via RR-BLUP, conforme especificado acima, em que:

$\hat{g} = W\hat{m} = W(W'R^{-1}W + I\lambda)^{-1}W'R^{-1}(y - X\hat{b})$, visto que $\hat{m} = (W'R^{-1}W + I\lambda)^{-1}W'R^{-1}(y - X\hat{b})$. O vetor aleatório de erros tem variância igual a $Var(e) = R\sigma_e^2$. R é uma matriz diagonal de pesos para ponderar y com diferentes confiabilidades. Com confiabilidades altas e homogêneas

(maiores que 0,85), pode-se considerar R = I e o sistema simplifica para $\hat{m} = (W'W + I\lambda)^{-1}W'(y - X\hat{b})$, em que $\lambda = \frac{\sigma_e^2}{\sigma_m^2} = \frac{\sigma_e^2}{(\sigma_g^2 / n_Q)}$.

(ii) Via BLP ou índice de seleção (com G genômica e $\hat{b}$ estimado via quadrados mínimos generalizados, o que é garantido quando y contem valores genéticos desregressados), em que: $\hat{g} = G[G + R(\sigma_e^2 / \sigma_g^2)]^{-1}(y - X\hat{b})$. Se necessário os efeitos dos marcadores podem ser obtidos por $\hat{m} = \{W'/[2\sum_i^n p_i(1-p_i)]\}[G + R(\sigma_e^2 / \sigma_g^2)]^{-1}(y - X\hat{b})$. Com R = I, uma observação por indivíduo e dividindo ambos os lados da equação por G o sistema simplifica para $\hat{g} = [I + G^{-1}(\sigma_e^2 / \sigma_g^2)]^{-1}(y - X\hat{b})$. Nesse caso, os métodos do índice de seleção (Henderson, 1963; Resende et al., 1990; Lopes, 2005) e de modelos mistos (Henderson, 1973) são idênticos para a seleção genômica.

(iii) Via BLUP Modelo Equivalente, em que:

$\hat{g} = [R^{-1} + G^{-1}(\sigma_e^2 / \sigma_g^2)]^{-1}R^{-1}(y - X\hat{b})$. Com R = I e uma observação por indivíduo o sistema simplifica para $\hat{g} = [I + G^{-1}(\sigma_e^2 / \sigma_g^2)]^{-1}(y - X\hat{b})$.

Na situação em que os marcadores não explicam toda a variação genética, o modelo pode ser estendido para englobar o efeito poligênico residual g* (variação genética não explicada pelos marcadores). Esse modelo é dado por $y = Xb + Wm + Tg^* + e$, em que T é a matriz de incidência para g*. Com o uso de mapa denso de marcadores a inclusão dos efeitos poligênicos, g* não aumenta a acurácia da GWS (Calus & Veerkamp, 2007). No entanto, para capitalizar o ganho genético no longo prazo, a inclusão desses efeitos é recomendada (Muir, 2007). No longo prazo, o BLUP tradicional obtém informação no genoma inteiro em cada geração. A GWS sem o efeito poligênico seleciona de forma muito acurada para a mesma parte do genoma em cada geração. Uma forma de aliviar esse problema é por meio da re-estimação dos efeitos de marcadores, freqüentemente, visando à exploração de novas associações de marcadores-QTL.

Para implementação do procedimento RR-BLUP/GWS são necessários: W, X, y e $\lambda = \sigma_e^2 / \sigma_{gi}^2 = \sigma_e^2 / (\sigma_g^2 / n_Q)$. O vetor y refere-se a fenótipos corrigidos; a matriz W refere-se à contagem de doses dos marcadores moleculares; X é um vetor conhecido composto de valores 1; λ depende de componentes de variância (herdabilidade ou confiabilidade da seleção) e do número de segmentos cromossômicos n_Q. A seguir são descritos cada um desses elementos, conforme Resende et al. (2010).

6.8 Formas de parametrização da matriz de incidência genotípica

Parametrização 1

A matriz de incidência W contém os valores 0, 1 e 2 para o número de alelos do marcador (ou do suposto QTL) em um indivíduo diplóide. Com marcadores codominantes a média e variância da variável indicadora W associada à matriz de incidência são dadas por:

Média da variável W = 0 x p^2 + 1 x 2p(1-p) + 2 x $(1-p)^2$ = 2p

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

Assumindo os alelos de cada marca como em equilíbrio de Hardy-Weinberg na população, o cálculo das frequências alélicas é realizado conforme o quadro a seguir, sendo p dado por $p = N_2/N + (1/2)\ N_1/N$, sendo o calculo realizado para cada coluna de marcador no arquivo de dados em que N_2 é o numero de códigos 2 na referida coluna no arquivo.

Genótipos	Código	Contagem	Frequencia	Cálculo da Frequencia de M
MM	2	N_2	p^2	$N_2/N = p^2$
Mm	1	N_1	2p(1-p)	$(1/2)\ N_1/N = p(1-p)$
mm	0	N_0	$(1-p)^2$	0
Soma	-	N	1	$p = N_2/N + (1/2)\ N_1/N$

Os valores de W 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 $\sigma_g^2 = 2\sum_i^m p_i(1-p_i)\sigma_m^2$. Dessa forma, os valores de W_i devem ser subtraídos pela média de W (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 $\hat{g} = W\hat{m}$. Para os indivíduos com dados perdidos de marcas, seus valores na matriz W devem ser o valor esperado 2p, que, centrados, transformam-se em zero.

É importante relatar que os efeitos dos QTLs via marcadores m são assumidos com distribuição normal $(m \sim (0, I\sigma_m^2))$ e os alelos marcadores são assumidos como amostras de uma distribuição Bernoulli com média p e variância p(1-p). O número de alelos em um indivíduo diploide (variável W) apresenta distribuição Binomial com média 2p e variância 2p(1-p) (2 provas Bernoulli).

Parametrização 2

Adicionalmente, pode-se padronizar (usando $Var(W_i) = 2p_i(1-p_i)$) os dados dos marcadores na matriz W, da seguinte forma para cada elemento W_i da matriz, referente ao loco i:

$W_i = (0 - 2p_i) / (Var(W_i))^{1/2}$ se o indivíduo é homozigoto para o primeiro alelo (mm);
$W_i = (1 - 2p_i)/(Var(W_i))^{1/2}$ se o indivíduo é heterozigoto (Mm);
$W_i = (2 - 2p_i)/2/(Var(W_i))^{1/2}$ se o indivíduo é homozigoto para o segundo alelo no loco marcador (MM);
$W_i = 0$ 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 W com W_i ajustado é igual a 1, obtendo-se uma variável com média zero e variância unitária. Sendo m o efeito do marcador na população, a variância devida ao marcador é dada por $Var(W_i m) = Var(W_i)\ Var(m)$. Com a transformação acima, $Var(W_i) = 1$ e portanto, $Var(W_i m) = Var(m)$. Em outras palavras, modelando a variância do efeito do marcador, modela-se diretamente a variância do marcador, independentemente de sua frequência. Assim, com centralização e padronização $\sigma_g^2 = n\sigma_m^2$. Dessa forma, no método RR-BLUP deve-se usar $n_Q = n$ e os efeitos genéticos aditivos dos indivíduos são dados por $\hat{g} = W\hat{m}$.

Essa padronização reflete positivamente na composição da matriz de parentesco genômico G usada no G-BLUP, a qual conterá a média ponderada das relações de parentesco estimadas de cada loco marcador, em que os pesos da ponderação são função da inversa da PEV (variância do erro de predição) associada à variável indicadora W em cada marcador. No caso, a PEV é dada por $PEV_i(W) = Var(W_i) = 2p_i(1-p_i)$. E a matriz G é dada por $G = WW'/n$. Essa parametrização é melhor do que a 1 e 3, segundo Meuwissen et al. (2011). Todavia é equivalente a parametrização $G = WD_pW'$, mencionada em tópico anterior, em que $diag(D_P) = 1/[n2p_i(1-p_i)]$. Pela parametrização 1, tem-se no G-BLUP: $G = WW'/[2\sum_i^m p_i(1-p_i)]$, a qual é melhor que a 2, segundo Endelman e Jannink (2012).

Parametrização 3

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, no método RR-BLUP, $n = 2\sum_i^m p_i(1-p_i)$ e o efeito genético aditivo do indivíduo j é dado por $\hat{g}_j = \sum_i^m [I(w_{ij} = 1)(2p_i\hat{m}_i) + I(w_{ij} = 0)(p_i\hat{m}_i - q_i\hat{m}_i) + I(w_{ij} = -1)(-2q_i\hat{m}_i)$.

Para garantir G como uma matriz positiva definida no G-BLUP, pode-se obter $G^P = G + 10^{-6} I$, em que I é uma matriz identidade, ou usar $G^P = \omega G + (1-\omega) A$, ou usar $G^P = \omega G + (1-\omega) I$, em que :

$\omega = \frac{Var(\rho_g)}{Var(\rho_g) + 0.125/n_m}$, *em que* $\omega = \frac{0.05^2}{0.05^2 + 0.125/n_m}$ *se* $Var(\rho_g) = 0.05^2$ (*bovinos*)

, conforme tópico 6.28. Assim, se nm = 1000, w = 0.95 é o peso dado a G.

Parametrizações para marcadores DArT são apresentadas por Resende et al. (2010).

6.9 Correção dos Fenótipos

Os fenótipos devem ser corrigidos para os efeitos ambientais e dos genitores. Assim, os valores genéticos devem ser preditos e posteriormente desregressados e corrigidos para os efeitos dos genitores. Devem ser desregressados por 3 motivos: não pode haver duas regressões, uma baseada em pedigree e outra baseada em marcadores; a matriz A baseada em pedigree é menos precisa que a WW′ baseada em marcas; presença de genes de grande efeito presentes em um dos genitores. Adicionalmente devem ser corrigidos para os efeitos genéticos dos genitores, trabalhando-se basicamente com o efeito da *segregação mendeliana desregressada*, já que o dado ideal para a população de treinamento deve ser o *mérito genético verdadeiro de indivíduos não aparentados*. E o efeito da segregação mendeliana proporciona isso: analise da associação de alelos de marcas e de QTL, ou seja, captura efeitos genéticos explicados pelo desequilíbrio de ligação e não pelo parentesco ou genealogia.

Uma forma explícita de se fazer isso, parcialmente, é a consideração do pedigree via ajuste de g*, o vetor de efeitos poligênicos por meio do modelo $y = Wb + Xm + Tg^* + e$, em que T é a matriz de incidência para a*. Sem a correção mencionada acima ou o ajuste de g*, os marcadores podem estar capturando apenas o parentesco (estrutura de população) entre os indivíduos e não necessariamente o desequilíbrio de ligação com os genes propriamente ditos. Nesse caso, a acurácia da validação em uma amostra independente (indivíduos de outras famílias) da população e, também, em indivíduos de outras gerações poderá ser baixa, ao contrário do que teria sido predito.

Outra forma de realizar esse ajuste para estrutura de população 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. Esse ajuste é adequado quando a acurácia da avaliação dos genitores é próxima de 1. Várias alternativas de correção de fenótipos são apresentadas no Capítulo 1, tópico 1.12.

Outra opção de correção para estrutura de família é segundo o modelo descrito no final do item 4.3. Nesse caso, ajusta-se os primeiros autovetores (associados aos maiores autovalores) de G como covariáveis de efeitos fixos, conforme descrito no início do item 6.30.

Quando se tem um catalogo de valores genéticos com diferentes acurácias, o procedimento de obtenção dos valores fenotípicos desregressados e corrigidos para os efeitos genéticos dos genitores envolve os seguintes passos (Garrick et al, 2009; Resende et al., 2010):

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

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

$\lambda^* = (1-h^2)/h^2$, em que h² é a herdabilidade ao nível de indivíduo.

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

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

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

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

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

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

$\delta = (0.5 - r_{gm}^2)/(1 - r_i^2)$

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

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

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

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

(iv) Obtenção da quantidade desconhecida y_i*:*

Resolução para y_i, do sistema $\begin{bmatrix} Z_{gm}'Z_{gm} + 4\lambda^* & -2\lambda^* \\ -2\lambda^* & Z_i'Z_i + 2\lambda^* \end{bmatrix} \begin{bmatrix} \hat{g}_{gm} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} y_{gm} \\ y_i \end{bmatrix}$. Assim, $y_i = (-2\lambda^*)\hat{g}_{gm} + (Z_i'Z_i + 2\lambda^*)\hat{g}_i$, o qual representa a informação do indivíduo, agora corrigida para o valor genético médio de seus genitores.

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

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

Assim, para obtenção de $\hat{g}_i^*$ necessita-se da herdabilidade h², das confiabilidades (quadrado da acurácia) das avaliações dos três indivíduos (r_{gj}^2, r_{gk}^2 e r_i^2) e dos efeitos genéticos preditos dos três indivíduos ($\hat{g}_i$, $\hat{g}_j$ e $\hat{g}_k$).

Considere um caráter com h² de 0.20 e a avaliação genética de 3 indivíduos onde foram obtidos os seguintes resultados: $\hat{g}_i$ = 18, $\hat{g}_j$ =13 e $\hat{g}_k$ = 5; r_i^2 = 0.70 r_{gj}^2 = 0.90 e r_{gk}^2 = 0.80. Assim, são obtidos:

$r_{gm}^2 = (r_{gj}^2 + r_{gk}^2)/4 = (0.90 + 0.80)/4 = 0.425;$

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

$\lambda^* = (1 - h^2)/h^2 = 0.8/0.2 = 4;$

$\alpha = 1/(0.5 - r_{gm}^2) = 1/(0.5 - 0.425) = 13.3333;$

$\delta = (0.5 - r_{gm}^2)/(1 - r_i^2) = (0.5 - 0.425)/(1 - 0.70) = 0.25.$

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

$Z_{gm}'Z_{gm} = \lambda^*(0.5\alpha - 4) + 0.5\lambda^*(\alpha^2 + 16/\delta)^{1/2} = 4(0.5 \;\; 13.3333 - 4) + 0.5 \;\; 4(13.3333^2 + 16/0.25)^{1/2} = 41.765$

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

$Z_i'Z_i = \delta Z_{gm}'Z_{gm} + 2\lambda^*(2\delta - 1) = 0.25\,41.765 + 2 \;\; 4\,(2 \;\; 0.25 - 1) = 6.4412.$

Computa-se agora, seguindo o passo (iv), a quantidade $y_i = (-2\lambda^*)\hat{g}_{gm} + (Z_i'Z_i + 2\lambda^*)\hat{g}_i = (-2 \;\; 4)9 + (6.4412 + 2 \;\; 4)18 = 187.9423.$

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

$\hat{g}_i^* = y_i/(Z_i'Z_i) = 187.9423/6.4412 = 29.1780.$ Esse é o valor do indivíduo, a ser usado na análise genômica integrando o vetor y. Tal quantidade é equivalente a $\hat{g}_i^* = (\hat{g}_{i-gm})/r_i^{2*}$, ou seja, ao valor genético individual corrigido para a média de seus genitores e desregressado pela quantidade $r_i^{2*} = 1 - \lambda^*/(Z_i'Z_i + \lambda^*) = 1 - 4^*/(6.4412 + 4) = 0.6169$, que é a acurácia da estimação do efeito da segregação mendeliana.

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

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

genótipos dos marcadores estão disponíveis), a seleção precoce deve basear-se diretamente em $u + \hat{g}_i = u + W_i\hat{m}$, mas a acurácia da seleção é calculada com base em $u + \hat{g}_i = u + W_i\hat{m}^*$, em que $\hat{m}^*$ é o vetor de efeitos preditos dos marcadores, obtido via $\hat{g}_i^*$, usando valores genéticos desregressados e corrigidos para os efeitos de genitores. Po outro lado, $\hat{m}$ é o vetor de efeitos preditos dos marcadores, obtido usando valores genéticos apenas desregressados.

6.10 Relação entre Variância Genética e Variância dos Marcadores

A relação entre variância genética aditiva e variância dos efeitos dos marcadores é essencial na predição genômica. Tem-se que $Var(g_i) = Var(W_im) = Var(W_i)\ Var(m) = 2p_i(1\text{-}p_i)\ Var(m_i) = 2p_i(1\text{-}p_i)\ m_i^2$ equivale à variância genética devida ao loco i. Para vários locos, a variância genética aditiva total é dada por $\sigma_g^2 = \sum_i^n 2p_i(1-p_i)m_i^2$, a qual pode ser expressa também por $\sigma_g^2 = \sum_i^n U_iV_i$, em que $U_i = 2p_i(1-p_i)$ e $V_i = m_i^2$. A covariância entre U e V, denominada C_{UV} é dada por $C_{UV} = (\sum_i^n U_iV_i)/n - (\sum_i^n U_i/n)(\sum_i^n V_i)/n)$ e refere-se à covariância entre frequências alélicas e magnitudes dos efeitos alélicos. Rearranjando essa expressão tem-se $\sum_i^n U_iV_i = nC_{UV} + (\sum_i^n U_i)(\sum_i^n V_i)/n)$, de forma que $\sigma_g^2 = \sum_i^n U_iV_i = nC_{UV} + [\sum_i^n 2p_i(1-p_i)](\sum_i^n m_i^2)/n$. Sendo $(\sum_i^n m_i^2)/n = \sigma_m^2$, tem-se $\sigma_g^2 = [2\sum_i^n p_i(1-p_i)\sigma_m^2] + nC_{UV}$.

Assim, a variância entre marcadores (σ_m^2) obtida por REML, as freqüências alélicas e os efeitos dos marcadores preditos por BLUP podem ser usados na obtenção da variância genética aditiva total. Em alguns casos C_{UV} tende a zero, revelando ausência de correlação entre frequências e efeitos alélicos (Resende et al., 2010). Em outros casos, a quantidade m_i^2 é substituída por σ_m^2, pois a esperança de m_i^2 é a variância do efeito do marcador, ou seja, $E(m_i^2) = \sigma_m^2$. Assim, muitas das aplicações usam $\sigma_g^2 = [2\sum_i^n p_i(1-p_i)\sigma_m^2]$ e a variância entre marcadores dada por $\sigma_m^2 = (\sigma_g^2 - nC_{UV})/[2\sum_i^n p_i(1-p_i)]$ é simplificada para $\sigma_m^2 = \sigma_{gi}^2 = \sigma_g^2/[2\sum_i^n p_i(1-p_i)]$.

Na predição RR-BLUP/GWS necessita-se da quantidade $\lambda = \sigma_e^2/\sigma_{gi}^2 = \sigma_e^2/(\sigma_g^2/n_Q)$, em que n_Q é o número de locos controlando o caráter (assumindo que cada loco está perfeitamente marcado), o qual é desconhecido a priori. Sendo $\sigma_{gi}^2 = \sigma_g^2/[2\sum_i^n p_i(1-p_i)]$, n_Q pode ser tomado como $[2\sum_i^n p_i(1-p_i)]$. Alternativamente, λ pode ser expresso como $\lambda = n_Q(1-h^2)/h^2 = [2\sum_i^n p_i(1-p_i)](1-h^2)/h^2$. Assim, de posse de h^2 e das freqüências alélicas nos locos marcadores, obtém-se λ para uso nas equações de modelo misto.

A variância genética e a herdabilidade (h^2) podem ser computadas via dados fenotípicos ou via dados de marcadores e fenotípicos conforme descrito acima no

cômputo de σ_g^2. A h² a ser usada no RR-BLUP deve ser a herdabilidade ajustada ou dos dados corrigidos ($h_{aj}^2 = \sigma_a^2 / \sigma_{yaj}^2$), em que σ_{yaj}^2 é a variância fenotípica ajustada. Se y é corrigido para a média dos genitores o numerador de h_{aj}^2 deve conter apenas a variância genética devida à segregação mendeliana, ou seja, $h_{aj}^{2*} = (1/2)\sigma_a^2 / \sigma_{yaj}^2$ ou $h_{aj}^{2*} = (3/4)\sigma_a^2 / \sigma_{yaj}^2$ quando se conhece os dois genitores (famílias de irmãos germanos) ou apenas um dos genitores (famílias de meios irmãos), respectivamente. Essas herdabilidades podem ser expressas também em função da herdabilidade individual h², por meio das expressões $h_{aj}^{2*} = (1/2\ h^2)/(1/2\ h^2 + (1-h^2))$ para progênies de irmãos germanos e $h_{aj}^{2*} = (3/4\ h^2)/(3/4\ h^2 + (1-h^2))$ para progênies de meios-irmãos (Resende, 2002). Essas fórmulas mostram que o denominador de h_{aj}^{2*} também contempla apenas a variância genética devida à segregação mendeliana e não a variância genética total. Outra forma de expressar h_{aj}^{2*} é usar diretamente a confiabilidade ou quadrado da acurácia dos efeitos da segregação mendeliana (r_i^{2*}). Para cômputo do RR-BLUP e da acurácia da GWS, h_{aj}^{2*} pode ser tomada como a média dos r_i^{2*} dos indivíduos em análise.

Recomenda-se analisar inicialmente todo o conjunto de marcadores codominantes em todos os indivíduos fenotipados (população de estimação completa). Esse procedimento visa identificar os marcadores com maiores efeitos em módulo, objetivando rodar análises com subgrupos menores de marcadores e determinar quantos e quais marcadores maximizam a acurácia seletiva. O número ótimo de marcadores é um compromisso entre maior informatividade (maior acurácia, pela maior captura de genes) e menor precisão (menor acurácia, pelo menor tamanho amostral por efeito estimado) com o aumento do número de marcadores. Posteriormente, a validação deve ser realizada usando apenas a fração de marcadores que maximiza a acurácia, usando n como o somatório $[2\sum_i^n p_i(1-p_i)]$ nesse subconjunto de marcadores. Também σ_m^2 e h² devem ser recalculadas, sendo que essa h² pode ser menor do que aquela calculada anteriormente. Mas a h² usada para computar a acurácia a partir da capacidade preditiva, via $r_{q\hat{q}} = r_{y\hat{y}} / h$, deve ser a h² total, estimada dos próprios dados fenotípicos. Essa tende a ser similar a h² estimada via marcadores, quando se usa o total de marcas em grande número. Esse procedimento de seleção indireta de covariáveis (denominado RR-BLUP_B por Resende et al. 2010 e Resende Jr. et al. 2012) é recomendável, pois tende a produzir acurácia mais alta, similar à obtida pelos métodos Bayesianos. Dessa forma, ambas as abordagens assumem que muitos dos marcadores apresentam efeitos zero. O aumento ou diminuição da acurácia da GWS via RR-BLUP é um compromisso ou balanço entre acréscimo da quantidade de informação útil via uso de maior número de locos marcadores e diminuição do tamanho de amostra efetivo para estimar o efeito de cada loco, ou seja, menor número de indivíduos por loco a ser estimado (menor N/n).

O número reduzido de marcadores explicando grande parte da variação genética ou da acurácia máxima possível é muito interessante do ponto de vista prático. Nesse caso, arranjos de DNA com baixa densidade de marcadores previamente selecionados poderiam ser usados nas populações de seleção. Na Austrália, a acurada predição de valores genéticos genômicos em gado leiteiro pode ser realizada com chips de SNP contendo 1000 (propiciando 85% da acurácia obtida com 42500 SNP) a 5000 (propiciando 95% da acurácia obtida com 42500 SNP) SNP igualmente espaçados (Moser et al., 2010). Uma alternativa ao uso de marcadores previamente selecionados é o uso de marcadores igualmente espaçados e em maior número do que aqueles selecionados. Isso permite atender a vários caracteres e pode conduzir ao uso generalizado da GWS em várias espécies e países.

6.11 Exemplo via RR-BLUP/GWS

Considere o pequeno exemplo a seguir, referente à avaliação de 5 indivíduos para o caráter diâmetro e genotipagem para 7 marcas, em que são apresentados o número de um dos alelos de cada loco marcador.

Individuo	Diâmetro	Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
1	9.87	2	0	0	0	2	0	0
2	14.48	1	1	0	0	1	1	0
3	8.91	0	2	0	0	0	0	2
4	14.64	1	0	1	0	1	0	0
5	9.55	1	0	0	1	1	1	0

Os efeitos genéticos dos marcadores são obtidos resolvendo-se

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

Tem-se as seguintes matrizes:

$$W=\begin{bmatrix}2&0&0&0&2&0&0\\1&1&0&0&1&1&0\\0&2&0&0&0&0&2\\1&0&1&0&1&0&0\\1&0&0&1&1&1&0\end{bmatrix}\quad y=\begin{bmatrix}9.87\\14.48\\8.91\\14.64\\9.55\end{bmatrix};\qquad X=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}.$$

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

$$X'X=[5];\ X'W=[5\quad 3\quad 1\quad 1\quad 5\quad 2\quad 2];\ W'X=(X'W)'=[5\quad 3\quad 1\quad 1\quad 5\quad 2\quad 2]'$$

$$W'W+I=\begin{bmatrix}8&1&1&1&7&2&0\\1&6&0&0&1&1&4\\1&0&2&0&1&0&0\\1&0&0&2&1&1&0\\7&1&1&1&8&2&0\\2&1&0&1&2&3&0\\0&4&0&0&0&0&5\end{bmatrix}\quad \text{X'y = [57.45]}\quad W'y=\begin{bmatrix}58.4100\\32.3000\\14.6400\\9.5500\\58.4100\\24.0300\\17.8200\end{bmatrix}$$

Assim, tem-se:

$$\begin{bmatrix}\hat{b}\\ \hat{m}\end{bmatrix}=\begin{bmatrix} 5 & 5 & 3 & 1 & 1 & 5 & 2 & 2\\ 5 & 8 & 1 & 1 & 1 & 7 & 2 & 0\\ 3 & 1 & 6 & 0 & 0 & 1 & 1 & 4\\ 1 & 1 & 0 & 2 & 0 & 1 & 0 & 0\\ 1 & 1 & 0 & 0 & 2 & 1 & 1 & 0\\ 5 & 7 & 1 & 1 & 1 & 8 & 2 & 0\\ 2 & 2 & 1 & 0 & 1 & 2 & 3 & 0\\ 2 & 0 & 4 & 0 & 0 & 0 & 0 & 5 \end{bmatrix}^{-1}\begin{bmatrix} 57.4500\\ 58.4100\\ 32.3000\\ 14.6400\\ 9.5500\\ 58.4100\\ 24.0300\\ 17.8200\end{bmatrix}.$$

Os resultados são $\begin{bmatrix}\hat{b}\\ \hat{m}\end{bmatrix}=\begin{bmatrix} 12.4519\\ -0.3526\\ 0.2761\\ 1.4467\\ -1.3701\\ -0.3526\\ 0.5436\\ -1.63765\end{bmatrix}$, em que 12,4519 é a média geral e os demais valores são as estimativas dos efeitos genéticos dos marcadores.

O valor genético genômico dos indivíduos de uma população de seleção podem ser obtidos por $VGG = \hat{y}_j = \sum_i w_{ij}\hat{m}_i$. No caso, as predições para os 5 indivíduos são $VGG=\begin{bmatrix} -1.4104\\ 0.1145\\ -2.7230\\ 0.7415\\ -1.5317\end{bmatrix}$. Outras formas de obtenção de W são apresentadas no tópico 6.8.

6.12 G-BLUP com Dominância e Interação GE: Avaliação Simultânea Global

Modelo BLUP Individual Fenotípico

O modelo linear misto convencional, contemplando os efeitos fixos (b), genéticos aleatórios (a) e ambientais aleatórios (e) é dado por: $y = Xb + Zg + e$.

Modelo de QTL

Incluindo os efeitos (q) dos QTLs para cada loco *j*, o modelo torna-se $y = Xb + Zg^* + \sum_j Q_j q_j + e$,em que Q_j é uma matriz de incidência que relaciona os indivíduos aos alelos do loco *j*, e *q* contém os efeitos alélicos para cada loco. As matrizes de incidência Q não são conhecidas e nem as suas dimensões, dadas pelo número de alelos em cada loco. Também não é conhecido o número de locos que afeta o caráter. Isto contrasta com o primeiro modelo, em que as matrizes de incidência para *b* e *g* (X e Z, respectivamente) são conhecidas. Se Q fosse conhecida as equações de modelo misto poderiam ser usadas sem qualquer alteração.

Modelo GWS

Um outro modelo melhor poderia ser $y = Xb + \sum_j Q_j q_j + e$, no qual todos os locos seriam individualizados e não haveria necessidade de inclusão do resíduo genético poligênico ou infinitesimal (g^*). Como se conhecem apenas os marcadores esse modelo é dado por $y = Xb + Z\sum_i W_i m_i + e$.

O que torna a análise de QTL e da GWS diferenciada do BLUP tradicional é o fato da matriz Q ser desconhecida. No entanto, ela pode ser estimada com base nas informações dos marcadores. Segundo Perez-Enciso e Misztal (2004), a forma como os marcadores são usados para estimar W e a forma de definição de *q* resulta em distintos modelos que contemplam os vários delineamentos para a análise de QTLs e formas de seleção genômica.

Modelo G-BLUP

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 (ver final do tópico 6.26). No entanto, a alternativa mais eficiente é realizar toda a predição em um único passo, conforme relatado por Misztal et al. (2009) 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: $y = Xb + Zg + e$, em que *y* é o vetor de observações fenotípicas, *b* é o vetor de efeitos fixos, *g* é o vetor dos efeitos genéticos aditivos individuais (aleatórios) e *e* refere-se ao vetor de resíduos aleatórios. *X* e *Z* são as matrizes de incidência para *b* e *g*. Usando informações fenotípicas e dos marcadores tem-se o modelo equivalente: $y = Xb + ZWm + e$, em que *m* é o vetor dos efeitos aleatórios de marcadores, W é a matriz de incidência para *m* e g = Wm. A matriz de incidência W 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 W é usar os valores -1, 0 e 1 (Resende, 2007; 2008; Resende et al., 2010).

As equações de modelo misto para a predição de g via o método G-BLUP equivalem a: $\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+G^{-1}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$, em que $G = (WW')/k = (WW')/[2\sum_i^n p_i(1-p_i)]$ e $k = 2\sum_i^n p_i(1-p_i)$. Com padronização prévia dos elementos de W (dividindo-os por $[2\sum_i^n p_i(1-p_i)]^{1/2}$) e centrando a média em zero tem-se $G = WW'/n$, em que n é o número

de marcas. 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) sugeriram o seguinte parâmetro de escala:

$k = \left((p_0 - q_0)^2 + 2\left([\sum_{i}^{n} p_i(1-p_i)]/n \right)((\alpha+\beta+2)/(\alpha+\beta)) \right) n$ 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 marcadores SNP. O estimador de g pode ser resumido em: $[\hat{g}] = \left[Z'Z + G^{-1}\frac{\sigma_e^2}{\sigma_g^2} \right]^{-1} [Z\acute{y}]$.

A matriz G é densa e sua inversão apresenta alta demanda computacional. Assim, é interessante evitar essa inversão. Isto pode ser feito modificando (multiplicando por G) as equações de modelo misto para

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ GZ'R^{-1}X & GZ'R^{-1}Z + I(1/\sigma_g^2) \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ GZ'R^{-1}y \end{bmatrix}$$

ou, na sua forma simplificada em função de R, para

$\begin{bmatrix} X'X & X'Z \\ GZ'X & GZ'Z + I\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ GZ'y \end{bmatrix}$. Esse sistema de equações é então resolvido pelo método de Gauss-Seidel ou por iteração nos dados. Mas, em muitos casos, o número N de indivíduos genotipados é baixo e, como a matriz G tem dimensão N x N, a mesma pode ser invertida diretamente.

Para a avaliação global das três classes de indivíduos em um único passo, o mesmo modelo $y = Xb + Zg + e$ pode ser usado, porém com uma alteração (substituição da matriz G pela matriz H) nas equações de modelo misto, conforme Misztal et al.(2009): $\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$.

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:

$H^{-1} = A^{-1} + \begin{vmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{vmatrix} = \begin{vmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} + G^{-1} - A_{22}^{-1} \end{vmatrix}$, 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 $\hat{g}_j = \sum_i w_{ij}\hat{m}_i$. 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 ($\hat{g}$) pelo G-BLUP, os efeitos estimados dos marcadores ($\hat{m}$) podem ser obtidos, conforme desenvolvido a seguir: $\hat{g} = W\hat{m} \Rightarrow W'\hat{g} = W'W\hat{m} \Rightarrow \hat{m} = (W'W)^{-1}W'\hat{g}$..

A análise pelo G-BLUP 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 $G = (WW')/k = (WW')/[2\sum_{i}^{n} p_i(1-p_i)]$, essa h^2 pode ser estimada por REML fazendo uso das equações de modelo misto para a estimação dos componentes de variância σ_g^2 e σ_e^2. Segundo outra parametrização, os elementos da matriz G representam o parentesco realizado médio multi-locos e são dados por $G_{jk} = (1/n)\sum_{i=1}^{n} \frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$. Outro ponto favorável do G-BLUP 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 apenas para a população de estimação. No G-BLUP, a população de validação (indivíduos que foram apenas genotipados) tem seus fenótipos substituídos por dados perdidos e, portanto, os indivíduos dessa população tem uma estimativa validada da acurácia.

Um modelo G-BLUP incluindo efeitos de dominância (d) e epistáticos do tipo aditivo x aditivo (aa) pode ser ajustado e é dado por $y = Xb + Zg + Zd + Zaa + e$, em que a estrutura de variâncias é dada por $g \sim N(0, G\sigma_g^2);\ d \sim N(0, G_d\sigma_d^2);\ aa \sim N(0, G_{aa}\sigma_{aa}^2);\ e \sim N(0, I\sigma_e^2)$ e os efeitos epistáticos apresentam matriz de covariância $G_{aa} = G\#G$, em que # denota o produto de Hadamard. Os efeitos de dominância apresentam matriz de incidência S e de covariância G_d, com variância $\sigma_d^2 = \sum_{i=1}^{n}[2p_i(1-p_i)][1-(2p_i(1-p_i))]\sigma_{md}^2$, em que σ_{md}^2 é a variância de dominância contribuída por um loco m. A relação σ_{md}^2/σ_d^2 é então dada por $\sigma_{md}^2/\sigma_d^2 = 1/\sum_{i=1}^{n}[2p_i(1-p_i)][1-(2p_i(1-p_i))]$.

A matriz S é análoga à W e é composta por valores de o, 1 e o (para os genótipos marcadores MM, Mm e mm, respectivamente), seguindo, portanto, distribuição Bernoulli com média $2p_i(1-p_i)$ e variância $[2p_i(1-p_i)][1-(2p_i(1-p_i)]$. Subtraindo os elementos de S pela média $(2p_i(1-p_i))$, obtém-se os seguintes valores de s_{ij}, para o marcador i no indivíduo j: $s_{ij} = 0 - [2p_i(1-p_i)]$, $s_{ij} = 1 - [2p_i(1-p_i)]$ e $s_{ij} = 0 - [2p_i(1-p_i)]$, respectivamente, obtendo-se uma variável com média zero. Para os indivíduos com dados perdidos de marcas, seus valores na matriz S devem ser o valor esperado $2p_i(1-p_i)$, que, centrados, transformam-se em zero. Assim, valores perdidos devem ser substituídos por $s_{ij} = 0$.

Sendo $Var(Sm_d) = SS'\sigma^2_{md}$ e pelo modelo equivalente $G_d\sigma^2_d = Var(Sm_d) = SS'\sigma^2_{md}$, a matriz de parentesco de dominância é, então, dada por $G_d = SS'\sigma^2_{md}/\sigma^2_d$. Sendo $\sigma^2_{md}/\sigma^2_d = 1/\sum_{i=1}^{n}[2p_i(1-p_i)][1-(2p_i(1-p_i))]$, tem-se $G_d = SS'/\sum_{i=1}^{n}[2p_i(1-p_i)][1-(2p_i(1-p_i))]$.

Os elementos da matriz G_d representam o parentesco realizado médio de dominância multi-locos e são então dados por $G_{d_{jk}} = (1/n)\sum_{i=1}^{n}\frac{[s_{ij}-(2p_i(1-p_i)][s_{ik}-(2p_i(1-p_i)]}{[2p_i(1-p_i)][1-(2p_i(1-p_i))]}$.

Modelos em nível de indivíduos contemplando as interações genótipos ambientes (ge) podem também ser ajustados, desde que existam indivíduos aparentados no mesmo ambiente e também entre ambientes. Neste caso, o modelo equivale a $y = Xb + Zg + Zge + e$, em que *ge* 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 *ge*. As equações de modelo misto para a predição de *a* e *ge* via o método BLUP equivalem a:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z+G^{-1}\frac{\sigma^2_e}{\sigma^2_g} & Z'Z \\ Z'X & Z'Z & Z'Z+G^{-1}_{ge}\frac{\sigma^2_e}{\sigma^2_{ge}} \end{bmatrix}\begin{bmatrix}\hat{b}\\ \hat{g}\\ \hat{ge}\end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z\acute{y}\end{bmatrix}$$, em que:

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

O método G-BLUP ou BLUP genômico pode também ser implementado considerando a heterogeneidade de variância entre marcadores. Nesse caso, a matriz G é dada por $G = (W^*DW^{*\prime})/[2\sum_{i}^{n} p_i(1-p_i)]$, em que p_i é a frequência de um dos alelos do loco i e W* refere-se à matriz W corrigida para suas médias em cada loco ($2p_i$). A matriz D é dada por $diag(D) = (\tau^2_1 \ldots \tau^2_n)$ e os elementos τ^2_i 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; (iv) permite considerar a heterogeneidade de variância genética entre marcadores.

6.13 GBLUP e Regressão Aleatória Multivariada (MRR)

Para caracteres associados a curvas de crescimento em função do tempo ou da idade de avaliação, os modelos de regressão aleatória multivariados (MRR) devem ser adotados considerando dois conjuntos de regressão dos fenótipos do caráter em função das idades mensuradas. O primeiro conjunto diz respeito à regressão fixa para os indivíduos pertencentes à mesma classe de efeitos fixos e o segundo contempla efeitos aleatórios que descrevem os desvios de cada indivíduo em relação à regressão fixa. As regressões fixas e aleatórias são representadas por funções contínuas.

Um modelo de regressão aleatória multivariado pode ser ajustado para os efeitos aleatórios genético aditivo e ambiente permanente cujas covariáveis podem ser descritas por polinômios de Legendre. Esse modelo é dado por y = Xb + Zg + Tp + e, em que p é o vetor dos efeitos de ambiente permanente com matriz de incidência T. Expresso de outra forma, o modelo é dado por $y = Xb + \phi_g g + \phi_p p + e$, em que ϕ_g e ϕ_p são matrizes de incidência para os coeficientes polinomiais dos efeitos genético aditivo e de ambiente permanente, respectivamente.

As distribuições dos coeficientes de regressão aleatória são dadas por: $g \sim N(0, A \otimes K_g)$, sendo A a matriz de parentesco entre os indivíduos e K_g uma matriz de dimensão (k_g +1) x (k_g+1) de covariâncias entre coeficientes de regressão aleatória para os efeitos genéticos aditivos; $p \sim N(0, I_n \otimes K_p)$, sendo I_n uma matriz identidade de ordem n e K_p uma matriz de dimensão (k_p+1) x (k_p+1) de covariâncias entre coeficientes de regressão aleatória para os efeitos de ambiente permanente. Maiores detalhes são apresentados no Capítulo 1. Com seleção genômica os modelos de regressão aleatória multivariados devem usar, em lugar de A, a matriz de parentesco genômico, dada por $G = (WW')/k = (WW')/[2\sum_i^n p_i(1-p_i)]$.

6.14 Comparação entre Métodos de Estimação Penalizada

Métodos de estimação penalizada

Em um problema de regressão tem-se que a variável dependente y é dada como função de uma variável preditora (w) e vetor de erros aleatórios (e), segundo o modelo $y = \beta' w + e$. No contexto da seleção genômica define-se w 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. Aqui, a notação β substitui a notação *m* usada nos tópicos anteriores.

Usando esperança condicional, a equação de regressão é dada por:
$\hat{y} = \hat{\beta}' w = E(y \mid w)$

Isso implica $\hat{\beta} = E(\beta \mid w, y) = [\int \beta\, p(\beta) p(y \mid \beta, w)\, d\beta] / [\int p(\beta) p(y \mid \beta, w)\, d\beta]$, em que $p(\beta)$ é a função densidade de probabilidade de β e $p(y \mid \beta, w)$ 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 (medida da falta de ajuste do modelo) da regressão e o segundo termo é a penalização, a qual depende da magnitude dos coeficientes de regressão via $\sum_{i=1}^n \beta_i^2$. 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 $\sum_{i=1}^n \beta_i^2 \le t$. A solução para esse problema de otimização conduz a $\hat{\beta} = [W'W + \lambda_{RR}(t)I]^{-1} W' y$.

Outro método relacionado é o LASSO, que combina *shrinkage* (regularização) com seleção de variáveis e envolve o seguinte problema de otimização, via minimização de $(1/N)\sum_j^N (y_j - \sum_{i=1}^n w_{ij}\beta_i)^2 + \lambda_L \sum_{i=1}^n |\beta_i|$, 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 $\sum_{i=1}^n |\beta_i| \le t$. 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 em que λ_L tende a zero a solução converge para método de regressão fixa via quadrados mínimos (FR-LS), ou seja, para $\hat{\beta} = (W'W)^{-1}W'y$. Nesse caso, não há seleção de covariáveis e 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) associada 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 subconjuntos de covariáveis (como o Garrote de Breiman) produz modelos interpretáveis, porém, com muita variabilidade nos resultados, pois, trata-se 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 0) como a seleção de subconjuntos. 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:

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

Função de penalização, restrição ou regularização:

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

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, tem-se o método RR-BLUP e

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

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, h² é a herdabilidade do caráter, σ_g^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 (G) e utilizada no método BLUP fenotípico tradicional, tem-se o método denominado G-BLUP ou BLUP genômico, que é equivalente ao RR-BLUP em termos da predição dos efeitos aditivos g. Assim, tem-se para o G-BLUP:

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

$G = (W^*W^{*\prime})/[2\sum_{i}^{n} p_i(1-p_i)]$, em que p_i é a freqüência de um dos alelos do loco i e W* refere-se à matriz W corrigida para suas médias em cada loco ($2p_i$).

Tem-se então a equivalência $\hat{g} = W\hat{\beta} = W[W'W + \lambda_{RR}(t)I]^{-1}W'y = [Z'Z + G^{-1}(\sigma_e^2/\sigma_g^2)]^{-1}y$.

b. LASSO

Função objetivo a ser minimizada:

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

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:

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

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 $\alpha = 0$, EN = LASSO ou se q = 1, EN = LASSO.
- Se $\alpha = 1$, EN = RR ou se q = 2, EN = RR.
- Se $1 \le q \le 2$ 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.
- Comporta semelhantemente ao Lasso, mas é robusta a extrema colinearidade entre as covariáveis.
- Permite seleciona 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 e Hastie, 2005).

Os métodos frequentistas Lasso e EN não são usados frequentemente devido ao surgimento dos Lassos Bayesianos, os quais apresentam uma série de vantagens e contornam os problemas associados aos referidos métodos frequentistas.

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

Solução para os coeficientes de regressão:

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

- 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 e usados para cômputo do método RR-BLUP-Het.

Distribuições normal (RR-BLUP) e exponencial (LASSO)

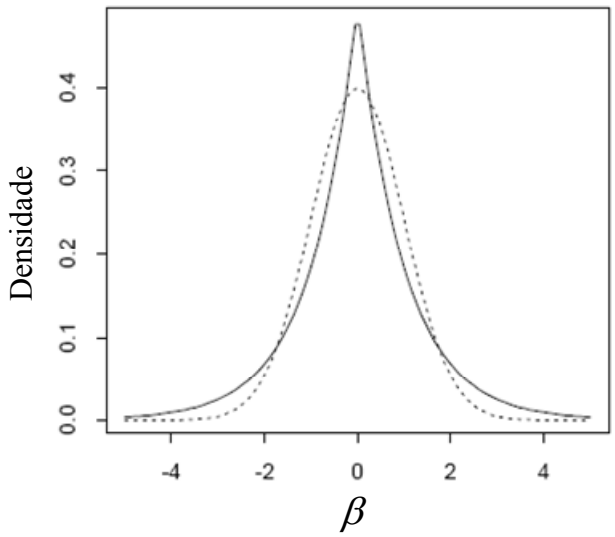


Figura 3. Densidades das distribuições normal (curva pontilhada) e exponencial dupla (curva cheia), ambas com médias iguais a zero e variâncias iguais à unidade.

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 zero e menor encurtamento sobre coeficientes de regressão distantes de zero (Figura 3).

6.15 Métodos Bayesianos

Os métodos de predição de valores genéticos genômicos RR-BLUP, Bayes A e Bayes B foram considerados por Meuwissen et al.(2001). Essas abordagens diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos; o método BayesA assume poucos genes de grandes efeitos e muitos genes com pequenos efeitos. No método Bayes B muitos efeitos de marcadores são assumidos como zero, *a priori*. Isso reduz o tamanho do genoma por meio da concentração nas partes do mesmo onde existem QTLs. O melhor método é aquele que reflete melhor a natureza biológica do caráter poligênico em questão, em termos de efeitos gênicos.

O método ideal de predição de valores genéticos genômicos equivale ao cálculo da média condicional do valor genético dado o genótipo do indivíduo em cada QTL. Essa média somente pode ser calculada usando uma distribuição *a priori* dos efeitos dos QTLs. Considerando, por meio de marcadores, cada QTL em separado, essa esperança condicional é dada por $\hat{\beta} = E(\beta|w)$. O estimador apropriado segue o teorema de Bayes e é dado por $\hat{\beta} = \frac{\int \beta \ f(w|\beta) \ f(\beta) \ d \ \beta}{\int \ f \ (w|\beta) \ f(\beta) \ d \ \beta}$, em que $f \ (w|\beta)$ é a função de verossimilhança dos dados (w), e $f(\beta)$ é a distribuição *a priori* dos efeitos β dos QTLs marcados. Esse estimador mostra que o método ideal depende da distribuição *a priori* dos efeitos de QTL. A presença de QTLs é testada em muitas posições (10 mil SNPs) e, portanto, não existe QTLs em muitas posições. Dessa forma, a distribuição *a priori* $f(\beta)$ deve ter uma alta probabilidade para $f(0)$. Para especificar essa alta probabilidade, deve-se ter uma noção de quantos QTLs controlam o caráter (Goddard & Hayes, 2007).

Nessa situação, com muitos efeitos β iguais a zero, o método RR-BLUP resulta em muitas estimativas de β próximas de zero, porém não iguais a zero. Na soma dessas estimativas, esse efeito acumulado introduz algum erro na predição. Os métodos bayesianos Bayes A e Bayes B relatados por Meuwissen et al. (2001) consideram mais adequadamente a distribuição *a priori* dos efeitos dos QTLs.

O método Bayes A é similar ao método BLUP com variâncias heterogêneas, pois as variâncias dos segmentos cromossômicos diferem para cada segmento e são estimadas sob esse modelo, considerando a informação combinada dos dados e da distribuição *a priori* para essas variâncias. Essa distribuição é tomada como uma qui-quadrado invertida e escalada. Os métodos Bayesianos propiciam acurácias mais altas porque forçam muitos efeitos de segmentos cromossômicos a valores próximos a zero (Bayes A) ou a zero (Bayes B, conduzindo a N/n mais favorável) 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.

A estimação Bayesiana maximiza a distribuição *a posteriori* do parâmetro ou distribuição condicional do parâmetro dado as observações (y) e é proporcional ao produto da função de verossimilhança pela distribuição *a priori* do parâmetro. Em outras palavras, a função de verossimilhança conecta a distribuição *a priori* à posteriori usando para isto os dados experimentais (amostrais). Dessa forma, a distribuição *a posteriori* contempla o grau de conhecimento prévio sobre o parâmetro

e também as informações adicionais propiciadas pelo experimento e é a base da estimação Bayesiana.

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 da distribuição conjunta *a posteriori* para os efeitos e componentes de variância; (iv) obtenção das distribuições condicionais completas *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, portanto métodos MCMC, como o amostrador de Gibbs, têm sido utilizados para obter amostras das distribuições marginais a posteriori por meio das distribuições condicionais completas *a posteriori* já citadas.

Nos métodos MCMC, as cadeias antes do equilíbrio fornecem amostras das distribuições condicionais completas *a posteriori* $f(\theta_i|y,\theta_2,\theta_3...)$ para os efeitos e componentes de variância. Após o equilíbrio fornecem amostras das distribuições marginais *a posteriori* $f(\theta_i|y)$ para as referidas variáveis aleatórias. Meuwissen et al. (2001) usaram 10.000 ciclos MCMC com descarte dos 1.000 primeiros como período de *burn in* (para se atingir o equilíbrio).

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 fenotípicos (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-quadrada inversa e escalonada conforme a seguir:

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

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

em que v_β é o número de graus de liberdades e S^2_β é o parâmetro de escala da distribuição.

Assim, tem-se que a distribuição marginal *a priori* dos efeitos genéticos dos marcadores, $\beta_i | v_\beta, S^2_\beta$, tem distribuição t de Student univariada, ou seja, $\beta_i | v_\beta, S^2_\beta \sim t(0, v_\beta, S^2_\beta)$. Assim, esta formulação resulta na modelagem dos efeitos dos marcadores como amostras de uma distribuição t de Student.

O valor de S^2_β pode ser derivado com base no valor esperado de uma variável aleatória com distribuição qui-quadrado invertida escalonada. Essa esperança

matemática, para um componente de variância genérico σ^2, 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 $E(\sigma^2_{\beta i}) = \frac{S^2_\beta v_\beta}{v_\beta - 2}$ e $S^2_\beta = \frac{E(\sigma^2_{\beta i})(v_\beta - 2)}{v_\beta}$. A esperança $E(\sigma^2_{\beta i})$ equivale $E(\sigma^2_{\beta i}) = \frac{\sigma^2_g}{\sum_{i=1}^{n} 2p_i(1-p_i)}$. Assim, $S^2_\beta = \frac{\sigma^2_g}{\sum_{i=1}^{n} 2p_i(1-p_i)} \frac{(v_\beta - 2)}{v_\beta}$, em que v_β = 4.012 ou 4.2, conforme Meuwissen et al. (2001), σ^2_g é a variância genética aditiva do caráter e p_i é a freqüência alélica do marcador i. Meuwissen et al. (2001) consideraram $S^2_\beta = 0.002 \; ou \; 0.0429$. Isto descreve uma distribuição moderadamente leptocúrtica. Qualquer valor maior que 4 pode ser usado para v_β. Valores menores ou iguais a 4 torna *a priori* "flat" (não informativa).

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

Assumindo $\beta_i \sim N(0, \sigma^2_{\beta_i})$, em que $\sigma^2_{\beta_i}$ é tomado de uma distribuição qui-quadrado invertida, segundo o enfoque bayesiano, 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 uma cauda mais longa que a distribuição normal. Este método pode ser implementado via amostragem de Gibbs, para obtenção dessa informação combinada (priori x verossimilhança) 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 distribuição *a priori* t escalada 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 = o, 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:
$y = Ju + W\beta + e$, onde:
y : vetor de dados fenotípicos.
u : média geral.
β : vetor de efeitos genéticos aditivos (aleatórios) de marcadores.
e : vetor de erros.
J, W : matrizes de incidência que associam u e β aos dados fenotípicos (y).

Considera-se, inicialmente, que a distribuição condicional dos dados fenotípicos, dados u, β e σ_e^2 é normal multivariada: $y|\mu,\ \beta,\ \sigma_e^2 \sim N\ (1\mu + W\beta,\ I\ \sigma_e^2)$, 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 acima. 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 escalada invertida, tem-se que a distribuição conjunta *a posteriori* pode ser reescrita:

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

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 $\sigma_{\beta i}^2$ é dada por uma qui-quadrado invertida escalonada por $S_\beta^2 + \beta_i' \beta_i$ e com graus de liberdade υ_β , ou seja $P(\sigma_{\beta i}^2 | \beta_i) = \chi^{-2}(\upsilon_\beta, S_\beta^2 + \beta_i' \beta_i)$. Não se pode usar essa distribuição *a posteriori* diretamente para estimar $\sigma_{\beta i}^2$, 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 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 u como único efeito fixo, pode-se calcular u 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 u, calcular os valores de *e* via $e = (y - J\mu - W\beta)$, em que W = [W_1 W_2 W_3] é a matriz de incidência para os efeitos de marcadores. Então, atualize a variância residual por meio da amostragem de $\chi^{-2}(N-2, e_i' e_i)$.
4. Amostrar, de uma distribuição normal com média (1/N) $(y - Wg)$ e variância σ^2_e / N, a média geral dado a atualizada variância residual.
5. Amostrar, de uma distribuição com média $\frac{W_{ij}' y - W_{ij}' W \beta_{ij=0} - W_{ij}' \mathrm{J}_n u}{W_{ij}' W_{ij} + \sigma^2_e / \sigma^2_{\beta i}}$ e variância $\sigma^2_e / (W_{ij}' W_{ij} + \sigma^2_e / \sigma^2_{\beta i})$, todos os efeitos de marcadores β_{ij} dado a amostragem mais recente da média, σ^2_e e $\sigma^2_{\beta i}$, em que W_{ij} é o vetor coluna de W 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.

BayesB

O método BayesB apresenta as mesmas suposições que o BayesA para uma fração π dos SNPs e assume que $(1-\pi)$ dos SNPs apresenta efeitos nulos. Um problema desse método é a escolha da fração π. Com a seleção de covariáveis baseada nos maiores módulos 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 $(1-\pi)$, 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 $\sigma^2_{\beta i} = 0$. No método BayesB, uma vez que não é possível uma amostragem de $\sigma^2_{\beta i} = 0$, o método da amostragem de Gibbs não pode ser usado, pois não move sobre todo o espaço de amostragem. Assim, o algoritmo de Metropolis-Hastings (HM) 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 dessa 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) faz 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 o autor 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 $\sigma^2_{gi} = 0$, não conduz necessariamente a $g_i = 0$, conforme intenção original de Meuwissen et al. (2001), em que g_i é o efeito genético do loco i. Sugere então que o estado zero seja especificado ao nível dos efeitos e não ao nível das variâncias. Assim, à probabilidade de mistura π poderia ser atribuída uma distribuição *a priori* Beta. Surge então, o método BayesCπ 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 zero os métodos BayesCπ e RR-BLUP são iguais.

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, a seleção de um grupo de marcas que está em associação com o caráter é necessária. O método BayesB determina π subjetivamente. 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 parsimonioso 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 no caráter.

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 27.

Tabela 27. Comparação entre os métodos bayesianos

Método	Modelo para os efeitos genéticos	Parâmetros que estima	Método se π = 1
BayesDπ	$a_j = \sum_{i=1}^{n} \beta_i w_{ij} \delta_i$	$\sigma^2_{\beta i}$, δ_i , σ^2_e , π	BayesD
BayesCπ	$a_j = \sum_{i=1}^{n} \beta_i w_{ij} \delta_i$	σ^2_{β} , δ_i , σ^2_e , π	BayesC
BayesC	$a_j = \sum_{i=1}^{n} \beta_i w_{ij} \delta_i$	σ^2_{β} , δ_i , σ^2_e	RR-BLUP bayesiano (δ_i = 1)
BayesB	$a_j = \sum_{i=1}^{n} \beta_i w_{ij} \delta_i$	$\sigma^2_{\beta i}$, δ_i , σ^2_e	BayesA
BayesA	$a_j = \sum_{i=1}^{n} \beta_i w_{ij}$	$\sigma^2_{\beta i}$, σ^2_e	-
RR-BLUP	$a_j = \sum_{i=1}^{n} \beta_i w_{ij}$	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 $\hat{y}_{-i} = y - \sum_{j \neq i}^{n} w_j \hat{\beta}_j$. Estimar a estatística suficiente $\hat{Y}_i = (w_i' y - \sum_{j \neq i}^{n} (w_i' w_j)\hat{\beta}_j)/N$ e $\sigma^2 = \sigma_e^2 / N$.

b) Calcular $\hat{\beta}_i = E[\beta_i \mid Y_i]$, que é usado para atualizar a solução para o marcador i. A expressão para cômputo de $\hat{\beta}_i = E[\beta_i \mid Y_i]$ 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.

6.16 Métodos Lasso

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 28.

Tabela 28. Características dos três métodos Lassos.

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

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 W 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 $\sigma^2_g = \sum_{i=1}^{m} 2p_i(1-p_i)\sigma^2_\beta$ (Gianola et al., 2009), tem-se $\sigma^2_g = \sum_{i=1}^{m} 2p_i(1-p_i)\,2/\lambda^2$, em que σ^2_g 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 $\lambda^2 = \sum_{i=1}^{m} 2p_i(1-p_i)\,2/\sigma^2_g$. 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

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

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

Com dois componentes de variância (σ^2_e e σ^2_β) o modelo torna-se

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

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^2_i 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 1000000.

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

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

$\beta_i \mid demais \propto N(w_i'(y - 1_i\tilde{\mu} - W\tilde{\beta}_{-i})\tilde{\sigma}_e^{-2} / LHS_I, 1 / LHS_I)$, em que $LHS_i = w_i' w_i \tilde{\sigma}_e^{-2} + \tau_i^{-2}$ e w_i é a linha de W 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

$y = 1u + W\beta + e$

$e \mid \sigma^2 \sim MVN(0, I\sigma^2)$

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

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

$p(\beta \mid \tau) \sim N(0, D\sigma^2); diag(D) = \tau_1^2 ... \tau_n^2);$

$p(\tau \mid \lambda) = \prod_i (\lambda^2 / 2) \exp(-\lambda^2 \tau_1^2 / 2).$

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:

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

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

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

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

G-BLUP com heterogeneidade de variâncias

O método G-BLUP ou BLUP genômico pode também ser implementado considerando a heterogeneidade de variância entre marcadores. Nesse caso, a matriz A é dada por $A = (W^* D W^{*\prime}) / [2\sum_i^n p_i(1 - p_i)]$, em que p_i é a frequência de um dos alelos do loco i e W* refere-se à matriz W 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 via os 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.

Relação entre RR-BLUP, BLASSO e IBLASSO

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.

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 π = 1, o BayesCπ é igual ao RR-BLUP.

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., 2011), 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. Assim, 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 e não a variância genética total do caráter. Assim 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

Com arquitetura genética homogênea, conforme Resende et al. (2011), 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:

$$h^2 = \frac{1}{1+\lambda_{BL}^2/(2n_Q)} = \frac{1}{1+2\lambda_{RR}/(2n_Q)} = \frac{1}{1+\lambda_{RR}/n_Q} = \frac{n_Q}{n_Q+\lambda_{RR}}, \quad \text{em que}$$

$$n_Q = 2\sum_i^n p_i(1-p_i).$$

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 pode ser especificado em função do parâmetro de penalização λ_{BL} do BLASSO, sendo dado por

$$\hat{h}^2 = \frac{1}{1+\hat{\lambda}_{BL}^2/(2n_Q)} = \frac{2n_Q}{2n_Q+\hat{\lambda}_{BL}^2}.$$

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

Relação entre RR-BLUP, BLASSO e IBLASSO

Para o IBLASSO, conforme Resende et al. (2011), tem-se:

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

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

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

$n_Q = 2\sum_i^n p_i(1-p_i)$. Assim, com arquitetura genética homogênea, a h² 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 h² 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.

6.17 Distribuições dos efeitos genéticos nos métodos RR-BLUP, Bayes e Lasso.

Na Tabela 29 são apresentadas as distribuições assumidas para os efeitos genéticos de marcadores nos diferentes métodos de GWS.

Tabela 29. 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 das variâncias
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 π = 0)	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-π) 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	qui-quadrado invertida
Lassos	Exponencial Dupla	Exponencial Dupla	Gama Invertida

As distribuições assumidas para os efeitos genéticos de marcadores nos diferentes métodos de GWS são: RR-BLUP: Normal com variância comum; Métodos Bayesianos: t dado priori qui-quadrado para as variâncias; Lassos: Exponencial Dupla. A Figura 4 ilustra as formas das distribuições normal (RR-BLUP), t (BayesA) e exponencial (LASSO).

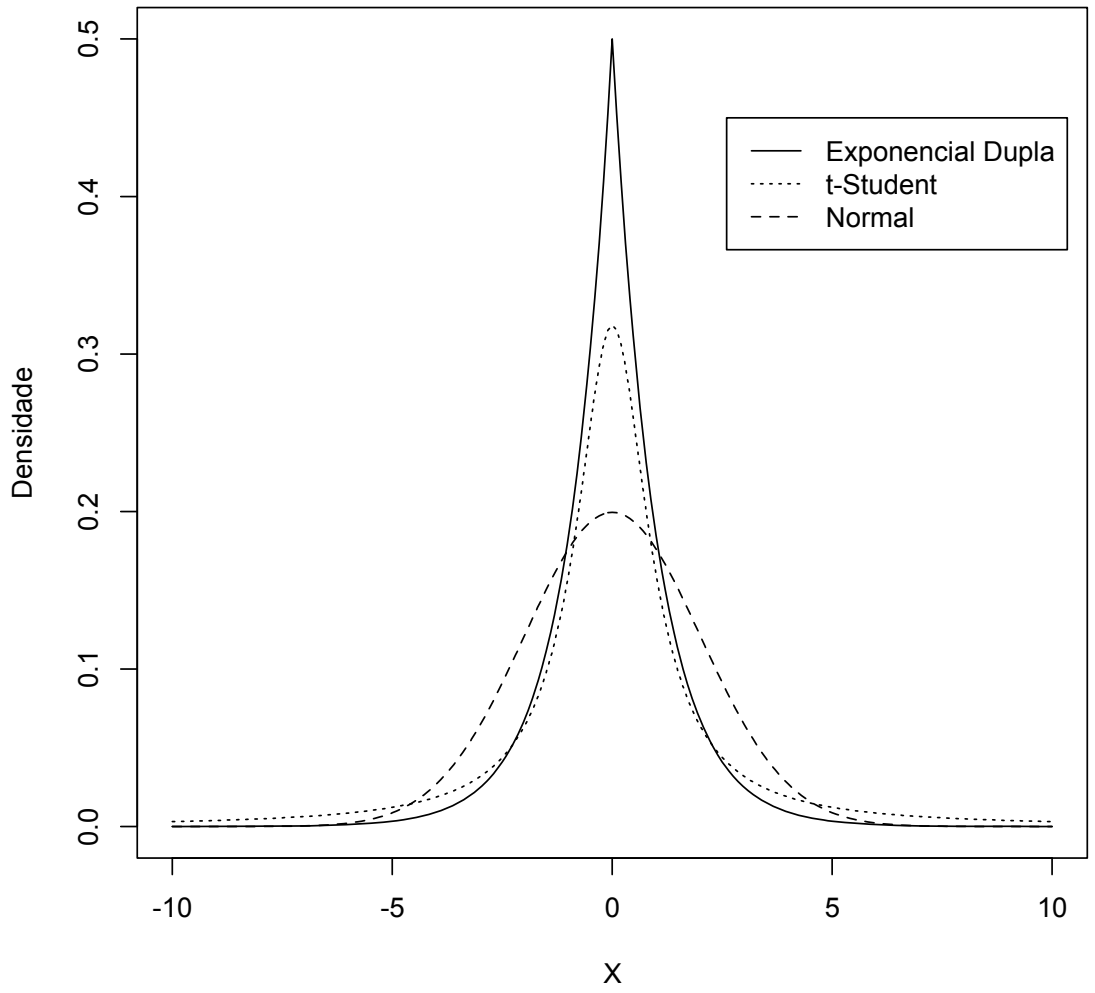


Figura 4 – Funções densidade de probabilidade das distribuições exponencial dupla, normal e t de Student, todas com médias iguais a zero e variâncias iguais à unidade (Resende Jr. et al., 2012c).

Observa-se que, em relação ao RR-BLUP, 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 zero e menor encurtamento sobre coeficientes de regressão distantes de zero. A densidade *a priori* utilizada no BayesA também apresenta maior massa de densidade no valor zero e caudas mais robustas do que a normal usada no RR-BLUP. O LASSO Bayesiano também exerce maior encurtamento sobre coeficientes de regressão próximos de zero do que o BayesA. Mas as caudas das distribuições são similares pelos dois métodos (Figura 4).

6.18 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 juste de modelos com muitas interações epistáticas e de dominância.

Gonzales-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 e 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 $y_j = u + g(w_j) + e_j$, 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 $g(w_j)$ é uma função desconhecida que relaciona os genótipos marcadores (covariáveis) com os fenótipos (variável dependente).

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

Função Objetivo a ser Minimizada:

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

Função de Penalização

$h\left\|g(w)\right\|_H^2$, em que h é o parâmetro de suavização e $\left\|g(w)\right\|_H^2$ é a norma de g(w) 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(w) que minimize a soma de quadrados penalizada $SS[g(w)] = \left\{ \sum_{j}^{N} [(y_j - u - g(w_i)]^2 + h \left\| g(w) \right\|_H^2 \right\}$. A solução para essa minimização é dada por:

$g(w) = \alpha_0 + \sum_{j=1}^{N} \alpha_j k(w - w_i)$, em que α_j são coeficientes desconhecidos (com total equivalente ao número N de indivíduos genotipados) e $k(w\text{-}w_j)$ 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, $k(w_i,w_j)$. 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 $(Cor[g(w_i),g(w_j)]$. 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 $Cov(g_i,g_j) \ \alpha \ K_{RKHS}(w_i,w_j)$, em que w_i,w_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 e b).

O modelo pode então ser expandido da seguinte forma:

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

$y_j = u + \sum_{j=1}^{N} \alpha_j k(w - w_i) + e_j$, em que α_0 faz parte de u.

Em termos vetoriais, tem-se:

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

$$T(h) = \begin{bmatrix} t_1(h) \\ t_2(h) \\ . \\ . \\ . \\ t_n(h) \end{bmatrix}, \; t_i(h) = {}_1[k_h(w_i - w_1)\; k_h(w_i - w_2) ... k_h(w_i - w_n)]_n \text{ e } \alpha' = {}_1[\alpha_1\; \alpha_2\; ...\alpha_n]_n .$$

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: $y = 1u + K_h\alpha + e$, 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 $K_t^* = [K_h^*(w_i - w_j)]$, com $K_h^*(w_i - w_j)$ sendo o kernel entre o genotipo 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 seguinte modelo $g_j = p_j + \alpha_j$, em que p_j é a regressão sobre o pedigree, 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, , $K_{RKHS}(w_i,w_j)$, avaliada nos genótipos marcadores, em que w_i e w_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:

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

Qualquer função positiva definida satisfazendo $\sum_i \sum_j \alpha_i \alpha_j K_{RKHS}(w_i, w_j)$ para todas as sequências não nulas $\{\alpha_i\}$ é uma escolha válida de kernel.

Pode-se escolher $K_{RKHS}(w_i, w_j)$ como um kernel Gaussiano $K_{RKHS}(w_i, w_j) = \exp\{-2(d_{ij} / q_{0.5})\}$, em que $d_{ij} = \sum_{k=1}^{p} (w_{ik} - w_{jk})^2$ 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 etal., 2010):

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

Amostras são retiradas dessa distribuição.

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

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

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

respectivamente.

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., 2009).

6.19 Regressão via quadrados mínimos parciais (PLSR)

A regressão via quadrados mínimos parciais (PLS) 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., 1985). 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.

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

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

e_j : termo residual.

Pelo PLS, a função $f(w_j)$ é definida como $f(w_j) = \sum_{l=1}^{h} t_{jl}\beta_l$, 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 $\beta_i = \sum_{l=1}^{h} \beta_l w_{li}$.

As variáveis latentes são componentes ortogonais (isso 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 W das variáveis originais por meio de T = WQ, em que Q é uma matriz de pesos. A diferença é que a PCR extrai componentes que explicam a variância de W e a PLSR extrai componentes que tem maior covariância com y. Na PLSR as colunas de pesos na matriz Q 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.

O método PLS é definido de acordo com as seguintes decomposições das matrizes W e Y, as quais são efetuadas de forma simultânea:

$W = TL' + E_1$,

$Y=Uq'+e_2$ (1),

, em que T e U são matrizes de componentes, L e q são matrizes de carregamento, E_1 e e_2 são vetores de resíduos. A decomposição não é independente, o que possibilita estabelecer uma relação entre componentes de W e Y de forma que para cada fator a relação abaixo é obtida:

$\hat{u}_\ell = \hat{b}_\ell t_\ell$

, sendo $\hat{b}_\ell = (u'_\ell t_\ell)/(t'_\ell t_\ell)$ $(\ell = 1,...,n_{pls})$ coeficientes estimados via quadrados mínimos ordinários (*Ordinary Least Squares* – OLS) e agrupados em uma matriz diagonal B. Maiores detalhes são apresentados por Azevedo (2012).

6.20 Regressão via componentes principais (PCR)

Conforme Azevedo et al. (2012), o método PCR é definido de acordo com a seguinte combinação de variáveis:

$Z_v = W\hat{P}$ (2),

, sendo P a matriz de autovetores da matriz de covariância entre as covariáveis (W) e, Z_v a matriz dos componentes principais (Z_ν ,ν=1,...,n_{pcr}), os quais representam combinações lineares das covariáveis originais.

Visando estabelecer a relação entre Y e os componentes utiliza-se a regressão linear múltipla para obter as equações de predição do PCR e do PLS, respectivamente:

$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1\hat{z}_1 + \hat{\alpha}_2\hat{z}_2 + ... + \hat{\alpha}_{n_{pcr}}\hat{z}_{n_{pcr}}$, (3)

$\hat{y}=\hat{T}\hat{B}\hat{q}'$, (4)

em que: α_ν é o coeficientes da regressão entre Y e Z, $\forall \nu$ obtidos por meio do método OLS.

Os coeficientes Bq' e $\hat{\alpha}$ não possuem interpretação biológica, porém é possível estimar os coeficientes associados às variáveis originais (efeitos dos marcadores) combinando as equações (2) e (3), (1) e (4) dos métodos PCR e PLS, respectivamente. Desta forma tem-se:

$\hat{m}_{pcr}=\hat{P}\hat{\alpha}$,

$\hat{m}_{pls}=\hat{L}\hat{B}\hat{q}'$.

Os métodos PLS e PCR podem também serem aplicados com seleção de covariáveis, gerando os métodos PLS esparso e PCR supervisionado (Long et al. 2011).

6.21 Regressão via componentes independentes (ICR)

A Regressão via Componentes Independentes (*Independent Component Regression*– ICR), proposto por Comon (1994), consiste em decompor a matriz de covariáveis W em combinações de componentes independentes, garantindo a retirada da multicolinearidade dos dados, além de reduzir a dimensionalidade. Por esse método não existe o pressuposto de que os dados sejam provenientes de uma distribuição normal. Desta forma, pode ser aplicado de forma eficiente à seleção genômica ampla (GWS), em que a matriz de marcas W é parametrizada com os valores 0, 1 e 2 (distribuição não normal). Dessa forma, conforme Azevedo et al. (2012), tem-se a decomposição dada por:

$W' = F'S$,

em que: F é definida como uma função $f(KR)$, sendo K uma matriz de ortogonalização de S obtida por meio da decomposição espectral e R uma matriz ortogonal que maximiza a independência estatística das colunas de S, em que S é a matriz dos componentes independentes S_κ $\kappa=1,...,n_{icr}$.

O algoritmo desenvolvido por Hyvärinen (1998b) é utilizado na ICR visando encontrar a matriz R baseando-se no princípio da máxima entropia ($J(r)$). Desta forma, obtém-se a seguinte aproximação:

$J(r) \propto [E\{G_i(r)\} - E\{G_i(v)\}]^2$,

sendo r e v variáveis padronizadas e $G_1(u) = -\exp\left(-u^2/2\right)$ em que u é uma variável normal padronizada. Após o processo iterativo tem-se a matriz de componentes dada por:

$\hat{S} = WKR$, (5)

sendo KR uma aproximação de F'. Assim, obtêm-se a equação de predição baseada no método ICR expressa por:

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1\hat{s}_1 + \hat{\gamma}_2\hat{s}_2 + ... + \hat{\gamma}_{n_{icr}}\hat{s}_{n_{icr}}, \quad (6)$$

em que: s_κ é o componente independente e γ_κ é o coeficiente da regressão determinado por meio do método OLS, $\kappa=1,...,n_{icr}$. Similarmente aos outros métodos, pode-se obter os efeitos de marcadores à partir das equações (5) e (6) por meio da seguinte equação:

$\hat{m}_{icr} = KR\hat{\gamma}$,

, sendo $\hat{\gamma}$ o vetor de estimativas dos coeficientes provenientes da regressão entre Y e S.

Um passo importante dos métodos de Redução Dimensional é a escolha do número ótimo de componentes a serem inseridos no modelo. Um critério de decisão para o PLS e PCR é adotar uma percentagem da variação total explicada pelos componentes, a qual neste trabalho foi de 70%. Tal porcentagem também foi considerada para o método ICR, uma vez que Cadavid et al. (2008) sugere que o número de componentes no método ICR pode ser o mesmo obtido no método PCR.

6.22 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 30).

Tabela 30. Parâmetros usados na simulação.

Caráter	Va	Ve	h^2	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 31).

Tabela 31. 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 32 (Resende et al., 2011).

Tabela 32. 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.

6.23 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 $\hat{m} = (W'W + \lambda_{RR} I)^{-1} W' y$. O estimador do efeito de uma marca i equivale a $\hat{m}_i = (w_i' w_i + \lambda_{RR})^{-1} w_i' w_i yd_i = f_i yd_i$, em que $f_i = (w_i' w_i + \lambda_{RR})^{-1} w_i' w_i$ e 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 $yd_i = w_i'(y - \mu - \sum_j w_j \hat{m}_j)$, i ≠ j. O valor genético aditivo do indivíduo k é dado por $\hat{g}_k = \sum_i w_i f_i yd_i$.

Pelos métodos bayesianos BayesA e BayesB existe um componente adicional resultante da amostragem da distribuição condicional *a posteriori* de β tal que $\hat{g}_k = \sum_i w_i f_i yd_i + N(\hat{m}_i, (w_i' w_i + \lambda_i)^{-1} \sigma_e^2)$. 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. Também, isto pode ser corrigido via parametrização com padronização em W.

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.

6.24 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 rescrita 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 modelo associado ao sistema de equações equivale a $c_Y = \mu + Mc_Y^* + e$, em que $c_Y^* = [c_Y \quad c_X]$.

O fator α é necessário para que o sistema tenha solução e é dado por $\alpha = \sigma_e^2 / \sigma_c^2$, em 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, $\alpha = \sigma_e^2 / \sigma_c^2 = 0.001 / 0.999 = 0.001001$.

As equações de modelo misto apresentadas são praticamente iguais às equações de quadrados mínimos. Para derivação do BLUP não é necessário a suposição de normalidade (o conteúdo c não tem distribuição normal) segundo alguns procedimentos como a minimização da variância do erro de predição; necessita-se apenas de componentes de variância conhecidos. Outras derivações como aquela via máximo a posteriori (MAP) assumem que y e g tem distribuição normal multivariada. Nesse caso, propriedades favoráveis adicionais são asseguradas ao BLUP (ver Resende, 2002, páginas 220 a 226).

Considere o seguinte exemplo, com quatro indivíduos genotipados (não aparentados e com contagem de alelos marcadores 1, 0, 2 e 2, respectivamente) e um 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:

$$\begin{aligned} c_X &= \begin{pmatrix} 1 & A_{XY} A_{YY}^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix} \\ &= \begin{pmatrix} 1 & A_{XY} I_{(4)} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix} = \begin{pmatrix} 1 & A_{XY} \end{pmatrix} \begin{pmatrix} \mu \\ c_Y - 1\mu \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0.5 \end{pmatrix} \begin{pmatrix} 1.25 \\ -0.25 \\ -1.25 \\ 0.75 \\ 0.75 \end{pmatrix} = 1.625 \end{aligned}$$

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

6.25 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, café) 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. A GWS explorando essas vantagens foi implementada por Resende Jr. (2010), Resende et al. (2012) e Resende Jr. et al. (2012 a e b) em espécies florestais, por Cavalcanti et al. (2012) em cajueiros, por Oliveira et al. (2012) em mandioca, por Simeão et al. (2013) em forrageiras e por Rocha (2011) e Azevedo (2012) em suínos.

Espécies vegetais alógamas anuais (milho, girassol)

Nessas espécies o benefício da GWS se dá devido a três fatores: aumento de $r_{gg\hat{}}$, 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 $r_{gg\hat{}}$ 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. A GWS explorando essas vantagens foi implementada por Fritsche Neto (2011), Fritsche Neto et al. (2012) e Oliveira et al. (2012).

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 $r_{gg\hat{}}$, 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 $r_{gg\hat{}}$ 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.

O aumento da eficiência do melhoramento de plantas anuais via aplicação da metodologia de modelos mistos é apresentada com detalhes na coleção de artigos publicados por Viana et al. (2010; 2011; 2012).

6.26 Redução no erro da inferência sobre os QTL via uso dos marcadores

(A) Método G-BLUP

O método G-BLUP foi inicialmente aplicado por Nejati-Javaremi et al. (1997) e Fernando (1998) e, no contexto da seleção genômica por Habier et al. (2007), Van Raden (2008), Goddard (2008), Goddard et al. (2009), Hayes et al. (2009) e Stranden & Garrick (2009). Assim, no contexto da GWS, o método G-BLUP emergiu vários anos após a proposição dos métodos RR-BLUP (também denominado SNP-BLUP), BayesA e BayesB por Meuwissen et al. (2001).

Modelo G-BLUP

$y = Xb + Zg + e$; $Var(g) = G_M \sigma_g^2$,

em que G_m é a matriz de parentesco genômico nos locos marcadores.

Modelo Equivalente G-BLUP

$g = Wm$

$y = Xb + ZWm + e$; $Var(Wm) = WI\sigma_m^2 W' = WW'\sigma_m^2$

, em que *m* é o vetor de efeitos genéticos (substituição alélica) dos marcadores.

Assim, $Var(g) = Var(Wm)$ e, portanto, $G_M \sigma_g^2 = WW'\sigma_m^2$ e $G_M = WW'\sigma_m^2 / \sigma_g^2$ e W é a respectiva matriz de incidência.

A vantagem da GWS advém da possibilidade de se acessar os genótipos dos próprios QTLs que controlam o caráter em questão e então estimar os seus efeitos nos fenótipos. De forma equivalente, uma vez lidos os genótipos dos QTLs nos vários indivíduos pode-se construir a matriz de parentesco exato (G_Q) entre os indivíduos em avaliação e produzir estimativas acuradas de seus valores genéticos genômicos. Nessa predição BLUP usando G_Q realiza-se intrinsecamente a associação QTL e seus efeitos nos fenótipos.

No entanto, na prática, tem-se a matriz de parentesco G_M baseada em marcadores e não tem-se G_Q. Assim, há uma distância ou erro (G_E) na inferência sobre G_Q baseada em G_M, ou seja, $G_Q - G_M = G_E$ e, portanto, $G_Q = G_M + G_E = G_M + (G_Q - G_M)$. Assim, $G_M = G_Q - G_E$, ou seja, G_M estima a diferença $G_Q - G_E$.

O valor esperado de G_M quando o número de marcadores tende a infinito é a matriz A obtida com base em pedigree. Assim a equação para G_Q pode ser rescrita como $G_Q = G_M + (G_Q - G_M) = E(G_M) + [G_Q - E(G_M)] = A + (G_Q - A)$. Nessa mesma condição (número muito grande de marcadores) e se os marcadores coincidem perfeitamente com os QTLs tem-se $G_Q = A$ e as seleções genômica e fenotípica se equivalem. Uma vez que o número de QTLs de um caráter é finito, tem-se $G_Q \neq A$ e a seleção genômica pode superar a fenotípica. Nesse caso, os desvios em $(G_Q - A)$ contemplam a segregação mendeliana dos alelos nos QTLs.

Sendo $G_Q = A + (G_Q - A)$ e como não se conhece G_Q, essa pode ser estimada por $\hat{G}_Q = A + \hat{\beta}(G_M - A)$, em que $\hat{\beta}$ é uma regressão matricial dos elementos de $(G_Q - A)$ nos elementos de $(G_M - A)$ e visa retirar de G_M quanto é devido a G_Q, separando-a de G_E. O coeficiente de regressão é dado por $\hat{\beta} = Cov[(G_Q - A),(G_M - A)]/Var(G_M - A) = Var(G_Q - A)/Var(G_M - A)$. O denominador de $\hat{\beta}$, $Var(G_M - A)$, pode ser expresso por $Var(G_M - A) = Var(G_Q - A) + Var(G_E)$ e, portanto, contempla dois componentes confundidos: (i) superioridade do uso de G_Q em lugar A; (ii) erro no uso de G_M no lugar de G_Q.

Como G_Q é desconhecida, a mesma deve ser estimada com base nos marcadores (G_M), assumindo que os QTLs tem as mesmas propriedades alélicas que os marcadores, isto é, desequilíbrio de ligação médio entre marcas igual desequilíbrio de ligação médio entre marcas e QTLs. Assim, esse último LD pode ser predito dividindo aleatoriamente o total de marcadores em dois grupos sem sobreposição, calculando as matrizes G_{M1} e G_{M2} associadas a esses dois grupos e computando a covariância entre essas duas matrizes.

Conforme Goddard et al. (2011), a quantidade $c = Cov[(G_{M1} - A),(G_{M2} - A)]$ estima $Cov[(G_Q - A),(G_M - A)]$ e consequentemente $Var(G_Q - A)$. O denominador de $\hat{\beta}$ equivale a $den = Cov[(G_{M1} - A),(G_{M2} - A)] + Var(G_{M1} - G_{M2})/2$. Assim, $\hat{\beta} = \frac{c}{den} = \frac{Cov[(G_{M1} - A),(G_{M2} - A)]}{Cov[(G_{M1} - A),(G_{M2} - A)] + Var(G_{M1} - G_{M2})/2}$. A quantidade $Var(G_{M1} - G_{M2})$ estima $2Var(G_E)$, pois $Var(G_{M1} - G_{M2}) = Var(G_Q) + Var(G_E) - 2Var(G_Q) + Var(G_Q) + Var(G_E) = 2Var(G_E).$ Conforme Yang et al. (2010), $Var(G_E) = 1/n_m$. Assim, $\hat{\beta} = \frac{c}{den} = \frac{Cov[(G_{M1} - A),(G_{M2} - A)]}{Cov[(G_{M1} - A),(G_{M2} - A)] + 1/n_m} = \frac{c}{c + 1/n_m}$. Isto pode ser simplificado para $\hat{\beta} = 1 - 1/(cn + 1)$.

A covariância $c = Cov[(G_{M1} - A),(G_{M2} - A)]$ é estimada como a covariância entre os elementos fora da diagonal das matrizes das diferenças $(G_{M1} - A)$ e $(G_{M2} - A)$. A quantidade c pode ser também computada como $c = Cov(G_M, G_{QMAF_{LOW}})$, em que $G_{QMAF_{LOW}}$ é a matriz de parentesco real ao nível dos QTLs ou variantes causais, formada somente com os SNPs de baixa MAF (minor allele frequency), os quais mimicam os referidos QTLs. Intrinsecamente tem-se $\beta = \frac{Cov(G_M, G_{QMAF_{LOW}})}{Var(G_M)}$.

O valor de $\hat{\beta}$ permite determinar quanto da diferença $(G_M - A)$ é devida à melhoria da GWS em relação à seleção fenotípica e quanto é devida à distância entre G_Q e G_M, ou seja, pela falta de determinação dos QTLs pelas marcas. Assim, conforme Goddard et al.(2011), $\hat{\beta}$ pode ser dado também por $\hat{\beta} = n_m/(n_m + Me)$, em que Me é o número efetivo de segmentos cromossômicos, cuja fórmula de cálculo é apresentada mais adiante. Essa proporção $\hat{\beta}$ mede a relação entre número de efeitos a estimar em relação ao número de efeitos a explicar, assumindo que todas as marcas

são diferentes dos QTLs. Esse estimador é $\hat{\beta} \sim E(r^2) = 1/(2+4NeL)$ (Tenesa et al., 2007) ou $\hat{\beta} \sim E(r^2) = 1/(1+4NeL)$ (Sved, 1971).

Quando os marcadores são os QTLs ou estão em desequilíbrio de ligação com os QTLs, G_M propicia mais informações sobre a covariância entre parentes do que a matriz A. Isto ocorre porque a matriz A não considera a variação no parentesco entre os irmãos completos. Uma segunda abordagem é usar o ajuste de g*, conforme detahado a seguir.

(B) Método GBLUP Melhorado

Modelo G-BLUP melhorado

$y = Xb + Zg + e$; $Var(g) = G_Q \sigma_g^2$

Modelo Equivalente G-BLUP melhorado

$q = Wm$

$y = Xb + ZWm + Zg* + e = Xb + Zq + Zg* + e$; $Var(q) = G_M \sigma_q^2$; $Var(g*) = A\sigma_{g*}^2$

, em que g* é o vetor de efeitos poligênicos não capturados pelos marcadores.
Assim, $\sigma_g^2 = \sigma_q^2 + \sigma_{g*}^2$ e, portanto, $Var(g) = G_M \sigma_q^2 + A\sigma_{g*}^2$.

Os componentes de variância σ_q^2 e σ_{g*}^2 podem ser estimados por REML e então tem-se $\hat{\beta} = \hat{\sigma}_q^2/(\hat{\sigma}_q^2 + \hat{\sigma}_{g*}^2)$. Pode-se então estimar G_Q como $\hat{G}_Q = G_M \hat{\sigma}_q^2 + A\hat{\sigma}_{g*}^2 = [A + \hat{\beta}(G_M - A)]\hat{\sigma}_g^2$, em que $\hat{\sigma}_g^2 = \hat{\sigma}_q^2 + \hat{\sigma}_{g*}^2$. A matriz $\hat{G}_Q$ estimada deve então ser usada no lugar de A nas equações de modelo misto para a predição dos valores genéticos dos indivíduos e cômputo de suas acurácias seletivas.

Esse modelo G-BLUP melhorado é equivalente ao RR-BLUP com ajuste do vetor de efeitos poligênicos residuais. Apresenta como vantagem a possibilidade de computar as acurácias seletivas dos indivíduos.

(C) Otimização do G-BLUP na predição de g (catálogo de valores genéticos dos indivíduos)

Um fator que contribui para a redução de G_E é a padronização de W, obtendo-se W_p. Essa padronização reflete positivamente na composição da matriz de parentesco genômico G_m, a qual conterá a média ponderada das relações de parentesco estimadas de cada loco marcador, em que os pesos da ponderação são função da inversa da PEV (variância do erro de predição) associada à variável indicadora W em cada marcador. No caso, a PEV é dada por $PEV = Var(W_i) = 2p_i(1-p_i)$. E a matriz G_m é dada por $G_M = W_p W_p'/n$, em que W_p contém elementos dados por $w_{ij_p} = \frac{(w_{ij} - 2p_i)}{[2p_i(1-p_i)]^{1/2}}$, em que j refere-se a indivíduos. Essa parametrização é também interessante porque não propicia pesos subestimados à informações dos alelos com baixa frequência. Assim, permite detectar alelos raros como não nulos. As parametrizações alternativas propiciam maior peso aos SNPs com alta heterozigose ($2p_i(1-p_i)$).

Nesse caso, os elementos da matriz G_m representam o parentesco realizado médio multi-locos e são dados por $G_{jk} = (1/n)\sum_{i=1}^{n} \frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$. Outro ponto favorável do G-BLUP estimado dessa maneira refere-se à possibilidade de estimação direta (via PEV) da acurácia da GWS.

Os seguintes modelos alternativos podem ser usados para maximizar a eficiência da GWS pelo método BLUP:

a. **RR-BLUP (SNP-BLUP) com ajuste do vetor de efeitos poligênicos residuais**

$y = Xb + ZWm + Zg^* + e = Xb + Zq + Zg^* + e$.

$\hat{g} = W\hat{m} + \hat{g}^* = \hat{q} + \hat{g}^*$

b. **Modelo G-BLUP melhorado 1: β estimado a partir dos marcadores e fenótipos**

$y = Xb + Zg + e$; $Var(g) = G_Q\sigma_g^2$

$y = Xb + ZWm + Zg^* + e = Xb + Zq + Zg^* + e$

$\hat{G}_Q = G_M\hat{\sigma}_q^2 + A\hat{\sigma}_{g^*}^2 = [A + \hat{\beta}(G_M - A)]\hat{\sigma}_g^2$, com $\hat{\beta} = \hat{\sigma}_q^2/(\hat{\sigma}_q^2 + \hat{\sigma}_{g^*}^2)$.

$\hat{g} = \hat{g}_{G_Q}$

c. **Modelo G-BLUP melhorado 2: β estimado a partir dos marcadores e Ne.**

$y = Xb + Zg + e$; $Var(g) = G_Q\sigma_g^2$

$\hat{G}_Q = A + \hat{\beta}(G_M - A)$, com $\hat{\beta} = c/(c + 1/n_m)$ ou $\hat{\beta} = n_m/(n_m + Me)$, em que Me é o número efetivo de segmentos cromossômicos sendo dado por $Me = (2NeL_cN_c)/[\ln(NeL_c)]$ (Goddard et al., 2011, em que L_c é o comprimento médio de um cromossomo em Morgans e N_c é o número de cromossomos).

$\hat{g} = \hat{g}_{G_Q}$

Tem-se também $\hat{\beta} = n_m/(n_m + Me) = 1/(2 + 4NeS/\ln(NeL_c))$ em que $S = L_cN_c/n$ é a distância média entre marcadores ou tamanho do segmento cromossômico, ou seja, $\hat{\beta} \sim E(r^2) = 1/(2 + 4NeS)$, que mede o r^2 entre pares de locos vizinhos, conforme Tenesa et al. (2007). Tomando por base a expressão de Sved (1971), $E(r^2) = 1/(1 + 4NeS)$, e não de Tenesa et al. (2007), tem-se $Me = (2NeL_cN_c)/[\ln(2NeL_c)]$.

(D) Otimização do G-BLUP na predição de m (catálogo de efeitos genéticos dos marcadores)

As mesmas recomendações referentes à obtenção de G_M devem ser seguidas. Mas os dados fenotípicos também devem ser corrigidos e, para $\hat{G}_Q$, o estimador mais adequado é $\hat{G}_Q = \hat{\beta}(G_M - A)$, capitalizando somente a segregação mendeliana.

d. RR-BLUP com correção prévia dos fenótipos (y_c) para os efeitos dos genitores.

$y_c = Xb + ZWm + Zg^* + e = Xb + Zq + Zg^* + e$.

$\hat{g}_s = W_s\hat{m} + 0.5(\hat{g}_m + \hat{g}_p)$: efeito genético predito para os novos indivíduos da geração seguinte (s), a partir de suas matrizes de incidência (W_s) e dos efeitos genéticos preditos de seus genitores maternos e paternos.

e. Modelo G-BLUP com correção prévia dos fenótipos (y_c) para os efeitos dos genitores.

$y_c = Xb + Zg + e$; $Var(g) = G_Q\sigma_g^2$

$\hat{m} = (W'W)^{-1}W'\hat{g}$

$\hat{g}_s = W_s\hat{m} + 0.5(\hat{g}_m + \hat{g}_p)$.

(E) Estimação do Me

O M_e pode ser estimado a partir de:

$$(1/Me) = Var(G_Q - A) = Var(G_M - A) - Var(G_E)$$
$$= Var(G_M - A) - PEV_{G_M} = Var(G_M - A) - (1/n_m)$$

ou $Me = (2NeL_cN_c)/[\ln(NeL_c)]$ (Goddard et al., 2011, em que L_c é o comprimento médio de um cromossomo em Morgans e N_c é o número de cromossomos) ou $(1/Me) = Var(G_Q - A) =$ valor médio de r^2 entre todos os pares de locos (r^2_{all-p}). Assim, a variância nos coeficientes de parentesco em torno de A equivale ao desequilíbrio de ligação médio.

A quantidade $Me = (1/r^2_{all-p})$ tem grande impacto na acurácia seletiva. Essa depende sobretudo de Nh^2/Me. Se Me é baixo essa fração terá valor alto, e a acurácia será alta. No caso, r^2_{all-p} refere-se ao desequilíbrio médio envolvendo todos os pares de SNPs e difere de r^2_{mq} que refere-se ao desequilíbrio envolvendo marcadores vizinhos. As quantidades $\hat{\beta} = \hat{\sigma}^2_q/(\hat{\sigma}^2_q + \hat{\sigma}^2_{g*})$ e $\hat{\beta} = n_m/(n_m + Me)$ são estimadores de r^2_{mq}. O desequilíbrio médio r^2_{all-p} equivale à variância dos coeficientes de parentesco associados aos elementos fora da diagonal de G_M, os quais apresentam média 0. Os elementos da diagonal de G_M apresentam média 1.

Outra abordagem para inferir sobre n_Q (número de locos gênicos) é usar o seu valor esperado, dado o tamanho efetivo (Ne) da população e o tamanho L do genoma da espécie. Com base no tamanho efetivo populacional (Ne), pode-se calcular o

número efetivo de locos ou segmentos cromossômicos (Me) devidos à ligação (segundo esse conceito, para dois gametas quaisquer, o genoma é quebrado em Me segmentos de tamanho igual). Nesse caso, n_Q é dado por n_Q = Me V(q) = Me k, sendo V(q) a heterozigose média de todos os segmentos cromossômicos independentes, ou seja, V(q) = 2p(1-p), em que p é a freqüência alélica média. V(q) é análogo a $V(Z_i)$, sendo que q refere-se aos locos gênicos e Z refere-se aos locos marcadores.

Segundo Goddard, (2008) e conforme apresentado por Resende (2008), a quantidade Me é dada por Me = (2NeL)/[Ln(4NeL)], em que L é o tamanho total do genoma em Morgans. Entretanto, Hayes et al. (2009) relata que o valor mais apropriado para Me situa-se entre 4NeL (que é o número real de segmentos) e (2NeL)/[Ln(4NeL)], sendo uma boa aproximação usar Me = 2NeL, ou seja, assumir o número efetivo de locos como 2NeL. Esse número efetivo de locos deve ser ponderado por uma função da freqüência alélica do gene (via freqüência do marcador), que está implícita em V(q). O valor de n_Q é dado então por n_Q = Me V(q) = Me k, em que V(q) = k é dado por k = 1/[Ln(2Ne)]. Dessa forma, n_Q = 2NeL 1/[Ln(2Ne)]. A quantidade Me V(q) refere-se ao número esperado de marcas com efeitos significativos.

Entretanto, segundo Daetwyler et al. (2010), a abordagem de Goddard (2008) propiciou, via simulação, resultados mais coerentes do que a abordagem de Hayes et al. (2009), embora Daetwyler et al. (2010) parece não ter feito a correção para k. Com dados reais (r^2_{mq} < 1), Hayes et al. (2009) concluíram o contrário.

Geralmente o número de SNPs significativos é maior do que o número de locos pois cada SNP rastreia um grande segmento cromossômico e então o efeito de cada segmento cromossômico é dividido em muitos SNPs. Em gado de leite, o número de SNPs com efeitos significativos variou de 3.000 a 4.000 entre caracteres, dentre cerca de 40.000 marcadores usados (Hayes et al, 2009).

O número máximo de SNPs com efeitos significativos é limitado pelo Ne. Com Ne mais baixo, menor é n_Q. O número real de segmentos cromossômicos total é 4NeL, ou seja, 120.000 em bovinos, que é bem maior que o número efetivo de segmentos.

Na Tabela 33 são apresentados valores de n_Q para bovinos (genoma com L = 30 Morgans) e eucalipto (genoma com L = 13,2 Morgans), para diferentes valores de Ne, usando várias abordagens.

Tabela 33. Número efetivo de segmentos cromossômicos (Me) e de locos (n_Q) em função do tamanho efetivo (Ne) e do comprimento do genoma (L) em bovinos e eucalipto.

Bovinos

Ne	Tam	Me Max.	Me Provavel	-	Me Min.	Correção	Me Provável Corrigido*	Me Min. Corrigido**	Correção 2	Me Prov Corrigido 2***
Ne	Ltot	4 Ne L	2NeL	Ln (4NeL)	2NeL/ Ln (4NeL)	1/*Ln*(2Ne)	2NeL/ *Ln*(2Ne)	2NeL / [Ln (4NeL) *Ln*(2Ne)]	Ln (Ne Lc)	2NeL/ Ln (Ne Lc)
15	30	1800	900	7.50	120.07	0.29	**261**	34.82	2.71	332.34
30	30	3600	1800	8.19	219.82	0.24	**432**	52.76	3.40	529.23
50	30	6000	3000	8.70	344.85	0.22	**660**	75.87	3.91	766.87
100	30	12000	6000	9.39	638.80	0.19	**1140**	121.37	4.61	1302.88
200	30	24000	12000	10.09	1189.79	0.17	**2040**	202.26	5.30	2264.87
500	30	60000	30000	11.00	2726.75	0.14	**4200**	381.75	6.21	4827.34
1000	30	120000	60000	11.70	5130.29	0.13	**7800**	666.94	6.91	8685.89

*** Hayes et al. (2009); **Goddard (2008); ***Goddard et al. (2011).**

Eucalipto

Ne	Tam	Me Max.	Me Provavel	-	Me Min.	Correção	Me Provável Corrigido*	Me Min. Corrigido**	Correção 2	Me Prov Corrigido 2***
Ne	Ltot	4 Ne L	2NeL	Ln (4NeL)	2NeL/ Ln (4NeL)	1/*Ln*(2Ne)	2NeL/ *Ln*(2Ne)	2NeL / [Ln (4NeL) *Ln*(2Ne)]	Ln (Ne Lc)	2NeL/ Ln (Ne Lc)
15	13.2	792	396	6.67	59.33	0.29	**115**	17.21	2.71	146.23
30	13.2	1584	792	7.37	107.50	0.24	**190**	25.80	3.40	232.86
50	13.2	2640	1320	7.88	167.54	0.22	**290**	36.86	3.91	337.42
100	13.2	5280	2640	8.57	307.99	0.19	**502**	58.52	4.61	573.27
200	13.2	10560	5280	9.26	569.90	0.17	**898**	96.88	5.30	996.54
500	13.2	26400	13200	10.18	1296.52	0.14	**1848**	181.51	6.21	2124.03
1000	13.2	52800	26400	10.87	2427.75	0.13	**3432**	315.61	6.91	3821.79

*** Lc = comprimento do cromossomo, aproximadamente igual (1 Morgan) para bovinos e eucalipto, que apresentam 30 e 13 pares de cromossomo respectivamente.**

Na Tabela 34 são apresentados cálculos de r^2_{mq} obtidos via $\hat{\beta} = n_m /(n_m + Me)$ considerando Me conforme Goddard et al. (2011).

Tabela 34. Valores de r² obtidos via $\hat{\beta} = n_m /(n_m + Me)$.

Bovinos

Ne	Me	M	r^2_{mq}
100	1302.88	10000	0.88
100	1302.88	20000	0.94
100	1302.88	30000	0.96
200	2264.87	10000	0.82
200	2264.87	20000	0.90
200	2264.87	30000	0.93
1000	4827.34	10000	0.67
1000	4827.34	20000	0.81
1000	4827.34	30000	0.86

Eucalipto

Ne	Me	M	r^2_{mq}
15	146.23	5000	0.97
15	146.23	10000	0.99
15	146.23	20000	0.99
50	337.42	5000	0.94
50	337.42	10000	0.97
50	337.42	20000	0.98
100	573.27	5000	0.90
100	573.27	10000	0.95
100	573.27	20000	0.97
500	2124.03	5000	0.70
500	2124.03	10000	0.82
500	2124.03	20000	0.90

Suínos

h^2_a	N	Nmarcas	L	Ne	r^2_{mq}	n_Q	râa*
0.2	1000	1000	28	15	0.79	247.0	0.55
0.2	1000	1000	28	30	0.69	410.3	0.42
0.2	**1000**	**1000**	**28**	**50**	**0.61**	**608.0**	**0.32**
0.2	1000	1000	28	100	0.47	1056.9	0.20
0.2	1000	1000	28	400	0.22	3351.0	0.05
0.2	1000	1000	28	**500**	0.19	4053.4	0.04
0.2	1000	1000	28	1000	0.12	7367.5	0.02
0.4	1000	10000	28	15	0.97	247.0	0.77
0.4	1000	10000	28	30	0.96	410.3	0.68
0.4	**1000**	**10000**	**28**	**50**	**0.94**	**608.0**	**0.60**
0.4	1000	10000	28	100	0.90	1056.9	0.48
0.4	1000	10000	28	400	0.74	3351.0	0.25
0.4	1000	10000	28	**500**	0.70	4053.4	0.21
0.4	1000	10000	28	1000	0.57	7367.5	0.13
0.2	2000	1000	28	15	0.79	247.0	0.66
0.2	2000	1000	28	30	0.69	410.3	0.53
0.2	**2000**	**1000**	**28**	**50**	**0.61**	**608.0**	**0.42**
0.2	2000	1000	28	100	0.47	1056.9	0.27
0.2	2000	1000	28	400	0.22	3351.0	0.08
0.2	2000	1000	28	**500**	0.19	4053.4	0.06
0.2	2000	1000	28	1000	0.12	7367.5	0.03
0.4	2000	10000	28	15	0.97	247.0	0.86
0.4	2000	10000	28	30	0.96	410.3	0.79
0.4	**2000**	**10000**	**28**	**50**	**0.94**	**608.0**	**0.72**
0.4	2000	10000	28	100	0.90	1056.9	0.60
0.4	2000	10000	28	400	0.74	3351.0	0.33
0.4	2000	10000	28	**500**	0.70	4053.4	0.29
0.4	2000	10000	28	1000	0.57	7367.5	0.18

*** Acurácia**

Outra forma de calcular Me é a partir da expressão da acurácia dada por $r_{\hat{g}g} = \sqrt{(Nh^2/Me)/[1+(Nh^2/Me)]} = \sqrt{(Nh^2)/(Me+Nh^2)}$ apresentada por Daetwyler et al. (2008). Rearranjando essa expressão tem-se $\hat{Me} = (Nh^2)(1-r^2_{\hat{g}gBLUP})/r^2_{\hat{g}gBLUP}$. Assim, Me pode ser computado a partir da acurácia estimada via método G-BLUP. Se métodos Bayesianos com seleção de covariáveis (BayesB, BayesCpi, BayesDpi, Blasso, Iblasso) são aplicados, o número de QTLs pode ser inferido via $\hat{n}_{QTL} = (Nh^2)(1-r^2_{\hat{g}gBayes})/r^2_{\hat{g}gBayes}$ (Daetwyler et al. (2010). Assim, $\hat{n}_{QTL}$ é uma fração de Me com efeitos mensuráveis sobre o caráter avaliado. A acurácia do G-BLUP, expressa como função de Me independe do número de QTLs governando o caráter, pois não há seleção de covariáveis visto que todos os marcadores são retidos no modelo e supõe-se que todos os Me segmentos possuem genes.

(F) G-BLUP-Het melhorado com heterogeneidade de variância entre SNPs

Com heterogeneidade de variância entre SNPs e sendo D uma matriz diagonal ($diag(D) = \tau_i$, sendo τ_i o componente de variância associado ao loco marcador i; $m \sim (0, D)$) contemplando essa heterogeneidade, a modelagem da estrutura de variância se modifica e as equações de modelo misto tornam-se:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \hat{G}^{*-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

, em que $\hat{G}^* = G_M\hat{\sigma}_q^2 + A\hat{\sigma}_{g^*}^2$, sendo $G_M = (W_p D W_P^{'})/\sigma_g^2$ quando W_p contém elementos dados por $w_{ij_p} = \frac{(w_{ij} - 2p_i)}{[2p_i(1-p_i)]^{1/2}}$. A matriz D é estimada por algum método Bayesiano, segundo o modelo $y = Xb + Wm + Zg^* + e$, com $m \sim (0, D)$.

Essa modelagem gera um método G-BLUP com heterogeneidade de variância e produz resultados similares aos obtidos pelo método BayesA.

(H) G-BLUP-Het melhorado com heterogeneidade de variância e modelagem espacial entre SNPs

Com dependência espacial entre efeitos de SNPs dentro de cromossomos devida ao desequilíbrio de ligação entre eles, modelos espaciais podem ser adotados. Nesse caso, a matriz D deve ser substituída por uma matriz de correlação autoregressiva (AR1) contemplando essa covariância espacial.

Um modelo autorregressivo com variâncias heterogêneas (ARH) pode ser ajustado. Nesse caso, tem-se $Var(g) = Var(W_p m) = W_p \Sigma_{mARH} W_p'$ e para 3 marcas a estrutura de covariância é:

$$\sum_{m\,ARH} = \begin{bmatrix} \sigma^2_{m_1} & \sigma_{m1}\sigma_{m2}\rho^1 & \sigma_{m1}\sigma_{m3}\rho^2 \\ Sim. & \sigma^2_{m_2} & \sigma_{m2}\sigma_{m3}\rho^1 \\ & & \sigma^2_{m_3} \end{bmatrix} \text{ e } G_M = (W_p \sum_{m\,ARH} W_P^{'}) / \sigma_g^2 .$$

Se parte da variância entre SNPs é correlacionada e parte é independente ou não correlacionada, tem-se $Var(g) = Var(W_p m) = W_p \Psi_{ARH} W_p$' e

$$\Psi_{ARH} = \begin{bmatrix} (\sigma^2_{m_{1c}} + \sigma^2_{m_{1nc}}) & \sigma_{m1c}\sigma_{m2c}\rho^1 & \sigma_{m1c}\sigma_{m3c}\rho^2 \\ Sim. & (\sigma^2_{m_{2c}} + \sigma^2_{m_{2nc}}) & \sigma_{m2c}\sigma_{m3c}\rho^1 \\ & & (\sigma^2_{m_{3c}} + \sigma^2_{m_{3nc}}) \end{bmatrix} \text{ e } G_M = (W_p \Psi_{ARH} W_P^{'}) / \sigma_g^2 .$$

Outra estrutura de correlação que pode ser usada é associada a modelos ante-dependência estruturados (SAD, que inclui também heterogeneidade de autocorrelações), em que $Var(g) = Var(W_p m) = W_p \Sigma_{mSAD} W_p$' e a estrutura da matriz de covariância é:

$$\sum_{mSAD} = \begin{bmatrix} \sigma^2_{m_1} & \sigma_{m1}\sigma_{m2}\rho_1 & \sigma_{m1}\sigma_{m3}\rho_1\rho_2 \\ Sim. & \sigma^2_{m_2} & \sigma_{m2}\sigma_{m3}\rho_2 \\ & & \sigma^2_{m_3} \end{bmatrix} \text{ e } G_M = (W_p \Sigma_{mSAD} W_P^{'}) / \sigma_g^2 .$$

Com SNPs correlacionados em parte e também independentes, tem-se $Var(g) = Var(W_p m) = W_p \Psi_{SAD} W_p$' e a estrutura da matriz de covariância é:

$$\Psi_{SAD} = \begin{bmatrix} (\sigma^2_{m_{1c}} + \sigma^2_{m_{1nc}}) & \sigma_{m1c}\sigma_{m2c}\rho_1 & \sigma_{m1c}\sigma_{m3c}\rho_1\rho_2 \\ Sim. & (\sigma^2_{m_{2c}} + \sigma^2_{m_{2nc}}) & \sigma_{m2c}\sigma_{m3c}\rho_2 \\ & & (\sigma^2_{m_{3c}} + \sigma^2_{m_{3nc}}) \end{bmatrix} \text{ e } G_M = (W_p \Psi_{SAD} W_P^{'}) / \sigma_g^2 .$$

Este modelo SAD pode ser estruturado para contemplar 10 atributos: diferentes precisões e heterogeneidade na variável indicadora W_p; diferentes precisões na variável fenotípica y; heterogeneidade de variâncias na variável aleatória m; efeitos autocorrelacionados em m; heterogeneidade de autocorrelações em m; simultaneamente efeitos autocorrelacionados e não correlacionados em m; cômputo da matriz de parentesco G_{FG} visando o uso da informação de ligação (LA); cômputo da matriz G* por meio do ajuste para a mesma endogamia base de G_{FG}; cômputo da matriz $\hat{G}^*$ por meio de regressão de G* em A; consideração da variância de G_{Mijj}, ou seja, o erro de amostragem associado a cada SNP. Alguns dos atributos mencionados aqui são abordados mais adiante.

(I) Correção de Fenótipos com Diferentes Precisões

Para considerar as diferentes precisões na variável fenotípica y, duas alternativas de correção podem ser usadas e essa correção deve ser usada mesmo após a desregressão e correção para os efeitos genéticos dos genitores (correção para estrutura de população). A primeira alternativa foi relatada por Van Raden (2008) e Legarra et al. (2011). A segunda foi relatada por Garrick et al. (2009).

O modelo misto tradicional pode ser especificado de duas maneiras:

(i) Modelo para fenótipos com iguais precisões e homogeneidade de variância residual

$$y = Xb + Zg + e;$$
$$g \sim N(0, A\sigma_g^2);\ e \sim N(0, I\sigma_e^2).$$

Esse modelo conduz à seguintes equações de modelo misto:

$$\begin{bmatrix} X'(I\sigma_e^2)^{-1}X & X'(I\sigma_e^2)^{-1}Z \\ Z'(I\sigma_e^2)^{-1}X & Z'(I\sigma_e^2)^{-1}Z + A^{-1}1/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'(I\sigma_e^2)^{-1}y \\ Z'(I\sigma_e^2)^{-1}y \end{bmatrix}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

(ii) Modelo para fenótipos com diferentes precisões e homogeneidade de variância residual

$$y = Xb + Zg + e;$$
$$g \sim N(0, A\sigma_g^2);\ e \sim N(0, R\sigma_e^2)$$

, em que R é uma matriz diagonal contendo os diferentes pesos associados às diferentes precisões dos fenótipos. Esse modelo conduz às seguintes equações equivalentes de modelo misto:

$$\begin{bmatrix} X'(R\sigma_e^2)^{-1}X & X'(R\sigma_e^2)^{-1}Z \\ Z'(R\sigma_e^2)^{-1}X & Z'(R\sigma_e^2)^{-1}Z + A^{-1}1/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'(R\sigma_e^2)^{-1}y \\ Z'(R\sigma_e^2)^{-1}y \end{bmatrix}$$

$$\begin{bmatrix} X'(R^{-1}1/\sigma_e^2)X & X'(R^{-1}1/\sigma_e^2)Z \\ Z'(R^{-1}1/\sigma_e^2)X & Z'(R^{-1}1/\sigma_e^2)Z + A^{-1}1/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'(R^{-1}1/\sigma_e^2)y \\ Z'(R^{-1}1/\sigma_e^2)y \end{bmatrix}$$

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + A^{-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}.$$

R foi definido da seguinte forma por Van Raden (2008):

$R_{ii} = \frac{1}{r_{\hat{g}gi}^2} - 1$, em que $r_{\hat{g}gi}^2$ é a confiabilidade associada ao fenótipo médio das filhas do do indivíduo i, corrigida para os efeitos dos genitores, a qual, quando associada a

progênies de meios irmãos, pode ser dada por $r^2_{\hat{g}gi} = \frac{(1/4)\sigma^2_g}{(1/4)\sigma^2_g + [(3/4)\sigma^2_g + \sigma^2_e]/n_{Fi}}$, em que n_{Fi} é o número de filhas de um genitor i. Desenvolvendo a expressão de R_{ii} obtem-se $R_{ii} = \frac{1}{r^2_{\hat{g}gi}} - 1 = \frac{(1/4)\sigma^2_g + [(3/4)\sigma^2_g + \sigma^2_e]/n_{Fi}}{(1/4)\sigma^2_g} - \frac{(1/4)\sigma^2_g}{(1/4)\sigma^2_g} = \frac{[(3/4)\sigma^2_g + \sigma^2_e]/n_{Fi}}{(1/4)\sigma^2_g} = \frac{(3/4)\sigma^2_g + \sigma^2_e}{(1/4)\sigma^2_g}\frac{1}{n_{Fi}}$.

As equações de modelo misto tornam-se então

$$\begin{bmatrix} X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})X & X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})Z \\ Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})X & Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})Z + A^{-1}(\sigma^2_e/\sigma^2_g) \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})y \\ Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})y \end{bmatrix}$$

Uma vez que R_{ii} já considera a herdabilidade do caráter, a fração (σ^2_e/σ^2_g) simplifica-se para um. Assim, tem-se

$$\begin{bmatrix} X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})X & X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})Z \\ Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})X & Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})Z + A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})y \\ Z'\frac{(1/4)\sigma^2_g}{(3/4)\sigma^2_g + \sigma^2_e}(I n_{Fi})y \end{bmatrix}.$$

Multiplicando-se todos os termos da equação por $\frac{(3/4)\sigma^2_g + \sigma^2_e}{(1/4)\sigma^2_g}$ obtém-se

$$\begin{bmatrix} X'(I n_{Fi})X & X'(I n_{Fi})Z \\ Z'(I n_{Fi})X & Z'(I n_{Fi})Z + A^{-1}\frac{(3/4)\sigma^2_g + \sigma^2_e}{(1/4)\sigma^2_g} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'(I n_{Fi})y \\ Z'(I n_{Fi})y \end{bmatrix},$$ que são as equações de modelo misto para um modelo de reprodutor, ponderadas pelo tamanho de progênie de cada um.

Garrick et al. (2009) relatam que as observações desregressadas apresentam heterogeneidade de variância quando os indivíduos apresentam diferentes confiabilidades. Sugerem então os seguinte peso para as informações: $P_{ii} = \frac{\sigma^2_e}{[(1-r^2_{mq})(1-r^2_{\hat{g}gi})/r^2_{\hat{g}gi}]\sigma^2_g} = \frac{1-h^2}{[(1-r^2_{mq})(1-r^2_{\hat{g}gi})/r^2_{\hat{g}gi}]h^2}$, em que r^2_{mq} refere-se ao desequilíbrio envolvendo marcadores vizinhos ou proporção da variação genética explicada pelos marcadores. As quantidades $\hat{r}^2_{mq} = \hat{\sigma}^2_q/(\hat{\sigma}^2_q + \hat{\sigma}^2_{g^*})$ e $\hat{r}^2_{mq} = n_m/(n_m + Me)$ são estimadores de r^2_{mq}. Tem-se também $E(r^2_{mq}) = 1/(2 + 4Ne\ S)$ (Tenesa et al., 2007), em que S é a distância média entre marcadores ou tamanho do segmento cromossômico que não sofre recombinação dentro dele.

Mas como $\frac{\sigma^2_e}{\sigma^2_g}$ é constante para todos os indivíduos tem-se que $P_{ii} = \frac{1}{[(1-r^2_{mq})(1-r^2_{\hat{g}gi})/r^2_{\hat{g}gi}]}$. Também $\frac{1}{[(1-r^2_{mq})}$ é constante para todos os indivíduos e tem-se que $P_{ii} = \frac{1}{(1-r^2_{\hat{g}gi})/r^2_{\hat{g}gi}} = \frac{1}{(1/r^2_{\hat{g}gi})-1}$. Verifica-se que esses pesos são equivalentes à

ponderação por R_{ii}^{-1} nas equações de modelo misto, conforme Van Raden et al. (2008), fato não notado por Garrick et al. (2009). Também os pesos não dependem da quantidade r_{mq}^2.

Outra opção é transformar o modelo para $R^{-1/2}y = R^{-1/2}Xb + R^{-1/2}Zg + R^{-1/2}e;\ g \sim N(0, A\sigma_g^2);\ e \sim N(0, I\sigma_e^2)$ e usar as tradicionais equações de modelo misto:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}.$$

(J) G-BLUP com Genotipagem dos Genitores e Fenotipagem dos Descendentes

Nesse caso, usando valores genéticos desregressados e corrigidos para os efeitos dos genitores, tem-se:

$$\left[Z'R^{-1}Z + G^{-1}(\sigma_e^2/\sigma_g^2)\right][\tilde{g}] = \left[Z'R^{-1}\ (y - X\hat{b})\right];$$

$$R_{ii} = \frac{1}{r_{\hat{g}gi}^2} - 1 = \frac{(3/4)\sigma_g^2 + \sigma_e^2}{(1/4)\sigma_g^2}\frac{1}{n_{Fi}}.$$

Com apenas uma observação desregressada por genitor, tem-se Z = I, e, portanto:

$$\left[R^{-1} + G^{-1}(\sigma_e^2/\sigma_g^2)\right][\tilde{g}] = \left[R^{-1}\ (y - X\hat{b})\right];$$

$$\left[I\frac{n_{Fi}(1/4)\sigma_g^2}{(3/4)\sigma_g^2 + \sigma_e^2} + G^{-1}(\sigma_e^2/\sigma_g^2)\right][\tilde{g}] = \left[I\frac{n_{Fi}(1/4)\sigma_g^2}{(3/4)\sigma_g^2 + \sigma_e^2}\ (y - X\hat{b})\right].$$

Uma vez que R_{ii} já considera a herdabilidade do caráter, tem-se $(\sigma_e^2/\sigma_g^2 = 1)$ e:

$$\left[I\frac{n_{Fi}(1/4)\sigma_g^2}{(3/4)\sigma_g^2 + \sigma_e^2} + G^{-1}\right][\tilde{g}] = \left[I\frac{n_{Fi}(1/4)\sigma_g^2}{(3/4)\sigma_g^2 + \sigma_e^2}\ (y - X\hat{b})\right].$$

Multiplicando-se todos os termos da equação por $\frac{(3/4)\sigma_g^2 + \sigma_e^2}{\sigma_g^2}$ obtém-se

$\left[I(n_{Fi}(1/4)) + G^{-1}\frac{(3/4)\sigma_g^2 + \sigma_e^2}{\sigma_g^2}\right][\tilde{g}] = \left[I(n_{Fi}(1/4))\ (y - X\hat{b})\right]$ que são as equações de modelo misto para um modelo individual reduzido (em que a matriz Z é composta por valores o e 0.5), ponderadas pelo tamanho de progênie de cada um. Esse modelo estima o valor genético aditivo total e não apenas o metade dele como o faz o modelo de reprodutor. Assim, embora usando indivíduos genotipados diferentes dos fenotipados, o uso da matriz R dada por $R_{ii} = \frac{1}{r_{\hat{g}gi}^2} - 1 = \frac{(3/4)\sigma_g^2 + \sigma_e^2}{(1/4)\sigma_g^2}\frac{1}{n_{Fi}}$ conduz à estimação dos valores genéticos aditivos totais para marcadores e indivíduos e não apenas metade deles.

(K) Otimização do G-BLUP Simultâneo em Indivíduos Genotipados e não Genotipados

A avaliação genética em um programa de melhoramento genético envolve simultaneamente indivíduos fenotipados e genotipados, apenas fenotipados e apenas genotipados.

Para a avaliação global das três classes de indivíduos em um único passo, o mesmo modelo $y = Xb + Zg + e$ pode ser usado, porém com uma alteração (substituição da matriz G pela matriz H) nas equações de modelo misto, conforme Misztal et al.(2009): $\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+H^{-1}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \end{bmatrix}$.

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 $H=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & G \end{bmatrix}=A+\begin{bmatrix} 0 & 0 \\ 0 & G-A_{22} \end{bmatrix}$, 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:
$H^{-1}=A^{-1}+\begin{bmatrix} 0 & 0 \\ 0 & G^{-1}-A_{22}^{-1} \end{bmatrix}=\begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22}+G^{-1}-A_{22}^{-1} \end{bmatrix}$, em que A_{22}^{-1} é a inversa da matriz de parentesco baseada em pedigree para os indivíduos somente genotipados. Métodos distintos para cômputo direto das inversas de matrizes de parentesco foram apresentados por Henderson (1976) e Thompson (1977).

Outra forma de expressar H é por meio de $H=\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}=\begin{bmatrix} A_{11}+A_{12}A_{22}^{-1}(G-A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}$. Verifica-se que os efeitos de G sobre os outros blocos da matriz H são determinados pelas regressões matriciais do tipo $A_{12}A_{22}^{-1}$, ou seja, são baseados inteiramente nas informações de pedigree e não fazem uso da informação genômica nessa regressão. No entanto, os genótipos marcadores podem também propiciar informação nessas regressões.

Meuwissen et al. (2011) relatam que essa forma de construção de H produz estimativas viesadas e menos acurada de valores genéticos devidas aos seguintes fatores: (i) ausência de regressão de $(G-A_{22})$ em A visando considerar os erros de estimação em G; (ii) não uso das informações de marcadores nas regressões matriciais usadas para propagar a informação genômica dos indivíduos genotipados para os não genotipados; (iii) diferentes escalas entre as informações genômicas e de pedigree.

Tais autores propuseram então o método LDLAb para construir uma matrix H que verdadeiramente combina todas as informações genômicas e de pedigree em uma abordagem unificada. O método LDLAb possui as seguintes características: (i) faz a regressão de $(G-A_{22})$ em A, conforme tópico anterior; (ii) permite propagar a informação genômica dos indivíduos genotipados para os não genotipados, por meio da matriz G_{FG}; (iii) substitui a matriz de parentesco A por uma matriz G_{FG} baseada

em análise de ligação (LA) conforme Fernando e Grossman (1989), usando a mesma população base de A, permitindo escalas iguais entre as informações genômicas e de pedigree. O método LDLAb maximiza a acurácia seletiva, embora seja computacionalmente mais oneroso.

Segundo o método LDLAb, a matriz H passa então a ser dada por

$$H_{LDLAb} = \begin{bmatrix} G_{FG11} & G_{FG12} \\ G_{FG21} & \hat{G}^* \end{bmatrix} = \begin{bmatrix} G_{FG11} + G_{FG12}G_{FG22}^{-1}(\hat{G}^* - G_{FG22})G_{FG22}^{-1}G_{FG21} & G_{FG12}G_{FG22}^{-1}\hat{G}^* \\ \hat{G}^* G_{FG22}^{-1}G_{FG21} & \hat{G}^* \end{bmatrix}.$$

Assim, o método segue os seguintes passos:

a. Cômputo da matriz de parentesco G_{FG} visando o uso da informação de ligação (LA).
b. Cômputo da matriz G* por meio do ajuste de $G = W_p W_p'/n$, em que W_p contém elementos dados por $w_{ij_p} = \frac{(w_{ij} - 2p_i)}{[2p_i(1-p_i)]^{1/2}}$, para a mesma endogamia base de G_{FG}.
c. Cômputo da matriz $\hat{G}^*$ por meio de $\hat{G}^* = A + \hat{\beta}(G^* - A)$, em $\hat{\beta}$ que foi definido em tópico anterior.
d. Construir a matriz $H_{LDLAb} = \begin{bmatrix} G_{FG11} + G_{FG12}G_{FG22}^{-1}(\hat{G}^* - G_{FG22})G_{FG22}^{-1}G_{FG21} & G_{FG12}G_{FG22}^{-1}\hat{G}^* \\ \hat{G}^* G_{FG22}^{-1}G_{FG21} & \hat{G}^* \end{bmatrix}$.

O método LDLAb utiliza completamente a informação LA contida nos dados de marcadores moleculares. Por usar estrutura de família, o método G-BLUP permite usar a informação LA. A regressão matricial $G_{FG12}G_{FG22}^{-1}$ substitui $A_{12}A_{22}^{-1}$, e, portanto, considera a informação molecular que está contida em G_{FG}. Fica provado então que a GWS usa ambos LA e LD. O método FG usa apenas LA.

O método LDLA puro exige que todos os indivíduos da população base sejam também genotipados. Na seleção genômica não se tem essas informações, de forma que o tradicional método LDLA não pode ser usado. Mas o método G-BLUP Simultâneo em Indivíduos Genotipados e não Genotipados fornece um meio de propagar a informação genômica dos indivíduos das gerações atuais até os indivíduos fundadores da população base por meio do pedigree.

O cômputo da matriz G* para a mesma endogamia base de G_{FG} é descrito a seguir. As matrizes G e A_{22} devem ser expressas na mesma escala. Caso contrário, haverão diferenças entre elas, mesmo se as relações de parentesco via marcadores e via pedigree forem as mesmas. A transformação para a mesma escala faz uso das estatísticas F de Wright referentes à coeficientes de endogamia, definidas a seguir:

F_{st}: endogamia da população base (endogamia antiga).
F_{is}: endogamia contribuída pela população corrente ou atual (endogamia nova).
F_{it}: endogamia total do indivíduo i.
$F_{it} = F_{st} + (1 - F_{st})F_{is}$.
$F_{is} = (F_{it} - F_{st})/(1 - F_{st})$.

O ajuste consiste em extrair de G o F_{st} e muda-lo para aquele calculado de A_{22} (chamado A_{st}) e recalcular a matriz G (obtendo G*) usando A_{st} como endogamia geral da população e calculando de G a quantidade F_{is}. Assim, as seguintes quantidades devem ser calculadas:

$F_{it} = (G_{ii} - 1)$: elementos da diagonal de G menos 1. Nesse caso, F_{st} é a endogamia média na população base, ou seja, a média dos elementos da diagonal de G menos 1.

$F_{is} = (G_{ii} - 1 - F_{st})/(1 - F_{st})$.

$G_{ii}^{*} = A_{st} + (1 - A_{st})F_{is} + 1$: endogamia total do indivíduo i calculado mudando a endogamia básica para aquela de A_{22}.

A_{st} : média dos elementos da diagonal de A_{22} menos 1.

Dessa forma, G_{ii}^{*} são os elementos da diagonal de G re-escalados. De maneira similar os elementos fora da diagonal de G são re-escalados usando os mesmos valores de F_{st} e A_{st}, baseados nas diagonais de G e A_{22}, respectivamente, e transformando os numeradores do parentesco em coancestrias (ϕ), ou seja, dividindo por 2 e posteriormente, transformando as coancestrias para o mesmo nível de endogamia por meio de $G_{ji}^{*} = 2[A_{st} + (1 - A_{st})\phi_{jis}]$, em que $\phi_{jis} = (G_{ji}/2 - F_{st})/(1 - F_{st})$ é a coancestria entre os indivíduos j e i, relativa à endogamia básica de F_{st}. A matriz $\hat{G}^{*} = \begin{bmatrix} G_{ii}^{*} & G_{ij}^{*} \\ G_{ji}^{*} & G_{jj}^{*} \end{bmatrix}$ é então usada na matriz $H_{LDLAb} = \begin{bmatrix} G_{FG11} + G_{FG12}G_{FG22}^{-1}(\hat{G}^{*} - G_{FG22})G_{FG22}^{-1}G_{FG21} & G_{FG12}G_{FG22}^{-1}\hat{G}^{*} \\ \hat{G}^{*}G_{FG22}^{-1}G_{FG21} & \hat{G}^{*} \end{bmatrix}$.

Uma outra abordagem que pode conduzir a melhoramento da GWS é o uso da teoria da coalescência. A ligação gênica conduz ao fato de que pontos próximos no cromossomo tenham a mesma árvore de coalescência. Alelos IBS na geração atual e que eram IBS na geração inicial são IBD e provavelmente estão em LD. A teoria da coalescência trata todos os alelos em um loco como sendo IBD e então modela a probabilidade de ocorrência de mutações causando-os a não serem IBS. As seguintes relações podem ser descritas:

Coalescência: IBD em LD: interesse do RR-BLUP.
IBS: IBD em LD e LE + mutantes novos: interesse do G-BLUP.
IBD: LE e LD: LE interesse do A-BLUP.

(L) Disponibilidade de duas estimativas de valor genético em cada indivíduo: BLUP fenotípico + BLUP GWS

Essa situação ocorre quando estão disponíveis os valores genéticos preditos para o caráter com base em dados fenotípicos (a) e genotípicos de marcas (g). Um índice de seleção pode ser estabelecido usando essas duas informações, cuja covariância equivale a $r^2_{\hat{g}g} r^2_{\hat{a}a}$, em que $r^2_{\hat{g}g}$ é a confiabilidade da seleção genômica e $r^2_{\hat{a}a}$ é a confiabilidade da predição dos valores genéticos usando dados fenotípicos.

Tal índice é dado por:
$I = b_1 g + b_2 a$
Os coeficientes de ponderação (b_i) do índice são dados por:
$b = P^{-1}C$, em que:

$P = \begin{bmatrix} r^2_{\hat{g}g} & r^2_{\hat{g}g} r^2_{\hat{a}a} \\ r^2_{\hat{g}g} r^2_{\hat{a}a} & r^2_{\hat{a}a} \end{bmatrix}$ $C = \begin{bmatrix} r^2_{\hat{g}g} \\ r^2_{\hat{a}a} \end{bmatrix}$ = vetor de covariância genética entre o valor genético e as duas fontes de informação.

Resolvendo o sistema de equações, obtêm-se os seguintes coeficientes de ponderação:

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} (1 - r^2_{\hat{a}a}) / (1 - r^2_{\hat{g}g} r^2_{\hat{a}a}) \\ (1 - r^2_{\hat{g}g}) / (1 - r^2_{\hat{g}g} r^2_{\hat{a}a}) \end{bmatrix}.$$

O aumento na acurácia pela inclusão da informação molecular é dado por
$r_{aum} = \{ r^2_{\hat{g}g} / (1 - r^2_{\hat{g}g} r^2_{\hat{a}a}) [(1 - r^2_{\hat{a}a})^2] \}^{1/2}$.

6.27 Genética de Populações Genômica Ampla (GWPG)

Com o advento da genotipagem ampla via marcadores SNPs e sequenciamento tem-se acesso a todos os locos do genoma e a Genética de Populações faz parte integral dos procedimentos de estimação de componentes de variância dos caracteres na população e da predição dos valores genéticos dos indivíduos. As seguintes quantidades são estimadas via análise apenas dos marcadores, sem associação com fenótipos:

- Call Rate:

Proporção dos indivíduos nos quais a genotipagem com determinada marca foi efetuada com sucesso.

- Freqüências Alélicas:

Assumindo os alelos de cada marca como estando em equilíbrio de Hardy-Weinberg na população, o cálculo das frequências alélicas (p_i) é realizado por $p_i = N_2/N + (1/2)\ N_1/N$, sendo o calculo realizado para cada coluna de marcador no arquivo de dados em que N_2 é o numero de códigos 2 na referida coluna no arquivo e N_1 é o numero de códigos 1.

- Frequência do Alelo menos Frequente (MAF):

MAF = min (p_i , 1-p_i)
Geralmente os marcadores úteis são aqueles com MAF maior que 5% ou 10%.

- Heterozigose média ou variância da variável indicadora W (Binomial):

H = 2 p_i (1-p_i). equivale também à média da variável indicadora T dos efeitos de dominância.

- Variância da heterozigose (h) ou da variável indicadora S (Bernoulli):

Var (h) = H (1-H)

- Desequilíbrio de ligação entre pares de locos vizinhos:

Usando as matriz de incidência W dos marcadores o valor de r pode ser dado por $r_{(a,b)} = \frac{Cov(W_{ia},W_{ib})}{[Var(W_{ia})]^{1/2}[Var(W_{ib})]^{1/2}}$, em que W é dada conforme abaixo.

Indivíduo	N. Alelos Loco Marcador a (W_a)	N. Alelos Loco Marcador b (W_b)
1	0	0
2	2	1
3	1	1
4	1	0
5	2	1
Correlação r	r = 0.76	r^2= 0.58

A quantidade r^2 estima r^2_{mq}, a partir do qual pode-se estimar Ne, via $r^2_{mq} = \frac{n_m}{n_m + 2N_e L}$, quando se conhece n_m e L (tamanho do genoma).

- Desequilíbrio de ligação entre todos os pares de locos:

Calculado de maneira similar ao r^2 acima, porém envolvendo todos os pares de locos ($\bar{r}^2_{tpl}$). Fornece uma estimativa do número efetivo de segmentos cromossômicos (Me) por meio da expressão $Me = \frac{1}{\bar{r}^2_{tpl}}$.

Goddard et al. (2011) acrescenta na expressão de r^2_{mq} uma divisão por $Ln(Ne\ L/k)$, em que k é o número de cromossomos. Quanto maior o tamanho L/k do cromossomo melhor (existem mais marcadores no cromossomo ajudando a capturar o mesmo QTL). A expressão torna-se então:

$r^2_{mq} = \frac{m}{m + 2N_e L/[\ Ln(Ne\ L/k)]}$, em que a fração $Ln(Ne\ L/k)$ advém do fato de se considerar o LD entre todos os marcadores dentro de cromossomo e a marca alvo e não apenas o vizinho mais próximo e o alvo.

- Estimação de r^2_{mq} via $\bar{r}^2_{tpl}$:

A partir da expressão $r^2_{mq} = \frac{n_m}{n_m + Me}$, r^2_{mq} pode ser estimado por $r^2_{mq} = \frac{n_m\ \bar{r}^2_{tpl}}{n_m\ \bar{r}^2_{tpl} + 1}$.

- Variância dos coeficientes de parentesco:

O desequilíbrio de ligação entre todos os pares de locos permite também estimar a variância ($Var(\rho_g)$) dos coeficientes de parentesco (ρ_g *ou* G_{jk}) na matriz de parentesco genômico G. Tem-se a igualdade $Var(\rho_g) = \bar{r}^2_{tpl} = \frac{1}{Me}$.

- Variância dos coeficientes de parentesco genéticos aditivos entre irmãos completos:

Genótipos Marcadores	Número de Alelos do Marcador (Binomial, n =2)	Proporção de uma Binomial com n =2 (r_g*)	Frequência Genotípica (f)	Medias por Genótipo: = r_g * f	Desvio de r	Desvio Quadrático	Desvio Quadrático * Frequência
MM	0	0	0.25	0	-0.5	0.25	0.0625
Mm	1	0.5	0.5	0.25	0	0	0
mm	2	1	0.25	0.25	0.5	0.25	0.0625
-				Média Geral $\bar{\rho}_g$ = 0.50			Variância ($Var(\rho_g)$) = 0.125

*** r_a: correlação genética aditiva entre indivíduos irmãos germanos.**

Verifica-se que a variância ($Var(\rho_g)$) equivale a 0.125 para um loco. Para n_{Qd} locos segregantes ou segmentos cromossômicos independentes dentro de família, tem-se $Var(\rho_g) = \frac{1}{8n_{Qd}}$ para famílias de irmãos completos.

- Variância dos coeficientes de parentesco genéticos de dominância entre irmãos completos:

Genótipos Marcadores	Efeitos de Dominância: Distribuição Bernoulli (f)	Corelação Genética de Dominância (r_d)	Medias por Classe de Genótipo: = r_d * f	Efeitos: Desvio da Bernoulli	Desvio Quadrático	Variância: Desvio Quadrático * Frequência (1-p) e p
MM e mm	0	0	0 x 0 = 0	1-0.25=-0.25	0.0625	0.046875
Mm	1	0.25	0.25 *1=0.25	1-0.25=0.75	0.5625	0.140625
-	-	-	**p=0.25**	-	-	**0.1875**

Verifica-se que a variância ($Var(\rho_d)$) equivale a 0.1875 para um loco. Para n_{Qd} locos tem-se $Var(\rho_d) = \frac{1}{0.1875\, n_{Qd}}$ para famílias de irmãos completos. Outra denominação para ρ_d é coeficiente de fraternidade.

Outra abordagem para cômputo de $Var(\rho_g)$ é apresentada por Stam (1980):

$$Var(\rho_g) = \left(\frac{0.5}{[2(L+k)]}\right)^2,$$

em que L é o tamanho do genoma em Morgans e k é o número de cromossomos. Para eucalipto (L = 13) tem-se $Var(\rho_g)$ = 0.0048 e o desvio padrão equivale a 0.0693. Assim, a correlação genética aditiva dentro de famílias de irmãos germanos varia de cerca de 0. 30 a cerca de 0.70.

A influência do número n_i de indivíduos por família na $Var(\rho_g)$, para o caso de um loco, dada por $Var(\rho_{g1}) = [n_i/(n_i-1)]\,0.125$, é apresentada a seguir.

$Var(\rho_{g1}) = [n_i/(n_i-1)]\,0.125$	n_i	% de 0.125
0.1667	4	0.75
0.1429	8	0.88
0.1364	12	0.92
0.1304	24	0.96
0.1277	**48**	**0.98**
0.1263	100	0.99
0.1256	200	1.00

Verifica-se que, com o aumento de n_i, a $Var(\rho_{g1})$ tende a 0.125. Entre 20 e 50 indivíduos por família já ocorre a estabilização de $Var(\rho_{g1})$. Tamanho de família muito pequeno é também um fator de aumento em $Var(\rho_g)$.

- Estimação do Tamanho Efetivo (Ne) via $\bar{r}^2_{tpl}$:

A partir da expressão $r^2_{mq} = \frac{n_m}{n_m + Me}$, estima-se $r^2_{mq} = \frac{n_m\, \bar{r}^2_{tpl}}{n_m\, \bar{r}^2_{tpl} + 1}$.

A partir de $r^2_{mq} = \frac{n_m}{n_m + 2Ne\, L}$ e conhecendo-se L, estima-se $Ne = \frac{n_m(1+r^2_{mq})}{2r^2_{mq}\, L}$.

6.28 Genética Quantitativa Genômica Ampla (GWQG) (198)

A superioridade da GWS sobre a seleção baseada em fenótipos pode ser atribuída a cinco fatores:

(i) uso da matriz de parentesco real e própria de cada caráter (desde que seja empregado um método de seleção de covariáveis), fato que aumenta a acurácia seletiva;

(ii) viabilização da seleção precoce direta (SPD), que aumenta o ganho genético por unidade de tempo;

(iii) permissão da avaliação repetida de cada alelo (propicia repetição experimental) sem o uso de testes clonais e de progênies, fato que aumenta a acurácia seletiva;

(iv) uso de maior número de informações, combinando três tipos de informação (fenotípica, genotípica e genealógica) para corrigir os dados e fazer a análise genômica, fato que aumenta a acurácia

(v) uso de uma Genética Quantitativa mais realística.

Generalização da Genome Wide

No contexto da genotipagem em larga escala surgiram os termos seleção genômica ampla (GWS: Genome Wide Selection), estudos de associação genômica ampla (GWAS: Genome Wide Association Studies), genética de populações genômica ampla (GWPG: Genome Wide Population Genetics) e genética quantitativa genômica ampla (GWQG: Genome Wide Quantitative Genetics). A GWAS e a GWS vieram substituir a análise de QTL e a MAS, respectivamente.

Variâncias dos coeficientes de parentesco dentro de famílias.

Conforme cálculos da variância da distribuição binomial mostrados acima, tem-se para um loco $Var(\rho_g) = \frac{1}{8}$ para famílias de irmãos germanos. Para n_{Qd} locos segregantes dentro de famílias tem-se $Var(\rho_g) = \frac{1}{8n_{Qd}}$. Para famílias de meios irmãos tem-se $Var(\rho_g) = \frac{1}{16}$ para um loco e $Var(\rho_g) = \frac{1}{16n_{Qd}}$ para n_{Qd} locos. No quadro a seguir essa questão é ilustrada para o caso de famílias de irmãos completos.

Irmãos Completos

n_{Qd}	Variância	Desvio	LIIC: 0.5 - 3 desvios	LSIC: 0.5 + 3 desvios
1	0.1250	0.354	-0.56	1.56
2	0.0625	0.250	-0.25	1.25
5	0.0250	0.158	0.03	0.97
10	0.0125	0.112	0.16	0.84
35	**0.0036**	**0.060**	**0.32**	**0.68**
100	0.0013	0.035	0.39	0.61
200	0.0006	0.025	0.43	0.58
300	0.0004	0.020	0.44	0.56
400	0.0003	0.018	0.45	0.55
500	0.0003	0.016	0.45	0.55
600	0.0002	0.014	0.46	0.54
700	0.0002	0.013	0.46	0.54
800	0.0002	0.013	0.46	0.54
900	0.0001	0.012	0.46	0.54
1000	0.0001	0.011	0.47	0.53
2000	0.0001	0.008	0.48	0.52
3000	0.0000	0.007	0.48	0.52
4000	0.0000	0.006	0.48	0.52

Verifica-se que com 35 locos segregando dentro de família, os coeficientes de parentesco entre pares de indivíduos dentro de família variam de 0.38 a 0.62. Portanto, podem se afastar bastante de 0.5.

No quadro a seguir essa questão é ilustrada para o caso de famílias de meios irmãos.

Meios Irmãos

n_{Qd}	Variância	Desvio	LIIC: 0.5 - 3 desvios	LSIC: 0.5 + 3 desvios	Fração da **Variância Dentro de Família:** LIIC	Fração da **Variância Dentro de Família:** LSIC
1	0.0625	0.2500	-0.50	1.00	1.50	0.00
2	0.0313	0.1768	-0.28	0.78	1.28	0.22
5	0.0125	0.1118	-0.09	0.59	1.09	0.41
10	0.0063	0.0791	0.01	0.49	0.99	0.51
100	0.0006	0.0250	0.18	0.33	0.83	0.68
200	0.0003	0.0177	0.20	0.30	0.80	0.70
300	0.0002	0.0144	0.21	0.29	0.79	0.71
400	0.0002	0.0125	0.21	0.29	0.79	0.71
500	0.0001	0.0112	0.22	0.28	0.78	0.72
600	0.0001	0.0102	0.22	0.28	0.78	0.72
700	0.0001	0.0094	0.22	0.28	0.78	0.72
800	0.0001	0.0088	0.22	0.28	0.78	0.72
900	0.0001	0.0083	0.23	0.28	0.78	0.73
1000	0.0001	0.0079	0.23	0.27	0.77	0.73
2000	0.0000	0.0056	0.23	0.27	0.77	0.73
3000	0.0000	0.0046	0.24	0.26	0.76	0.74
4000	0.0000	0.0040	0.24	0.26	0.76	0.74

Verifica-se que com 100 locos segregando dentro de família, os coeficientes de parentesco entre pares de indivíduos dentro de família variam de 0.18 a 0.33. Portanto, podem se afastar bastante de 0.25. Também, a fração da variância genética dentro de família retida, afasta-se de 0.75.

A $Var(\rho_g)$ genômica ampla em k cromossomos para meios irmãos pode também ser dada conforme Hill (1993):

$$Var(\rho_g) = [1/(128\ L^2)]\ [4L - k + \sum e^{-4\ell_i}]$$

em que:

L: é o tamanho total do genoma;

ℓ_i : é o comprimento do cromossomo i.

O termo $\sum e^{-4\ell_i}$ tende a zero para ℓ_i grande, de forma que tem-se $Var(\rho_g) = 1/(32\ L) - k/128\ L^2$.

Em humanos (L = 35 e k = 22) tem-se:

$$Var(\rho_g) = 1/(32\ L) - 22/128\ L^2$$

$$Var(\rho_g) = 0.00075$$

.

Usando a expressão $Var(\rho_g) = \frac{1}{16 n_{Qd}}$ tem-se:

$$n_{Qd} = \frac{1}{16\ Var(\rho_g)} = 83.05$$

Assim, 83 locos estão segregando dentro de famílias de meios irmãos em humanos.

Para irmãos germanos, em humanos, tem-se:

$$Var(\rho_g) = 1/(16\ L) - 22/64\ L^2$$

$$Var(\rho_g) = \frac{1}{8 n_{Qd}} = 0.001505$$

$$n_{Qd} = \frac{1}{8\ Var(\rho_g)} = 83.05$$

Para irmãos germanos, em Eucalyptus (L = 13,2), tem-se:

$$Var(\rho_g) = 1/(16\ L) - 13.2/64\ L^2$$

$$Var(\rho_g) = \frac{1}{8 n_{Qd}} = 0.003569$$

$$n_{Qd} = \frac{1}{8\ Var(\rho_g)} = 35.02$$

Esse valor confere com a Tabela acima.

A $Var(\rho_g)$ dado n_{Qd} pode ser usada para computar a proporção dessa variação no parentesco, capturada pelos marcadores por meio da expressão:

$$\omega = \frac{Var(\rho_g)}{Var(\rho_g) + 0.125/n_m}$$

A quantidade $0.125/n_m$ surge em analogia a $\frac{1}{8 n_{Qd}}$, substituindo n_{Qd} por n_m. Tem-se

$$\omega = \frac{0.05^2}{0.05^2 + 0.125/n_m} \quad se\ Var(\rho_g) = 0.05^2\ (bovinos).$$

Em termos do n_{Qd}, o Me pode ser dado por:
Me = n_{Qd} Nfam = 35 Ne/2, em que Nfam é o número de famílias de irmãos completos, onde cada família tem tamanho efetivo 2.

Assim, se Ne = 50, tem-se Me = 875 e, portanto, próximo a 2Ne L. A quantidade Me = n_{Qd} x Nfam pode ser usada alternativamente na expressão de Sved para r^2_{mq}.

Para meios irmãos, em Eucalyptus (L = 13,2), tem-se:

$$Var(\rho_g) = 1/(32\ L) - 13.2/128\ L^2$$

$$Var(\rho_g) = \frac{1}{16 n_{Qd}} = 0.001785$$

$$n_{Qd} = \frac{1}{16\ Var(\rho_g)} = 35.02$$

A seguir são apresentados alguns estimadores úteis em Genética Quantitativa Genômica.

Efeitos aditivo e de dominância e sua Covariância

A covariância $Cov(\rho_g, \rho_d) = Var(\rho_g)$ só é maior que zero para n_{Qd} maior que 1000. Assim, o modelo $y = Xb + Zg + Zd + e$ deve ser ajustado com estrutura de variância $G = \begin{pmatrix} G_g\sigma_g^2 & G_g\sigma_{gd} \\ G_g\sigma_{gd} & G_d\sigma_d^2 \end{pmatrix}$ associada ao vetor $\begin{bmatrix} g \\ d \end{bmatrix}$, por meio das equações de modelo misto:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_g^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z + G_g^{-1}\frac{\sigma_e^2}{\sigma_{gd}} \\ Z'X & Z'Z + G_g^{-1}\frac{\sigma_e^2}{\sigma_{gd}} & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z\acute{y} \end{bmatrix}.$$

Herdabilidade Genômica

Esse procedimento G-BLUP permite estimar os componentes de variância:

σ_g^2: variância genética aditiva;

σ_d^2: variância genética de dominância;

σ_{gd}: covariância genética entre efeitos aditivos e de dominância;

σ_e^2: variância residual;

h^2: herdabilidade genômica no sentido restrito estimada por G-REML/G-BLUP.

Matrizes de Parentesco Genômicos

$$G_{g_{jk}} = (1/n)\sum_{i=1}^{n} \frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$$: aditivo.

$$G_{d_{jk}} = (1/n)\sum_{i=1}^{n} \frac{[s_{ij} - (2p_i(1-p_i)][s_{ik} - (2p_i(1-p_i)]}{[2p_i(1-p_i)][1-(2p_i(1-p_i))]}$$: dominância.

Coeficientes de Endogamia Genômicos

- coeficientes de endogamia por indivíduo ($F_{gjj} = G_{gjj}$-1)
- coeficiente de endogamia médio da população (media de F_{gjj})

G_{gjj}: elementos da diagonal de $G_{g\,jk}$.

- Alternativa: $\hat{F} = (1/n)\sum_{i=1}^{n} \hat{F}_i$, em que $\hat{F}_i = \{(w_i - 2p_i)^2 / [(2p_i(1-p_i)]\} - 1$ (item 6.30).

Estimação do Tamanho Efetivo (Ne) via Endogamia F

Estima-se $Ne = 1/(2F)$.

Valores genéticos genômicos dos indivíduos

Soluções para g e d nas equações de modelo misto acima

Valores genéticos genômicos das marcas

Soluções de $\hat{m} = (W'W)^{-1}W'\hat{g}$.

Acurácias na predição de valores genéticos genômicos dos indivíduos j

$r_{\hat{g}_j g_j} = [1 - (PEV_j\ \sigma_e^2)/((1 + F_{jj})\ \sigma_g^2)]^{1/2} = [1 - (d_j\ \sigma_e^2)/(G_{gjj}\ \sigma_g^2)]^{1/2}$

d_i : i-ésimo elemento da diagonal de C^{22}.

A matriz dos coeficientes das equações de modelo misto equivale a

$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{bmatrix}$ e a inversa generalizada de C é igual a

$C^{-} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$.

A partir dessa acurácia podem ser estimados:

- n_Q dado N e h^2;
- Ne dado N, n_m, h^2 e Me (função de L e Ne);
- r^2_{mq} dado N, n_m e h^2.

Por exemplo, se $r^2_{mq} = 1$, $r_{g\hat{g}} = \sqrt{1/(1 + \frac{1}{Nh^2/Me})}$ ou $r_{g\hat{g}} = \sqrt{1/(1 + \frac{Me}{Nh^2})}$

Assim, Me pode ser calculado desta equação e, posteriormente, calcula-se o Ne, substituindo Me e L na expressão $Me = \frac{2N_e\ L}{Ln(2Ne)}$.

Redução Dimensional via Seleção de Covariáveis

- G-BLUP reduzido (ou supervisionado), também denominado RR-BLUP_B (Resende et al., 2010; Resende Jr. et al., 2012)
- PCR supervisionado
- BayesCpi: seleção de marcas e GWAS

6.29 Software Selegen Genômica para GWS e GWAS

O software Selegen Genômica teve seu início em 2007 (Resende, 2007) e contempla pelo menos quatro métodos de GWS, quatro de GWAS e também o método G-REML/G-BLUP para a estimação da herdabilidade recuperada pelos marcadores, conforme a Tabela 35. Essas sete abordagens foram aplicadas a dados simulados (caráter com herdabilidade individual de 30%, controlado por 98 genes menores e dois genes menores explicando 30% da variação genética). Foram simulados 300 indivíduos e 500 marcas moleculares codominantes (Resende et al., 2011).

Tabela 35. Modelos do Selegen Genômica para a GWAS e GWS.

GWAS				
Método*	**Modelo para Efeitos de Marcas**	**Fenótipos**	**Penalização λ**	**N Marcas Selecinadas**
1 GWAS-FR-OBS	Fixo	y_c	$\lambda = 0$	95
2 GWAS-PSE-FR-EST	Fixo	$\hat{y}$	$\lambda = 0$	139
3 GWAS-PSE-RR- OBS	Aleatório	y_c	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	6
4 GWAS-PSE*-RR-EST	Aleatório	$\hat{y}$	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	97
GWS				
Método	**Modelo para Efeitos de Marcas**	**Fenótipos Corrigidos**	**Penalização λ**	**Acurácia**
1 FR-LS	Fixo	y_c	$\lambda = 0$	0.44
2 RR-BLUP	Aleatório	y_c	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	0.78
3 RR-BLUP-Het	Aleatório	y_c	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	0.80
4 G-REML/G-BLUP	Aleatório	y_c ou y	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	0.80

*** RF - Regressao fixa; RR - Regressao Aleatória; EST – Estimado; OBS – Observado; y_c - vetor de dados corrigidos; $\hat{y}$ - vetor de fenótipos estimados.**

*** PSE: Pós Simultânea Estimação.**

Para a GWAS adotou-se um nível de significância 5% pelo teste F, gerando os resultados mostrados na Tabela 35. Verifica-se que o método 3 conduz à seleção do menor número de marcas e os métodos 1 e 4 são mais equilibrados. Os métodos 2 e 4 baseiam-se em fenótipos estimados com base em efeitos de marcas estimados em conjunto. Tais métodos apresentaram maior numero de marcadores úteis, significando que determinados marcadores podem ser uteis quando em conjunto mas inúteis isoladamente. As vantagens do método GWAS-PSE-RR-EST são: estimação simultânea dos efeitos de marcas, consideração dos efeitos de marcas como aleatórios, consideração da heterogeneidade de variâncias entre marcas, estimação após validação cruzada.

Verifica-se que os métodos retiveram numero 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 (Resende et al., 2011). 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 (Resende et al., 2011). No Selegen Genômica tem-se a seguinte

correspondência entre os métodos descritos e os modelos estatísticos do software (Tabela 36).

Tabela 36. Correspondência entre os Métodos e os Modelos do Selegen Genômica para a GWAS e GWS.

GWAS			
Método*	**Modelo Estatístico do Selegen Genômica**	**Arquivo de Fenótipos**	**Arquivo de Lambdas**
1 GWAS-FR-OBS	Modelo Estatístico 4	Observados	Zeros na 2ª coluna
2 GWAS-PSE-FR-EST	Modelo Estatístico 4	Estimados	Zeros na 2ª coluna
3 GWAS-PSE-RR- OBS	Modelo Estatístico 3	Observados	Lambda de cada marca na 2ª coluna
4 GWAS-PSE-RR-EST	Modelo Estatístico 3	Estimados	Lambda de cada marca na 2ª coluna
GWS			
Método	**Modelo Estatístico do Selegen Genômica**	**Fenótipos Corrigidos**	**Arquivo de Lambdas**
1 FR-LS	Modelo Estatístico 5	Observados	-
2 RR-BLUP	Modelo Estatístico 1	Observados	-
3 RR-BLUP-Het	Modelo Estatístico 2	Observados	Lambda de cada marca na 2ª coluna
4 RR-BLUP-Het	Modelo Estatístico 6	Observados	-

*** RF - Regressao fixa. RR - Regressão Aleatória. EST – Estimado. OBS – Observado.**

O programa exige um arquivo de dados com a seguinte seqüência de colunas: *Observação Família Bloco Indivíduo Fenótipos Variáveis*. As colunas Família, Bloco e Indivíduo podem ser preenchidas com o numero 1 na versão atual. Alguns modelos exigem adicionalmente um arquivo de lambdas dados por $\lambda = \hat{\sigma}_e^2 / \hat{\sigma}_{gi}^2$, em que $\hat{\sigma}_e^2$ é a estimativa da variância residual e $\hat{\sigma}_{gi}^2$ é a estimativa da variância genética aditiva de cada loco marcador. As quantidades $\hat{\sigma}_{gi}^2$ podem ser estimadas pelos métodos IBLASSO, BLASSO, BayesA, BayesB e BayesCPi, conforme Resende et al. (2011).

O Selegen Genômica tem sido usado na UFV e na Embrapa para o desenvolvimento de teses e artigos científicos com o eucalipto (Resende et al., 2012), milho (Fritsche Neto, 2011; Fritsche Neto et al., 2012a e b), suínos (Rocha, 2011), cajueiros (Cavalcanti et al., 2012) e mandioca (Oliveira et al., 2012).

Correção dos Dados de Testes Clonais

No modelo 1 do Selegen Genômica deve ser usada a opcão BLUP e ser fornecida a herdabilidade (h^2) da característica. Essa herdabilidade deve ser aquela associada a segregação mendeliana dos dados corrigidos y_c, conforme Resende et al. (2010). Os dados podem ser corrigidos a partir da predição de valores genéticos via metodologia de modelos mistos. Para o caso de teste clonal com estrutura de família, instalado no delineamento de blocos incompletos com uma planta por parcela, o seguinte modelo pode ser analisado no software Selegen Reml/Blup: y = Xr + Zg + Wc + Sf + Tb + e, em que y é o vetor de dados, r é o vetor dos efeitos de repetição (assumidos como fixos) somados à média geral, g é o vetor dos efeitos genéticos aditivos individuais (assumidos como aleatórios), c é o vetor dos efeitos de clone dentro de família de irmãos completos (aleatórios), f é o vetor dos efeitos de dominância de família de irmãos germanos (aleatórios), b é o vetor dos efeitos de bloco (aleatórios), e é o vetor de erros ou resíduos (aleatórios). As letras maiúsculas representam as matrizes de incidência para os referidos efeitos.

Esse modelo pode ser encontrado na janela "Clones Aparentados / Matriz A Completa", modelo 169 do software Selegen Reml/Blup. Deve ser fornecido um arquivo de dados com a seguinte seqüência de colunas: *Observação Indivíduo Repetição*

Clone Família Bloco Árvore Variáveis. Deve ser fornecido também um arquivo de pedigree com a seguinte seqüência de colunas: *Indivíduo Pai Mãe*. Todos os indivíduos devem constar na primeira coluna mesmo os que são apenas genitores. Genitores de genitores devem receber códigos zero na segunda e terceira colunas.

Os componentes de variância associados ao modelo, conforme notação do Selegen-Reml/Blup, são assim interpretados:

Va: variância genética aditiva.

Vclone/fam: variância entre clones dentro de famílias de irmãos completos, ajustada para variância genética aditiva total.

Vfam: variância da capacidade específica de combinação ou variância genética de dominância entre famílias de irmãos germanos.

Vbloc: variância entre blocos.

Ve: variância ambiental.

Vf: variância fenotípica individual.

h2a = h2: herdabilidade individual no sentido restrito no bloco, ou seja, dos efeitos aditivos.

h2g: herdabilidade individual no sentido amplo, ou seja, dos efeitos genotípicos totais.

c2clone/fam = c2: coeficiente de determinação dos efeitos de clones dentro de famílias de irmãos completos, ajustado para variância genética aditiva total.

c2fam = c21: coeficiente de determinação dos efeitos da capacidade específica de combinação.

c2bloc = c22: coeficiente de determinação dos efeitos de bloco.

u = Média geral do experimento.

Os componentes de variância apresentados acima podem ser decompostos da seguinte forma:

$Va = \sigma_g^2$: variância genética aditiva.

$Vfam = (1/4)\sigma_d^2 + (1/4)\sigma_{gg}^2 + (1/8)\sigma_{gd}^2 + (1/16)\sigma_{dd}^2 ...$

$Vclone/fam = (3/4)\sigma_d^2 + (3/4)\sigma_{gg}^2 + (7/8)\sigma_{gd}^2 + (15/16)\sigma_{dd}^2 ...$, em que d refere-se a efeitos de dominância e gg, gd e dd referem-se a efeitos epistáticos.

A herdabilidade da segregação mendeliana desregressada dos dados corrigidos y_c é dada por: $h_m^2 = \dfrac{n\ 0.5h^2}{n\ 0.5h^2 + 1 - h^2 - c_{clone/fam}^2 - c_{fam}^2 - c_{bloc}^2}$, em que n é o número de repetições do clone. Expressa de outra forma tem-se: $h_m^2 = \dfrac{n\ 0.5\sigma_g^2}{n\ 0.5\sigma_g^2 + \sigma_e^2}$.

Sob esse modelo o valor genotípico dos clones é dado por $VG = \hat{u} + \hat{g} + \hat{c} + \hat{f} = \hat{u} + \hat{g} + \hat{c}lone/fam + \hat{f}am$.

Nível de Significância na GWAS

Em problemas onde a inferência probabilística exata não está disponível, a função de verossimilhança observada pode ser usada diretamente para inferência. Isto pode ser feito por meio da razão de riscos (*odds ratio*), a qual é a própria razão direta entre os valores da função maximizada por dois conjuntos distintos de valores paramétricos a serem avaliados, ou seja, OD = $(L(U))/L(V))$. A inferência verossimilhança pura pode ser usada quando a teoria de grandes amostras não for adequada ao caso analisado. Esse é o caso de amostras pequenas com distribuição não normal.

Uma derivação do OD, muito usada no contexto da genética é o teste do LOD escore. LOD significa "*log of odds ratio*", ou seja, logaritmo na base 10 da razão de riscos (odds ratio). Riscos, no caso, quantificados pela verossimilhança de dois modelos a serem comparados. O LOD é dado por $LOD = Log_{10}\ OD = Log_{10}\ (L(U))/L(V)) = \lambda / [2\ Log\ (10)] = \lambda / 4.61$. Portanto, existe uma relação direta entre o LOD e o LRT ou λ, ou seja, $LOD = LRT / 4.61$. Alternativamente, $LRT = 4.61\ LOD$.

Com base nessa última expressão, pode-se associar valores de LOD e p–valores aproximados do LRT. Os valores críticos (λ) de qui-quadrado nos níveis de significância 10%, 5%, 1% e 0.5% são 2.71, 3.84, 6.64 e 7.88, respectivamente. Esses valores estão associados aos seguintes LOD´s, dados por $LOD = LRT / 4.61$: 0.588, 0.833, 1.440 e 1.709, respectivamente. Assim, uma inferência aproximada é de que LOD´s maiores que 1.71 já estão associados a elevados (menores do que 0.5 %) níveis de significância. Um LOD escore de 3 significa que uma hipótese é mil vezes mais plausível que a outra. Neste caso, a inferência é baseada apenas na razão de verossimilhança, sem invocar as propriedades distribucionais dos estimadores de máxima verossimilhança. As relações aproximadas entre LOD e significância pelo LRT são apresentadas na Tabela 37.

Tabela 37. Relações aproximadas entre LOD e significância pelo LRT.

LOD*	Número de vezes em que H_1 é mais provável do que H_0	LRT	Significância
0.588	3.87	2.71	10.00%
0.833	**6.81**	**3.84**	**5.00%**
1	10.00	**4.61**	3.17%
1.09	12.27	5.02	2.50%
1.44	27.54	6.64	1.00%
1.71	51.29	7.88	0.50%
2	100.00	9.22	0.23%
3	**1000.00**	**13.83**	**0.02%**

H_0: hipótese de ausência de ligação marcador – QTL; H_1: hipótese de presença de ligação marcador – QTL; * *Potência de 10 cujo resultado indica quantas vezes H_1 é mais provável do que H_0.

O nível de significância adotado pelo Selegen é 5% e parece adequado para a GWS mas não para a GWAS. Nesse caso, valores maiores de F devem ser procurados como ponto de corte nos resultados emitidos pelo Selegen, visando adotar significâncias da ordem de menos que 1% para a GWAS, geralmente 10^{-5}. Em termos de LRT, o valor de corte muda de 3.84 (equivalente ao F para grande número de graus de liberdade do resíduo) para 13.83 visando alterar a significância de 5% para 0.02% (Tabela 37).

Resultados Gerados pelo Selegen Genômica

Catálogo de estimativas dos efeitos genéticos de marcas ($\tilde{m}$) dados pela resolução dos sistemas:

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + I\sigma_e^2/\sigma_m^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}$$ (RR-BLUP ou SNP-BLUP)

ou

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + D_{\tau_i^2}^{-1}\sigma_e^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}$$ (RR-BLUP-Het), em que τ_i^2 são variâncias específicas para cada loco e presentes na diagonal da matriz $D_{\tau_i^2}$.

Catálogo de estimativas dos efeitos genéticos genômicos de indivíduos ($\tilde{g}$) dados pela resolução dos sistemas:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$ (G-BLUP)

ou

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{*-1}\sigma_e^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$ (G-BLUP-Het)

, em que $G = WW'/n_Q$ e $G^* = WD_{\tau_i^2}W'/n_Q$.

Estes são similares ao tradicional sistema

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\sigma_e^2/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$ (A-BLUP).

O método G-BLUP-Het produz resultados similares aos dos métodos BayesA e IBLASSO.

O vetor de dados fenotípicos a ser fornecido ao Selegen deve ser: (i) apenas corrigidos para os efeitos fixos (ou valores genéticos desregressados) caso o interesse seja no catálogo de estimativas dos efeitos genéticos genômicos dos indivíduos; (ii) corrigidos para os efeitos fixos (ou valores genéticos desregressados) e para os efeitos genéticos de genitores (estrutura de população) caso o interesse seja no catálogo de estimativas dos efeitos genéticos de marcas visando uso em outros indivíduos da população nas gerações atual e/ou subsequentes.

6.30 Software GCTA para G-REML em Genética Humana e Animal

Uma forma muito usada para a estimação da h^2 em humanos é via análise de ligação usando toda a genealogia (Almasy e Blangero, 1998). O software Solar (*Sequential Oligogenic Linkage Analysis Routines*) tem sido usado para essa estimação.

Atualmente isso está mudando e a h^2 tem sido estimada via análise de desequilíbrio de ligação. O método de Yang et al. (2010) e Visscher et al. (2010) tem sido usado por meio de sua implementação no software GCTA (Genome-wide Complex Trait Analysis) desenvolvido pelos mesmos autores (Yang et al., 2011). O referido software implementa o método REML genômico (G-REML ou G-REML/G-BLUP) usando a matriz de parentesco genômico. Esse aplicativo estima a variância genética aditiva total, a variância da interação efeitos aditivos por ambiente e a herdabilidade total capturada pelos marcadores.

Uma vez que a matriz G genômica é tipicamente densa, as técnicas computacionais que usam matrizes esparsas apresentam um custo computacional extra. Assim, com grande número de SNPs (> 10.000) o software ASREML é mais lento (demanda vários dias) do que o GCTA (demada poucas horas). Número de SNPs superior a 600.000 tem sido analisados com o GCTA.

Para corrigir para estrutura populacional ou de famílias, o programa apresenta uma função que exclui iterativamente um indivíduo de todo par que apresenta parentesco maior que 0.025 enquanto mantem o número máximo possível de indivíduos no conjunto de dados. Ou seja, esta opção é usada para excluir indivíduos muito aparentados. A não realização dessa correção conduz à captura da contribuição de todos os variantes causais e não apenas daqueles em desequilíbrio de ligação com os SNPs.

A correção para estrutura de população é importante em GWAS, pois o objetivo é detectar variantes causais. É importante também em GWS aplicada a distantes gerações futuras, sem a re-estimação dos efeitos das marcas. Para uso de um catálogo de valores genéticos de marcas em novos indivíduos na mesma geração ou na geração subsequente, a correção para estrutura de população não é tão necessária, visto que, nesse caso, o parentesco também é capitalizado na seleção.

A predição genômica de indivíduos aparentados é baseada mais em ligação do que em desequilíbrio de ligação. Por outro lado, a predição de indivíduos geneticamente distantes requer LD entre marcadores e QTL. Ligação é também uma forma de LD, mas LD fundamenta-se em consistência da fase de ligação entre marca e QTL. Essa consistência é demandada ao menos em toda a população e, possivelmente, mesmo entre populações (Daetwyler et al., 2012). Em vez de concentrar a predição genômica apenas no LD entre marca e QTL, na prática algumas vezes deve-se utilizar ambos ligação (capturada pelo parentesco na genealogia) e LD visando maximizar a acurácia da GWS.

Uma maneira de verificar se a acurácia da predição genômica é predominantemente devida ao parentesco ou ao desequilíbrio de ligação, refere-se ao

ajuste de milhares de marcadores presentes em apenas um cromossomo. Se esse ajuste propiciar a quase totalidade da acurácia conseguida com o ajuste de todos os SNPs em todos os cromossomos, isto significa que a acurácia da predição genômica é predominantemente devida ao parentesco. Explica-se isso pelo fato de que um cromossomo com muitos marcadores podem capturar bem as parentescos mas, nunca abrange todos os QTL. Meuwissen (2009) relata que o número de SNPs para a predição de indivíduos não aparentados é dado por 10 Ne L. Para eucalipto (L = 13 Morgans), em uma população com tamanho efetivo 100, seriam necessários 13000 SNPs.

O GCTA propicia também os autovetores da matriz G (similar a análise de componentes principais). Esses autovetores (componentes principais) associados aos maiores autovalores podem então ser incluídos no modelo como covariáveis de efeitos fixos, visando capturar a variância devida à estrutura de população. É importante notar que G é uma matriz N x N, referente aos N indivíduos e não aos n locos ou covariáveis. Assim, os autovetores associados aos maiores autovalores informam sobre os indivíduos que dominam (possuem maior parentesco com os demais) as relações de parentesco e agrupam os indivíduos em subgrupos estruturados. Dessa forma, o ajuste dos autovetores principais de G como covariáveis fornece uma correção para essa estruturação.

Geralmente, para milhares de marcadores (acima de 20.000), 10 a 200 autovetores (explicando no máximo 70% da variação total dos coeficientes de parentesco) são ajustados como efeitos fixos. Os modelos utilizados podem ser:

$$y = Xb + \sum_{i=1}^{v} U_i \alpha_i + Wm + e \text{ (RR-BLUP);}$$

$$y = Xb + \sum_{i=1}^{v} U_i \alpha_i + Zg + e \text{ ou } y = Xb + \sum_{i=1}^{v} U_i \alpha_i + Z\,Wm + e \text{ (G-BLUP),}$$

em que v é o número (tipicamente 10 ou 20) de autovetores (U_i), com os maiores autovalores, ajustados como efeitos fixos e, α_i são os coeficientes de regressão.

Os coeficientes α_i são ajustados sem shrinkage por OLS, a partir da matriz G. Janss et al. (2012) criticam essa abordagem, relatando que a mesma sofre do problema de contagem dupla, pois os autovetores (estimados a partir de G e genéricos para qualquer caráter) são incluídos duas vezes na análise: como efeitos fixos e, implicitamente, como efeitos aleatórios na parte aleatória do modelo (matriz G). Propõem então, uma nova abordagem para o problema, via estimação simultânea baseada em decomposição espectral, a qual propicia inferência simultânea sobre efeitos de marcas específicos para cada caráter e a latente (não observada) estrutura de população. Decomposições adequadas de matrizes de covariância do tipo da G são apresentadas por Ledoit e Wolf (2004) e Schafer e Strimmer (2005).

A expressão $G = (WW')/n$, em que W contém elementos dados por $w_{ij_p} = \dfrac{(w_{ij} - 2p_i)}{[2p_i(1-p_i)]^{1/2}}$ e G contém elementos dados por

$G_{jk} = (1/n)\sum_{i=1}^{n}\frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$, ignora o erro de amostragem ($Var(G_{ijk})$ e $Var(G_{ijj})$) associado a cada SNP. Para o SNP i nos indivíduos j e k, $j \neq k$, tem-se $Var(G_{ijk}) = \frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{4p_i^2(1-p_i)^2} = 1$, de forma que essa variância é a mesma para todos os SNPs independentemente de suas frequências e não há necessidade de correção em G_{ijk}. Entretanto, para $j = k$ tem-se $Var(G_{ijj}) = \frac{Var[(w_{ij} - 2p_i)^2]}{4p_i^2(1-p_i)^2} = \frac{1 - 2p_i(1-p_i)]}{2p_i(1-p_i)}$ e depende da frequência do alelo do SNP. Isto conduz ao uso de $G_{ijj} = 1 + [w_{ij}^2 - (1+2p_i)w_{ij} + 2p_i^2]/[2p_i(1-p_i)]$, o qual é um estimador não viesado de 1 + F, conforme mostrado abaixo. Se F = 0, $Var(G_{ijj}) = 1$ e não há necessidade de correção em G_{ijj}. Assim, considerando todos os SNPs tem-se $G_{jk} = (1/n)\sum_{i=1}^{n}\frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$ para $j \neq k$ e $G_{jk} = 1 + (1/n)\sum_{i=1}^{n}\frac{[w_{ij}^2 - (1+2p_i)w_{ij} + 2p_i^2]}{2p_i(1-p_i)}$ para $j = k$.

Meuwissen et al. (2011) não recomenda essa expressão para G_{jj}. Essa formulação de G_{jj} como diferente de G_{jk} é realizada visando minimizar a variação de amostragem. Mas isso pode conduzir a autovalores negativos na matriz G. Se G é toda estimada via $G = (WW')/n$ ela é semi-positiva definida e acrescentando um pequeno valor positivo aos elementos da diagonal (fazendo $G = (WW')/n + I * 10^{-4}$) torna-a positiva definida.

As frequências genotípicas associadas aos três genótipos de um SNP em uma espécie diploide é dada por $p_i^2 + p_i(1-p_i)F$; $[2p_i(1-p_i)](1-F)$ e $(1-p_i^2) + p_i(1-p_i)F$, para MM, Mm e mm, respectivamente. Assim, a estrutura populacional é dada por $[p_i^2 + p_i(1-p_i)F]MM$ + $\{[2p_i(1-p_i)](1-F)\}Mm$ + $[(1-p_i^2) + p_i(1-p_i)F]mm$.

O coeficiente de endogamia (F) multilocos estimado de todos os marcadores é dado por $\hat{F} = (1/n)\sum_{i=1}^{n}\hat{F}_i$, ou seja, pela média das estimativas através de todos os SNPs, em que $\hat{F}_i = [w_i^2 - (1+2p_i)w_i + 2p_i^2]/[2p_i(1-p_i)]$ e w_i é número de cópias do alelo de referência para o SNP i. Este estimador é o usado no GCTA. Isto significa que, no loco i, os elementos da diagonal da matriz G são dados por $G_{jj} = 1 + \hat{F}_i = 1 + [w_{ij}^2 - (1+2p_i)w_{ij} + 2p_i^2]/[2p_i(1-p_i)]$. Na matriz G o parentesco não é uma probabilidade (conforme definição clássica de IBD) mas sim uma correlação entre valores genéticos aditivos. Outro estimador de $\hat{F}_i$ é dado por $\hat{F}_i = \{(w_i - 2p_i)^2/[(2p_i(1-p_i)]\} - 1$.

A estimação da herdabilidade fundamenta-se nos parentescos ao nível dos variantes causais ou QTLs. Mas esses parentescos são estimados, via SNPs, com erros devidos ao desequilíbrio imperfeito. E o erro na predição é dado por $c+(1/n)$, em que c depende da distribuição da MAF dos variantes causais. Os autores desenvolveram um método baseado em regressão para corrigir para esse erro de predição. Nesse caso, os elementos da matriz G_m que representam o parentesco realizado médio multi-locos são dados por $G_{jk}^{*}=\begin{cases}1+\beta(G_{jk}-1), j=k\\ \beta G_{jk}, j\neq k\end{cases}$, em que

$G_{jk}=(1/n)\sum_{i=1}^{n}\frac{(w_{ij}-2p_i)(w_{ik}-2p_i)}{2p_i(1-p_i)}$ para $j\neq k$ e

$G_{jk}=1+(1/n)\sum_{i=1}^{n}\frac{[w_{ij}^{2}-(1+2p_i)w_{ij}+2p_i^{2}]}{2p_i(1-p_i)}$ para $j=k$;

$\beta=1-\frac{1/n}{c+(1/n)}=\frac{c+(1/n)}{c+(1/n)}-\frac{1/n}{c+(1/n)}=\frac{c}{c+(1/n)}$.

Expressão similar foi apresentada por Van Raden (2008).

Após esse ajuste, a estimativa proporcional da variância aditiva explicada por todos os marcadores é um estimador não viesado da herdabilidade se a suposição sobre a distribuição da MAF dos variantes causais for correta. A quantidade c é computada como $c=Cov(G_{jk},Q_{jk})$, em que Q_{jk} é a matriz de parentesco real ao nível dos QTLs ou variantes causais, formada somente com os SNPs de baixa MAF, os quais mimicam os referidos QTLs. Intrinsecamente tem-se $\beta=\frac{Cov(G_{jk},Q_{jk})}{Var(G_{jk})}$.

O termo $(1/n)$ advem do fato de se estimar G usando somente n SNPs. Isto corresponde ao erro de amostragem $Var(G_{ijk})$ igual a 1 calculado para G_{ijk} em um só SNP.

A expressão $G_{jk}^{*}=\beta G_{jk}, j\neq k$ difere de $G_{jk}^{*}=A_{jk}+\beta(G_{jk}-A_{jk}), j\neq k$ apresentada anteriormente. Isso se deve ao fato de ter-se aplicado a correção prévia para estrutura populacional ou de famílias. Isto culmina com valores zero nos elementos fora da diagonal de A.

6.31 Variação Epigenética e Covariância entre Parentes

Variação Epigenética refere-se à todas as mudanças reversíveis e herdáveis no genoma funcional que não alteram a seqüência de nucleotídeos do DNA. Existem três mecanismos principais de alterações epigenéticas: metilação do DNA, modificações de histonas e ação de RNAs não codificadores. Os padrões de metilação de DNA são os mais importantes.

A metilação afeta a concretização da matriz W e dados relativos à probabilidade de metilação em porções específicas do DNA já estão disponíveis para a análise genética em conjunto com dados fenotípicos, genealógicos e de marcadores genéticos. Algumas definições importantes são apresentadas a seguir.

Herança Epigenética: transmissão de variação fenotípica entre gerações não por meio da variação de sequências de DNA.

Transmissibilidade epigenética: probabilidade de transmissão de fenótipos ancestrais.

Coeficiente de Reset ou Reversão (ν): probabilidade de mudança de estado epigenético durante a gametogênese ou fase de desenvolvimento inicial.

Coeficientes de Transmissibilidade Epigenética $(1-\nu)$: é o complemento do coeficiente de reset, retorno ou reversão.

Ambiente indutor: sinal ambiental ou agente de estresse que causa a mudança de estado epigenético.

(a) Covariância entre Parentes para Sistemas de Reprodução Sexuada (Tal et al., 2010)

Modelo Fenotípico em Presença de Variação Epigenética

$y = Xb + Zg + Z\xi + e$

$\sigma_y^2 = \sigma_g^2 + \sigma_\xi^2 + \sigma_e^2$: variância fenotípica total.

Covariância entre Parentes com Variação epigenética

$$COV(P,F) = (1/2)\sigma_a^2 + (1/2)(1-\nu)\sigma_\xi^2$$
$$COV(MI) = (1/4)\sigma_a^2 + (1/4)(1-\nu)^2\sigma_\xi^2$$
$$COV(TSC) = (1/4)\sigma_a^2 + (1/4)(1-\nu)^3\sigma_\xi^2$$

Verifica-se que a variação epigenética inflaciona as covariâncias genéticas entre parentes.

Estimação dos Componentes de Variância

$$(1-\nu) = \frac{2[COV(MI) - COV(TS)]}{COV(P,F) - 2COV(MI)}$$

$$\sigma^2_{\xi} = \frac{2[COV(P,F) - 2COV(MI)]}{\nu(1-\nu)}$$

$$\sigma^2_g = 2COV(P,F) - (1-\nu)\sigma^2_{\xi}$$

Herdabilidade Epigenética

$$h^2_{\xi} = \frac{\sigma^2_{\xi}}{\sigma^2_y}$$

O modelo pode ser ajustado por meio das equações de modelo misto:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + A^{-1}\frac{\sigma^2_e}{\sigma^2_g} & Z'Z \\ Z'X & Z'Z & Z'Z + \Lambda^{-1}\frac{\sigma^2_e}{\sigma^2_{\xi}} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'\acute{y} \end{bmatrix}.$$

Esse procedimento REML/BLUP permite estimar os componentes de variância:

σ^2_g: variância genética aditiva;

σ^2_{ξ}: variância epigenética;

σ^2_e: variância residual;

h^2_{ξ}: herdabilidade epigenética;

A: matriz de correlação genética aditiva entre indivíduos;

Λ : matriz de transmissibilidade epigenética.

(b) Covariância entre Parentes para Sistemas de Reprodução Assexuada (Tal et al., 2010)

$\sigma^2_y = \sigma^2_{gt} + \sigma^2_{\xi} + \sigma^2_e$: variância fenotípica total.

Covariância entre parentes

$$COV(P,F) = \sigma^2_{gt} + (1-\nu)\sigma^2_{\xi}$$

$$COV(RAM) = \sigma^2_{gt} + (1-\nu)^2\sigma^2_{\xi}$$

$$COV(TSC) = \sigma^2_{gt} + (1-\nu)^3\sigma^2_{\xi}$$

Estimação dos Componentes de Variância

$$(1-\nu) = \frac{COV(RAM) - COV(TSC)}{COV(P,F) - COV(RAM)}$$

$$\sigma^2_{\xi} = \frac{COV(P,F) - COV(RAM)}{\nu(1-\nu)}$$

$$\sigma^2_{gt} = COV(P,F) - (1-\nu)\sigma^2_{\xi}$$

σ^2_{gt}: variância genotípica total.

7 Scripts em R para Modelos Mistos, Inferência Bayesiana e Seleção Genômica

7.1 R para Modelos Mistos

Método BLUP no Pacote Pedigreemm (escrito por Inez Vasquez, University of Wisconsin)

Modelo Animal com Efeito Aleatório Adicional (Parcela)

```
# carregamento do pacote pedigreemm
library(pedigreemm)

# leitura dos dados BLUP-INDIVIDUO
setwd("C:\\R 2013")
dados=read.table("modelo1Ru3.txt",h=T)


# listagem dos 6 dados iniciais e finais
head(dados)
tail(dados)

# início do arquivo de dados

id sire Dam  Bloc Parc Alt
1   NA  NA    1 999999 0.0
2   NA  NA    2 999999 0.0
3   NA  NA    3 999999 0.0
4   NA  NA    4 999999 0.0
5   NA  NA    5 999999 0.0
6    1  NA    6      1 2.6

#definição de blocos como efeito fixo
dados1=data.frame(dados,Blo=factor(dados$Bloc))


# montagem do pedigree

pedCows=read.table("modelo1Ru3.txt",h=T)

sire = pedCows[,2]
dam = pedCows[,3]
id = pedCows[,1]

pedCows = pedigree(sire=as.integer(sire),dam=as.integer(dam),label=as.character(id))


# Modelo para predição de efeitos genéticos de indivíduos
ajuste= pedigreemm(Alt ~ -1 + Blo + (1|id) + (1|Parc), data = dados1, pedigree = list(id
= pedCows))
summary(ajuste)


# predições dos efeitos de parcela
p=ranef(ajuste)$Parc
p

# predições dos efeitos genéticos de indivíduos
i=ranef(ajuste)$id
i

# definições do cabeçalho

id = indivíduo
sire = pai
dam = mãe
Bloc = bloco
```

Parc = parcela
Alt = variável

comentários

No pedigreemm deve-se duplicar a ultima linha do arquivo e atribuir um novo código (por exemplo, 999999)para bloco e parcela;
O pedigreemm só inclui na análise genitores com dados próprios. Assim, no caso de um teste de progênies e usando apenas dados do teste, os genitores devem ser incluídos no arquivo de dados com dados (y) fictícios e posteriormente os resultados devem ser corrigidos. Os valores genéticos preditos dos genitores corrigidos são dados por $\hat{a}_{gen-correto} = (y - \hat{B}loc)$, em que $\hat{B}loc$ referem-se aos efeitos estimados de blocos (os quais devem receber im código diferente para cada genitor (ver início do arquivo de exemplo).

Para correção dos valores genéticos preditos dos indivíduos, os autores derivaram a seguinte correção, válida para progênies de meios irmãos em delineamento com várias plantas por parcela:

$$\hat{a}_i = \hat{a}_{iR-Pedigreemm} + \hat{a}_{gen-incorreto}\left(\frac{3h^2}{16-4h^2-16c^2}\right) + \hat{a}_{gen-correto}\left(\frac{2-2c^2-2h^2}{4-h^2-4c^2}\right)$$

, em que:

$\hat{a}_i$: vetor de valores genéticos corretos preditos dos indivíduos;

$\hat{a}_{iR-Pedigreemm}$: vetor de valores genéticos incorretos dos indivíduos preditos pelo pedigreemm;

$\hat{a}_{gen-incorreto}$: vetor de valores genéticos incorretos dos genitores preditos pelo pedigreemm;

$\hat{a}_{gen-correto}$: vetor de valores genéticos corretos dos genitores obtidos por $\hat{a}_{gen-correto} = (y - \hat{B}loc)$;

h^2 : herdabilidade individual no sentido restrito;

c^2 : coeficiente de determinação dos efeitos de parcela.

Para progênies de meios irmãos em delineamento com uma planta por parcela, a correção é dada por:

$$\hat{a}_i = \hat{a}_{iR} + \hat{a}_{gen-incorreto}\frac{(3/4)h^2}{(4-h^2)} + \hat{a}_{gen-correto}\frac{(1/2)(1-h^2)}{(1/4)(4-h^2)}$$

Para progênies de irmãos germanos em delineamento com várias plantas por parcela, a correção é dada por:

$$\hat{a}_i = \hat{a}_{iR} + (a_{sire-nao-corrigido} + a_{dam-nao-corrigido})\frac{h^2}{8-4h^2-8c^2} + (a_{sire-corrigido} + a_{dam-corrigido})\left(\frac{1-c^2-h^2}{2-h^2-2c^2}\right)$$

Para progênies de irmãos germanos em delineamento com uma planta por parcela, a correção é dada por:

$$\hat{a}_i = \hat{a}_{iR} + (a_{sire-nao-corrigido} + a_{dam-nao-corrigido})\frac{(1/4)h^2}{(2-h^2)} + (a_{sire-corrigido} + a_{dam-corrigido})\frac{(1-h^2)}{(2-h^2)}$$

7.2 R para Inferência Bayesiana

Função *rhierLinearModel* no Pacote *Bayesm*: (escrito por Peter Rossi, University of Califórnia)

```
setwd("C:\\R 2013")

library(bayesm)

y=as.matrix(read.table("fen_final.txt"))
iota=rep(1,nrow(y))
x=cbind(iota,as.matrix(read.table("touros_final.txt")))

regdata=NULL
for(reg in 1:1)
{
regdata[[reg]]=list(y=y,X=x)
}

Data1=list(regdata=regdata)
Prior1=list(nu=52,V=diag(32,ncol(x)))
Mcmc1=list(R=20000,keep=4)

fit=rhierLinearModel(Data=Data1, Prior=Prior1,Mcmc=Mcmc1)

var=diag(matrix(apply(fit$Vbetadraw,2,mean),ncol(x),ncol(x))) #[-1]
var
dim(fit$Vbetadraw)
beta=apply(fit$betadraw,2,mean)
beta
vare=mean(fit$taudraw)
vare


# comentários
```

O vetor beta fornece os valores genéticos preditos dos genitores ($\hat{g}_{pi}$) e VBeta fornece as estimativas de variâncias genéticas aditivas ($\sigma^2_{gVBetai}$) associadas à segregação de cada genitor. O escalar vare fornece a estimativa da variância residual comum a todos os indivíduos.

As estimativas dos valores genéticos preditos dos indivíduos não genitores é dada na página 80. Toda essa abordagem produz estimativas de componentes de variância e valores genéticos pelo método Blup Bayesiano Melhorado (I-BAYES-BLUP ou BBM) proposto por Resende, Silva e Viana (2012).

Para modelos com mais fatores de efeitos aleatórios pode-se rodar o REML/BLUP tradicional no Selegen, obter os componentes de variância e coeficientes de determinação (c2), fixá-los nesses valores e fixar h2 em 1 e rodar o BLUP. Esses BLUPs obtidos com h2 igual a 1 são então submetidos ao sript do Bayesm acima.

7.3 R para Seleção Genômica

7.3.1 Método BayesA: escrito por Rohan Fernando (University of Iowa)

```
#BayesA
setwd("C:\\R 2013")
# Parameters
nmarkers = 2000;   #number of markers
numiter  = 200;   #number of iterations
vara     = 1.0/20.0;

# input training data
data     = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data)[1];
startMarker = 1800;
x = cbind(1,data[,startMarker:nmarkers]);  #this is the mean and then the markers
y = data[,nmarkers+1];

# inital values

nmarkers = nmarkers - startMarker + 1;
mean2pq = 0.5;                                          # just an approximation
scalea  = 0.5*vara/(nmarkers*mean2pq);                  # 0.5 = (v-2)/v for v=4

size = dim(x)[2];
b = array(0.0,size);
meanb = b;
b[1] = mean(y);
var  = array(0.0,size);

# adjust y
 ycorr = y - x%*%b;

# mcmc sampling
for (iter in 1:numiter){

# sample vare
       vare = ( t(ycorr)%*%ycorr )/rchisq(1,nrecords + 3);

# sample intercept
       ycorr = ycorr + x[,1]*b[1];
       rhs    = sum(ycorr)/vare;
       invLhs = 1.0/(nrecords/vare);
       mean = rhs*invLhs;
       b[1] = rnorm(1,mean,sqrt(invLhs));
       ycorr = ycorr - x[,1]*b[1];
       meanb[1] = meanb[1] + b[1];

# sample variance for each locus

       for (locus in 2:size){
              var[locus] = (scalea*4+b[locus]*b[locus])/rchisq(1,4.0+1)
       }
# sample effect for each locus
       for (locus in 2:size){
              ycorr = ycorr + x[,locus]*b[locus];   #unadjust y for this locus
              rhs = t(x[,locus])%*%ycorr/vare;
              lhs = t(x[,locus])%*%x[,locus]/vare + 1.0/var[locus];
              invLhs = 1.0/lhs;
              mean = invLhs*rhs;
              b[locus]= rnorm(1,mean,sqrt(invLhs));
              ycorr = ycorr - x[,locus]*b[locus];   #adjust y for the new value of this
locus
              meanb[locus] = meanb[locus] + b[locus];
       }

}

meanb = meanb/numiter;
plot(meanb)
gebv=x%*%meanb
plot(hist(gebv))
```

```
#test population
nmarkers=2000
data1     = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data1)[1];
x = cbind(1,data1[,startMarker:nmarkers]);
yt = data1[,nmarkers+1];
yHat_t   = x %*% meanb;
corr = cor(yt,yHat_t);
corr
```

7.3.2 Método BayesB: escrito por Rohan Fernando (University of Iowa)

```
#BayesB
setwd("C:\\R 2013")
# Parameters
nmarkers    = 2000;   #number of markers
numiter     = 200;   #number of iterations
numMHIter   = 200;    #use 1 for Bayes A
pi          = 0.95;   #Change this to run Bayes B rather than Bayes A
vara        = 1.0;

# input training data
data     = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data)[1];
startMarker = 1800;
x = cbind(1,data[,startMarker:nmarkers]);
y = data[,nmarkers+1];
a =  data[,nmarkers+2];

# inital values
nmarkers = nmarkers - startMarker + 1;
mean2pq     = 0.5;
scaleb  = 0.5*vara/(nmarkers*(1-pi)*mean2pq);

b = array(0.0,nmarkers+1);
meanb = b;
b[1] = mean(y);
var  = array(0.0,nmarkers);
ppa  = array(0.0,nmarkers);

# adjust y
ycorr = y - x%*%b;

# mcmc sampling
for (iter in 1:numiter){

# sample vare
       vare = ( t(ycorr)%*%ycorr )/rchisq(1,nrecords + 3);

# sample intercept
       ycorr = ycorr + x[,1]*b[1];
       rhs    = sum(ycorr)/vare;
       invLhs = 1.0/(nrecords/vare);
       mean = rhs*invLhs;
       b[1] = rnorm(1,mean,sqrt(invLhs));
       ycorr = ycorr - x[,1]*b[1];
       meanb[1] = meanb[1] + b[1];

# sample variance and effect for each locus
    nLoci = 0;
       for (locus in 1:nmarkers){
              ycorr = ycorr + x[,1+locus]*b[1+locus];
              rhs = t(x[,1+locus])%*%ycorr;
              totalSS = sum(ycorr^2)/vare;
              xpx = t(x[,1+locus])%*%x[,1+locus];
              v1  =  (xpx^2*var[locus] + xpx*vare);          # slide 47
              v2  =  xpx*vare;
              logDataNullModel = -0.5*(log(v2) + rhs^2/v2);   # slide 47
              if (var[locus] > 0.0){
                     logDataOld   = -0.5*(log(v1) + rhs^2/v1);
              }
              else {
```

```
                        logDataOld = logDataNullModel;
                }
                for (mhiter in 1:numMHIter){
                        u = runif(1);
                        varCandidate = 0;
                        if (u > pi){
                                varCandidate = scaleb*4/rchisq(1,4);
                        }
                        if (varCandidate > 0.0){
                                v1  =  (xpx^2*varCandidate + xpx*vare);
                                logDataNew =  -0.5*(log(v1) + rhs^2/v1);
                        }
                        else{
                                logDataNew = logDataNullModel;
                        }
                        acceptProb = exp(logDataNew-logDataOld); # slide 45
                        u = runif(1);
                        if(u <acceptProb) {
                                var[locus] = varCandidate;
                                logDataOld = logDataNew;
                        }
                }
                if(var[locus]) {
                        nLoci = nLoci + 1;
                        lhs = xpx/vare + 1.0/var[locus];
                        invLhs = 1.0/lhs;
                        mean = invLhs*rhs/vare;
                        b[1+locus]= rnorm(1,mean,sqrt(invLhs));
                        ycorr = ycorr - x[,1+locus]*b[1+locus];
                        meanb[1+locus] = meanb[1+locus] + b[1+locus];
                        ppa[locus] = ppa[locus] + 1;
                }
                else b[1+locus] = 0.0;
        }
}

meanb = meanb/numiter;
plot(meanb)
gebv=x%*%meanb
plot(hist(gebv))

#test population
nmarkers=2000
data1     = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data1)[1];
x = cbind(1,data1[,startMarker:nmarkers]);
yt = data1[,nmarkers+1];
yHat_t   = x %*% meanb;
corr = cor(yt,yHat_t);
corr
```

7.3.3 Método BayesCPi: escrito por Rohan Fernando (University of Iowa)

```
#BayesCPi
setwd("C:\\R 2013")
# Parameters
nmarkers    = 2000;    #number of markers
numiter     = 200;
pi          = 0.5;
vara        = 1.0;
logPi     = log(pi);
logPiComp = log(1-pi);
mean2pq     = 0.5;
nua       = 4;

# input training data
data     = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data)[1];
startMarker = 1800;
#startMarker = 1;
x = cbind(1,data[,startMarker:nmarkers]);
y = data[,nmarkers+1];
a =  data[,nmarkers+2];
```

```
storePi = array(0.0,numiter);
# inital values
nmarkers = nmarkers - startMarker + 1;
varEffects  = vara/(nmarkers*(1-pi)*mean2pq);
scalec      = varEffects*(nua-2)/nua;

cat ("scale : ",scalec);

# Hyper parameters of Scale factor
theta = 1;
beta  = 1;
meanscalec = 0;

b = array(0.0,nmarkers+1);
meanb   = b;
b[1]   = mean(y);
var    = array(0.0,nmarkers);
ppa    = array(0.0,nmarkers);
piMean = 0.0;

# adjust y
ycorr = y - x%*%b;

# mcmc sampling
for (iter in 1:numiter){

# sample vare
       vare = ( t(ycorr)%*%ycorr )/rchisq(1,nrecords + 3);

# sample intercept
       ycorr = ycorr + x[,1]*b[1];
       rhs    = sum(ycorr)/vare;
       invLhs = 1.0/(nrecords/vare);
       mean = rhs*invLhs;
       b[1] = rnorm(1,mean,sqrt(invLhs));
       ycorr = ycorr - x[,1]*b[1];
       meanb[1] = meanb[1] + b[1];

# sample effect for each locus
    nLoci = 0;
       for (locus in 1:nmarkers){
              ycorr = ycorr + x[,1+locus]*b[1+locus];
              rhs = t(x[,1+locus])%*%ycorr;
              xpx = t(x[,1+locus])%*%x[,1+locus];
              v0  =  xpx*vare;
              v1  =  (xpx^2*varEffects + xpx*vare);
              logDelta0 = -0.5*(log(v0) + rhs^2/v0) + logPi;
              logDelta1 = -0.5*(log(v1) + rhs^2/v1) + logPiComp;
              probDelta1 = 1.0/(1.0 + exp(logDelta0-logDelta1));
              u = runif(1);
              if(u < probDelta1) {
                     nLoci = nLoci + 1;
                     lhs = xpx/vare + 1.0/varEffects;
                     invLhs = 1.0/lhs;
                     mean = invLhs*rhs/vare;
                     b[1+locus]= rnorm(1,mean,sqrt(invLhs));
                     ycorr = ycorr - x[,1+locus]*b[1+locus];
                     meanb[1+locus] = meanb[1+locus] + b[1+locus];
                     ppa[locus] = ppa[locus] + 1;
                     var[locus] = varEffects;
              }
              else {
                     b[1+locus] = 0.0;
                     var[locus] = 0.0;
              }

       }
# sample common variance
       countLoci = 0;
       sum = 0.0;
       for (locus in 1:nmarkers){
              if(var[locus]>0.0){
                     countLoci = countLoci + 1;
                     sum = sum + b[1+locus]^2;
              }
```

```
       }
       varEffects = (scalec*nua + sum)/rchisq(1,nua+countLoci);


# sample Pi
       aa = nmarkers-countLoci + 1;
       bb = countLoci + 1;
       pi = rbeta(1, aa, bb);
       storePi[iter] = pi;
       piMean = piMean + pi;
#      scalec =  (nua-2)/nua*vara/((1-pi)*nmarkers*mean2pq)
       logPi     = log(pi);
       logPiComp = log(1-pi);


# sample Scale factor

#   shape = countLoci*(nua/2) + theta;
#   scale = countLoci*(nua/2)*(1/varEffects) + beta;

   shape = (nua/2) + theta;
#  scale = 1.0/((nua/2)*(1/varEffects) + beta);
   scale = (nua/2)*(1/varEffects) + beta;
   scalec = rgamma(1,shape,scale)

   meanscalec = meanscalec + scalec

   if ((iter %% 100)==0) {
              cat ("iteration ",iter," number of loci in model = ", nLoci,"\n");
              cat ("iteration ",iter," Scale Param : ", scalec, "\n");
       }
}

piMean = piMean/numiter;
meanb = meanb/numiter;
plot(meanb)
gebv=x%*%meanb
plot(hist(gebv))

#test population
nmarkers=2000
data1     = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data1)[1];
x = cbind(1,data1[,startMarker:nmarkers]);
yt = data1[,nmarkers+1];
yHat_t   = x %*% meanb;
corr = cor(yt,yHat_t);
corr
```

7.3.4 Método BLASSO no Pacote BLR (escrito por Gustavo de los Campos, University of Wisconsin)

```
#Bayesian LASSO
setwd("C:\\R 2013")

# Example of whole-Genome prediction by Lasso using BLR package

library(BLR)

nmarkers=2000;
startMarker=1801;

# training data
data = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);

x  = data[,startMarker:nmarkers];
y  = data[,nmarkers+1]

prior=list( varE=list(S=4.5,df=3),
            varBR=list(S=.009,df=3),
            lambda=list(type='random', value=30,shape=.52,rate=2e-5))
nIter<-200
```

```
burnIn<-1
fmL1<-BLR(y=y,XL=x,nIter=nIter, burnIn=burnIn,thin=1,prior=prior)

meanb = fmL1$bL;
plot(meanb)
gebv=x%*%meanb
plot(hist(gebv))

#test population
nmarkers=2000
data1     = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data1)[1];
x = data1[,startMarker:nmarkers]
yt = data1[,nmarkers+1];
yHat_t   = x %*% meanb;
corr = cor(yt,yHat_t);
corr
```

7.3.5 Método Regressão via Quadrados Mínimos Parciais (PLSR) no pacote pls (escrito por Gaston Sanchez e Laura Trinchera)

```
#PLSR
library(pls)

setwd("C:\\R 2013")

nmarkers=2000;
startMarker=1801;

# training data
data = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);

x = data[,startMarker:nmarkers];
y = data[,nmarkers+1]

nc=20 # number of components

pls_20 = plsr(y ~ x, ncomp=nc)
summary(pls_20)

# efeito de marcadores estimados na população de treinamento
eff=pls_20$coefficients[,,20]
plot(pls_20$coefficients[,,20])

# test data
nmarkers=2000
data1 = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
x = data1[,startMarker:nmarkers]
yt = data1[,nmarkers+1]
yHat_t = x %*%eff
corr = cor(yt,yHat_t)
corr
```

7.3.6 Método Regressão via Componentes Principais (PCR) no pacote pls (escrito por Gaston Sanchez e Laura Trinchera)

```
#PCR
setwd("C:\\R 2013")

library(pls)

nmarkers=2000;

# população de treinamento
data = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);
startMarker=1801;

x = data[,startMarker:nmarkers];
y = data[,nmarkers+1]
```

```
nc=67 # número de componentes
pcr = pcr(y ~ x,67)


# efeito de marcadores estimados na população de treinamento
b=pcr$coefficients[,,67]
plot(b)

# população de treinamento
data1 = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
nrecords = dim(data1)[1];
x = data1[,startMarker:nmarkers]
yt = data1[,nmarkers+1];
yHat_t = x %*%b;
corr = cor(yt,yHat_t);
corr
```

PCR Supervisionado

```
# leitura do arquivo de marcadores moleculares
setwd("C:\\R 2013")

library(pls)

dados=read.table("Veracel-A-BV-ALT-TOT.txt",h=T)

#leitura do arquivo de marcadores

snp=dados[,-(1:5)]
head(snp)
tail(snp)

#leitura do arquivo fenotípico
fenotipo=dados$EBV
head(fenotipo)
tail(fenotipo)

#transformar os dados do arquivo snp em uma matriz
M=as.matrix(snp)

# entrada: vetor y1
# M: matriz de marcadores

y=as.matrix(fenotipo)

# critério para a escolha do número de componentes
model=pcr(y~M,validation="CV")
rmsep.cv=sqrt(model$validation$PRESS/nrow(M))
nc=which(rmsep.cv==min(rmsep.cv))

# efeitos de marcadores
a=model$coefficients[,,nc]

# rank dos efeitos de marcadores
rank=cbind(colnames(M),matrix(rank(abs(a))))
colnames(rank)=c("marker","rank")

n=300 # n snps menores efeitos retirados
num_snp=matrix(seq(n+1,ncol(M)))
colnames(num_snp)=c("rank")


# merge entre o arquivo total de marcadores e os marcadores selecionados
merge_rank=merge(rank,num_snp, by=intersect("rank","rank"))
merge_rank1=matrix(merge_rank[,2])
colnames(merge_rank1)=c("marker")

marc=cbind(noquote(matrix(colnames(as.matrix(M)))),noquote(t(M)))
colnames(marc)=c("marker",1:nrow(M))

# marcadores selecionados
snp_new=merge(marc,merge_rank1, by=intersect("marker","marker"))
write.table(snp_new,"snp_new.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)
```

```
# leitura do arquivo de marcadores selecionados
snp=read.table("snp_new.txt")

M_new=t(snp[,-1])

# jacknife : validação
gbv=NULL
eff=matrix(0,ncol(M_new),nrow(M))
for(i in 1:nrow(M))
{
eff[,i]=(pcr(y[-i]~M_new[-i,]))$coefficients[,,nc]
gbv[i]=M_new[i,]%*%eff[,i]
}

# Vetor de valores genômicos dos indivíduos
gbv

#Vetor de efeito de marcadores
mean_eff=NULL
for(i in 1:nrow(eff))
{
mean_eff[i]=sum(eff[i,])/nrow(eff)
}
write.table(mean_eff,"eff_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

mean_eff


par(mfrow=c(3,1)) # divide a tela gráfica em 3
plot(mean_eff)
plot(abs(mean_eff), type = "l")
# visualização dos mais informativos
plot(density(mean_eff))

# Correlação
cor=cor(gbv,y)
cor
```

PCR Supervisionado + RR-BLUP

```
setwd("C:\\R 2013")

# Pacotes utilizados
library(rrBLUP)
library(MASS)

# leitura do arquivo de dados
dados=read.table("Veracel-A-BV-ALT-TOT.txt",h=T)

# leitura do arquivo de marcadores selecionados
snp=read.table("snp_new.txt")
colnames(snp)=c("marker",1:(ncol(snp)-1))
# head(snp)
# tail(snp)

#leitura do arquivo fenotípico
fenotipo=dados$EBV
# head(fenotipo)
# tail(fenotipo)

#leitura do arquivo da frequência das marcas
freq=read.table("allele.txt")
p=cbind(colnames(dados[,-(1:5)]),freq)
colnames(p)=c("marker","p","q")

# merge entre o arquivo total de marcas e de marcas selecionadas
merge=merge(p,snp, by=intersect("marker","marker"))
write.table(merge,"merge.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

# transformar os dados do arquivo snp selecionados em uma matriz
snp_new=read.table("merge.txt")
M=as.matrix(t(snp_new[,-(1:3)]))
```

```
# transformar os dados do arquivo fenotípico
y=as.matrix(fenotipo)

# jacknife: Validação
gbv=NULL
eff=matrix(0,ncol(M),nrow(M))
for(i in 1:nrow(M))
{
eff[,i]=(mixed.solve(y[-i], Z=M[-i,],K=diag(ncol(M))))$u
gbv[i]=M[i,]%*%eff[,i]
}
write.table(gbv,"gbv_spcr_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

# Vetor de valores genômicos dos indivíduos
gbv

#Vetor de efeito de marcadores
mean_eff=NULL
for(i in 1:nrow(eff))
{
mean_eff[i]=sum(eff[i,])/ncol(eff)
}
write.table(mean_eff,"eff_spcr_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

mean_eff

par(mfrow=c(3,1)) # divide a tela gráfica em 3
plot(mean_eff)
plot(abs(mean_eff), type = "l")
# visualização dos mais informativos
plot(density(mean_eff))

# Cálculo da herdabilidade da característica
model=mixed.solve(y, Z=M,K=diag(ncol(M))) # todos indivíduos

Va=model$Vu # variância genética aditiva explicada por 1 SNP
p_new=snp_new[,2] # frequencia dos marcadores selecionados

Vu=sum(2*(t(p_new%*%(1-p_new))*Va) ) # variância genética aditiva
Ve=model$Ve # variância residual
h2=Vu/(Vu+Ve) # herdabilidade
h2
write.table(h2,"h2_spcr_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

# Correlacao e Acurácia
cor=cor(gbv,y) # capacidade preditiva
cor
ac=cor/sqrt(h2) # acurácia
ac
```

7.3.7 Método Regressão via Componentes Independentes (ICR) pelo Pacote caret (escrito por Max Kuhn, da Pfizer)

```
library(caret)
setwd("C:\\R 2013")

# população de treinamento
nmarkers=2000;

data = matrix(scan("trainData.out0"),ncol=nmarkers+2,byrow=TRUE);
startMarker=1801;

# leitura do arquivo de marcadores moleculares
snp=data[,startMarker:nmarkers]
head(snp)
tail(snp)

#leitura do arquivo fenotípico
fenotipo=data[,nmarkers+1]
head(fenotipo)
tail(fenotipo)

#transformar os dados do arquivo snp em uma matriz
```

```
M=as.matrix(snp)
#transformar os dados do arquivo fenotípico
y=as.matrix(fenotipo)

# entrada: vetor y
# M: matriz de marcadores

fit=icr(M,y,n.comp=nc) # nc: número de componentes

# Matriz de branqueamento
K=fit$ica$ica$K

# Matriz ortogonal
R= fit$ica$ica$W

# Matriz transposta da matriz de misturas
A_trans=K%*%R

# Vetor de coeficientes associados aos componentes
gamma= fit$model$coefficients[-1]

#Vetor de efeito de marcadores
eff_snp=A_trans%*%gamma
eff_snp
par(mfrow=c(3,1)) # divide a tela gráfica em 3
plot(eff_snp)
plot(abs(eff_snp), type = "l")
# visualização dos mais informativos
plot(density(eff_snp))

# população de validação
nmarkers=2000
data1 = matrix(scan("testData.out0"),ncol=nmarkers+2,byrow=TRUE);
x = data1[,startMarker:nmarkers]
yt = data1[,nmarkers+1];
yHat_t = x %*% eff_snp;
corr = cor(yt,yHat_t);
corr
```

7.3.8 Método Regressão Ridge-BLUP (RR-BLUP) no pacote rrBLUP (escrito por Endelman)

```
library(rrBLUP)
setwd("C:\\R 2013")

# leitura do arquivo de marcadores moleculares
snp=read.table("geno.dat",h=T)
head(snp)
tail(snp)

#leitura do arquivo fenotípico
fenotipo=read.table("feno.dat",h=T)
head(fenotipo)
tail(fenotipo)

#transformar os dados do arquivo snp em uma matriz
M=as.matrix(snp)

# entrada: vetor y1
# z: matriz de marcadores
# k: é uma matriz diagonal com número de colunas igual a da matriz M
# output:  "Vu"   "Ve"   "beta" "u"    "LL"
y=fenotipo$vfen
fit2=mixed.solve(y, Z=M, K=diag(ncol(M)))

#Vetor de efeito de marcadores
fit2$u
par(mfrow=c(3,1))  # divide a tela gráfica em 3
plot(fit2$u)
plot(abs(fit2$u), type = "l")
# visualização dos mais informativos
plot(density(fit2$u))
```

```
# Vetor de valores genômicos
a=as.matrix(fit2$u)
rownames(a)<-c(colnames(snp))
gbv_rr= M%*%a
gbv_rr

# R2 vfen, vgen e vgenômico
seq <-(1:1:length(fenotipo$vfen))
seq
cor(gbv_rr,fenotipo$vfen)^2
cor(gbv_rr,fenotipo$vgen)^2
cor(fenotipo$vgen,fenotipo$vfen)^2
par(mfrow=c(3,1))  # divide a tela gráfica em 3
plot(seq,gbv_rr,xlab="Ind",ylab="EVGB",type = "l")
plot(seq,fenotipo$vgen,xlab="Ind",ylab="Vgen", type = "l")
plot(seq,fenotipo$vfen,xlab="Ind",ylab="Vfen",  type = "l")
```

RR-BLUP Completo com Validação Jacknife

```
setwd("C:\\R 2013")

# Pacotes utilizados
library(rrBLUP)

# leitura do arquivo de dados
dados=read.table("Veracel-A-BV-ALT-TOT.txt",h=T)

#leitura do arquivo de marcadores
snp=dados[,-(1:5)]
# head(snp)
# tail(snp)

#leitura do arquivo fenotípico
fenotipo=dados$EBV
# head(fenotipo)
# tail(fenotipo)

#leitura do arquivo da frequência das marcas
p=read.table("allele.txt")

# transformar os dados do arquivo snp em uma matriz
M=as.matrix(snp)

# transformar os dados do arquivo fenotípico
y=as.matrix(fenotipo)

# jacknife: Validação
gbv=NULL
eff=matrix(0,ncol(M),nrow(M))
for(i in 1:nrow(M))
{
eff[,i]=(mixed.solve(y[-i], Z=M[-i,],K=diag(ncol(M))))$u
gbv[i]=M[i,]%*%eff[,i]
}

write.table(gbv,"gbv_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

# Vetor de valores genômicos dos indivíduos
gbv

#Vetor de efeito de marcadores
mean_eff=NULL
for(i in 1:nrow(eff))
{
mean_eff[i]=sum(eff[i,])/nrow(eff)
}
write.table(mean_eff,"eff_rrblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

mean_eff

par(mfrow=c(3,1)) # divide a tela gráfica em 3
```

```
plot(mean_eff)
plot(abs(mean_eff), type = "l")
# visualização dos mais informativos
plot(density(mean_eff))

# Cálculo da herdabilidade da característica
model=mixed.solve(y, Z=M,K=diag(ncol(M))) # todos indivíduos
Va=model$Vu # variância genética aditiva explicada por 1 SNP
Vu=sum(2*(t(p[,1])%*%p[,2])*Va) # variância genética aditiva
Ve=model$Ve # variância residual
h2=Vu/(Vu+Ve) # herdabilidade
h2
write.table(h2,"h2.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

# Correlação e Acurácia
cor=cor(gbv,y) # capacidade preditiva
cor

ac=cor/sqrt(h2) # acurácia
ac
```

7.3.9 Método G-BLUP no pacote rrBLUP (escrito por Endelman)

```
library(rrBLUP)
setwd("C:\\R 2013")

# leitura do arquivo de marcadores moleculares
snp=read.table("geno.txt",h=T)
head(snp)
tail(snp)

#leitura do arquivo fenotípico
fenotipo=read.table("feno.txt",h=T)
head(fenotipo)
tail(fenotipo)

#transformar os dados do arquivo snp em uma matriz
M=as.matrix(snp)


# saida: "g.train" "beta"    "Vg"      "Ve"
# g.train: vetor de valores genômicos
#beta: estimativa do efeito fixo: neste caso é a média do fenótipo
#Vg: variancia genética
#Ve: variancia ambiental
y=fenotipo$vfen
fit1 = kinship.BLUP(y, G.train=M)
names(fit1)

#Média da característica (corrigida para efeitos fixos)
fit1$beta
mean(y)

# Vetor de valores genômicos
fit1$g.train

# R2 vfen, vgen e vgenômico
seq <-(1:1:length(fenotipo$vfen))
seq
cor(fit1$g.train,fenotipo$vfen)^2
cor(fit1$g.train,fenotipo$vgen)^2
cor(fenotipo$vgen,fenotipo$vfen)^2
par(mfrow=c(3,1))  # divide a tela gráfica em 3
plot(seq,fit1$g.train,xlab="Ind",ylab="EVGB",type = "l")
plot(seq,fenotipo$vgen,xlab="Ind",ylab="Vgen", type = "l")
plot(seq,fenotipo$vfen,xlab="Ind",ylab="Vfen",  type = "l")

# Cálculo da herdabilidade da característica
h2=(fit1$Vg)/(fit1$Ve+fit1$Vg)
h2

#estruturação dos valores genômicos em matriz para uso do sort
gbv=as.matrix(fit1$g.train)

# identificação dos fenotipos
```

```
rownames(gbv)<-c(fenotipo$id)

#ordenamento dos valores genômicos
gbv1=sort(as.matrix(gbv)[,],decreasing=TRUE) #ordenamento dos valores
plot(hist(gbv1))
gbv_10=quantile(gbv1, probs =c(90)/100) #identificação dos 10% melhores
top_10=gbv1[gbv1>=gbv_10]
top_10


#Vetor de efeito de marcadores
#a=inv(t(M)*M)*(t(M)*u)  sendo u = gbv
library(MASS)
a0=ginv(t(M)%*%M)%*%(t(M)%*%gbv)
par(mfrow=c(3,1))  # divide a tela gráfica em 3
plot(a0)
plot(abs(a0), type = "l")
# visualização dos mais informativos
plot(density(a0))
```

Método G-REML/G-BLUP no pacote rrBLUP (escrito por Endelman)

```
M=as.matrix(read.table("snp_2011_fim.txt",h=T))
dim(M)
M[1:5,1:5]

#Calculando freq alélicas

library(genetics)

M1=M+1 #transformando genótipos -1, 0 e 1 em 0, 1 e 2

Q=matrix(0,nrow(M1),ncol(M1))
Q[M1==0]<-"D/D"
Q[M1==1]<-"D/I"
Q[M1==2]<-"I/I"

p=matrix(0,ncol(Q),1)
q=matrix(0,ncol(Q),1)
pq=matrix(0,ncol(Q),1)

for(i in 1:ncol(Q))
{
p[i,]=matrix(summary(genotype(Q[,i]))$allele.freq[1,2],1,1)
q[i,]=matrix(summary(genotype(Q[,i]))$allele.freq[2,2],1,1)
pq[i,]=p[i,]*q[i,]
}
sum2pq=2*sum(pq)

fenotipo=read.table("fenotipos_2011.txt",h=T)
dim(fenotipo)
fenotipo[1:5,]

#corrigindo fenótipo para efeitos fixos

ym_slg=factor(fenotipo$ym_slg)
farm=factor(fenotipo$farm)
hcw=fenotipo$hcw
y=fenotipo$y
y1=mean(y) + lm(y~ ym_slg + farm + hcw)$residuals

#ajustando RR-BLUP

library(rrBLUP)
rrblup = mixed.solve(y1,Z=M)

Va=rrblup$Vu       #variância genética aditiva explicada por 1 SNP
sigma2_a=sum2pq*Va  #variância genética aditiva
Ve=rrblup$Ve       #Variância residual
h2_r=sigma_a/(sigma_a+Ve)

a_hat=rrblup$u     #vetor de efeitos estimados SNPs
plot(a_hat)
```

```
u_hat=M%*%rrblup$u #vetor de GBV estimados
plot(hist(u_hat))

#ajustando G-BLUP

gblup = mixed.solve(y1,K=A.mat(M))
gblup$Vu
gblup$Ve
h2_g=gblup$Vu/(gblup$Vu+gblup$Ve)

cor(gblup$u,u_hat)
library(MASS)
a_hat_g=ginv(t(M)%*%M)%*%t(M)%*%gblup$u
cor(a_hat_g,a_hat)
```

GBLUP Reduzido R

```
# leitura do arquivo de marcadores moleculares

setwd("C:\\R 2013")

library(rrBLUP)
library(MASS)

dados=read.table("Veracel-A-BV-ALT-TOT.txt",h=T)

#leitura do arquivo de marcadores
snp=dados[,-(1:5)]
head(snp)
tail(snp)

#leitura do arquivo fenotípico
fenotipo=dados$EBV
head(fenotipo)
tail(fenotipo)

#transformar os dados do arquivo snp em uma matriz
M=as.matrix(snp)

# entrada: vetor y1
# M: matriz de marcadores
# k: é uma matriz diagonal com número de colunas igual a da matriz M
# output: "Vu" "Ve" "beta" "u" "LL"

y=as.matrix(fenotipo)

u=as.matrix(mixed.solve(y,K=A.mat(M))$u) # valor genético dos indivíduos
a=ginv(t(M)%*%M)%*%(t(M)%*%u) # efeitos de todos os marcadores
rank=cbind(colnames(M),matrix(rank(abs(a))))
colnames(rank)=c("marker","rank")

n=300 # n de marcadores de menor efeito retirados
num_snp=matrix(seq(n+1,ncol(M)))
colnames(num_snp)=c("rank")

# merge entre o arquivo total de marcadores e os marcadores selecionados
merge_rank=merge(rank,num_snp, by=intersect("rank","rank"))
merge_rank1=matrix(merge_rank[,2])
colnames(merge_rank1)=c("marker")

marc=cbind(noquote(matrix(colnames(as.matrix(M)))),noquote(t(M)))
colnames(marc)=c("marker",1:nrow(M))

# marcadores selecionados
snp_new=merge(marc,merge_rank1, by=intersect("marker","marker"))
write.table(snp_new,"snp_new.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

#leitura dos marcadores selecionados

snp=read.table("snp_new.txt")

M_new=t(snp[,-1])
```

```
# jacknife : Validação2
gbv=NULL
eff=matrix(0,ncol(M_new),nrow(M))
for(i in 1:nrow(M))
{
a=as.matrix(mixed.solve(y[-i],K=A.mat(M_new[-i,]))$u)
eff[,i]=ginv(t(M_new[-i,])%*%M_new[-i,])%*%(t(M_new[-i,])%*%a)
gbv[i]=M_new[i,]%*%eff[,i]
}
# Vetor de valores genômicos dos indivíduos
gbv

#Vetor de efeito de marcadores

mean_eff=NULL
for(i in 1:nrow(eff))
{
mean_eff[i]=sum(eff[i,])/ncol(eff)
}

write.table(mean_eff,"eff_rgblup.txt",row.names=FALSE,col.names=FALSE,quote=FALSE)

mean_eff
par(mfrow=c(3,1)) # divide a tela gráfica em 3
plot(mean_eff)
plot(abs(mean_eff), type = "l")
# visualização dos mais informativos
plot(density(mean_eff))

# Cálculo da herdabilidade da característica
model=mixed.solve(y, K=A.mat(M_new)) # todos indivíduos

Vu=model$Vu  # variância genética aditiva
Ve=model$Ve   # variância residual
h2=Vu/(Vu+Ve)   # herdabilidade
h2
# Correlacao, Acuracia
cor=cor(gbv,y)
cor
ac=cor/sqrt(h2)
ac
```

7.3.10 Análise Espacial no Método RR-BLUP: Função *rhierLinearModel* no Pacote *Bayesm*: escrito por Peter Rossi (University of Califórnia)

```
setwd("C:\\R 2013")
library(bayesm)

y=as.matrix(read.table("fen_final.txt"))
iota=rep(1,nrow(y))
x=cbind(iota,as.matrix(read.table("touros_final.txt")))

regdata=NULL
for(reg in 1:1)
{
regdata[[reg]]=list(y=y,X=x)
}

Data1=list(regdata=regdata)
Prior1=list(nu=52,V=diag(32,ncol(x)))
Mcmc1=list(R=20000,keep=4)

fit=rhierLinearModel(Data=Data1, Prior=Prior1,Mcmc=Mcmc1)

var=diag(matrix(apply(fit$Vbetadraw,2,mean),ncol(x),ncol(x))) #[-1]
var
dim(fit$Vbetadraw)
beta=apply(fit$betadraw,2,mean)
beta
vare=mean(fit$taudraw)
vare
```

* Recomenda-se o ajuste de um modelo contemplando efeitos aleatórios de cromossomos individuais, com estrutura espacial dentro de cromossomos, com ajuste de todos os cromossomoos simultaneamente.

7.3.11 Método Regressão Kernel Hilbert Spaces (RKHS) (escrito por Gustavo de los Campos, University of Wisconsin)

```
setwd("C:\\R 2013")

data = read.table("fen_final.txt"))
phe = as.matrix(data[,5])
gen = matrix(as.numeric(as.matrix(data[,-c(1:5)])),nrow = nrow(data))

D<-as.matrix(dist(gen,method="euclidean"))^2
D<-D/mean(D)
h<-c(1e-2,.1,.4,0.5,.8,1.5,2,3,5)
R2<-numeric()
PMSE<-numeric()
VARE<-numeric()
VARU<-numeric()

for(i in 1:length(h)){
       print(paste('Working with h=',h[i],sep=''))
       # COMPUTES THE KERNEL
       K<-exp(-h[i]*D)
       # FITS THE MODEL
       prefix<- paste(h[i], "_",sep="")
       fm<-RKHS(y=phe,K=list(list(K=K,df0=5,S0=2)),
              nIter=12000,burnIn=2000,df0=5,S0=2,
              saveAt=prefix)
       R2[i] = cor(fm$yHat,phe)
       PMSE[i]<-mean((phe-fm$yHat)^2)
       VARE[i]<-fm$varE
       VARU[i]<-fm$K[[1]]$varU
}
plot(R2~h,xlab="Bandwidth",
             ylab="Residual Variance",type="o",col=4)
plot(VARE~h,xlab="Bandwidth",
             ylab="Residual Variance",type="o",col=4)

plot(PMSE~h,xlab="Bandwidth",
             ylab="PMSE",type="o",col=2)

plot(I(VARE/VARU)~h,xlab="Bandwidth",
      ylab="Ratio of variances(noise/signal)",
      type="o",col=4)
```

8 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.

AITKEN, A. C. Studies in practical mathematics: the evaluation of the latent roots and latent vectors of a matrix. **Proceedings of the Royal Society of Edinburgh**, v. 57, p. 269-304, 1937.

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

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.

ANDERSON, L.; GEORGES, M. Domestic animal genomes: deciphering the genetics of complex traits. **Nature Reviews Genetics**, v. 5, n.3, p.202-212, 2004.

ARANGO, J.; MISZTAL, I.; TSURUTA S.; CULBERTSON, M.; HERRING, W. Estimation of variance components including competitive effects of Large White growing gilts. **Journal of Animal Science**, v. 83, p. 1241-1246, 2005.

AUER, P.L.; DOERGE, R.W. Statistical design and analysis of RNA sequencing data. Genetics. 185: 405-416, 2010.

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.

AYROLES, J. F.; GIBSON, G. Analysis of variance of microarray data. **Methods in Enzymology**, v. 411, p. 214-233, 2006.

AZEVEDO, C. F. **Métodos de redução de dimensionalidade aplicados na seleção genômica para características de carcaça em suínos.** 2012. Dissertação (Estatística Aplicada e Biometria) - Universidade Federal de Viçosa.

AZEVEDO, C. F.; RESENDE, M.D.V.; SILVA, F.F.; LOPES, P.S.; GUIMARAES, S.E.F. Regressão via Componentes Independentes (ICR) para redução de dimensionalidade na seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, 2012.

AZEVEDO, C. F. ; SILVA, F. F ; PETERNELLI, L. A. ; RESENDE, M. D. V. ; GUIMARÃES, S. E. F. Regressão via componentes principais aplicada a seleção genômica ampla. In: XI MGEST, 2012. XI MGEST, 2012.

AZEVEDO, C. F. ; SILVA, F. F ; RESENDE, M. D. V. ; PETERNELLI, L. A. ; GUIMARÃES, S. E. F. Quadrados mínimos parciais multivariado: uma aplicação a seleção genômica considerando características de carcaça em suínos. In: SINAPE, 2012, João Pessoa. 20º SINAPE, 2012.

AZEVEDO, C. F. ; SILVA, F. F ; PETERNELLI, L. A. ; RESENDE, M. D. V. ; GUIMARÃES, S. E. F. Quadrados mínimos parciais aplicado a seleção genômica considerando características de carcaça em suínos. In: Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 2012, Piracicaba. 57ª RBRAS, 2012.

BAYES, T. An essay towards solving a problem in the doctrine of chances. ***Philos. Trans. R. Soc.***, London, v. 53, p. 370-418, 1763.

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

BISHOP, C.M. **Pattern recognition and machine learning**. Springer, 2006.

BOX, G. E. P.; TIAO, G. C. ***Bayesian inference in statistical analysis***. Reading: Addison-Wesley Publ. Co., 1973. 588 p.

BROTHERSTONE, S.; WHITE, I.M.S.; SYKES, R.; THOMPSON, R.; CONNOLLY, T.; LEE, S.; WOOLLIAMS, J. Competition Effects in a Young Sitka Spruce (Picea sitchensis, Bong. Carr) Clonal Trial. **Silvae Genetica**, v. 60, n. 3-4, p. 149-155, 2011.

BUENO FILHO, J. S. S.; VENCOVSKY, R. Selection in several environments by BLP as an alternative to pooled ANOVA in crop breeding. **Ciência e Agrotecnologia**, v. 33, p. 1342-1350, 2009.

BULMER, M. G. ***The mathematical theory of quantitative genetics***. Oxford: Charedon Press, 1980. 254 p.

CADAVID, A. C.; LAWRENCE, J. K.; RUZMAIKIN, A.; KAYLENG–KNIGHT. Principal components and independent component analysis of solar and space data. **Solar Phys**, v. 248, p. 247-261, 2008.

CALUS, M. P. L.; MEUWISSEN, T. H. E.; ROOS, A. P. W.; VEERKAMP, R. F. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 178, p. 553-561, 2008.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of breeding value when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. **Journal of Animal Breeding and Genetics**, v. 124, p. 362-368, 2007.

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.

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, 2009.

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, 2009.

CAVALCANTI, J. J. V.; RESENDE, M. D. V. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, v.34, p., 2012.

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

CHURCHILL, G. A.; DOERGE, R. W. Empirical threshold values for quantitative trait mapping. **Genetics**, v. 138, p. 963-971, 1994.

COCHRAN, W. G. Improvement by means of selection. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 2., 1951, Berkeley. **Proceedings**... Berkeley: University of California Press, 1951. p. 449-470.

COMON, P. Independent component analysis – a new concept. **Sigmal Processing**, v. 45, p. 59-83, 1994.

BIJMA P. A general definition of the heritable variation that determines the potential of a population to respond to selection. **Genetics** 189, 1347-1359, 2011.

COSTA e SILVA, J.; KERR, R.J. (2012). Accounting for competition in genetic analysis, with particular emphasis on forest genetic trials. **Tree Genetics and Genomes** (DOI 10.1007/s11295-012-0521-8).

COSTA e SILVA, J.; POTTS, B.M.; BIJMA, P.; KERR, R.J.; PILBEAM, D. Genetic control of interactions amongst individuals: Contrasting outcomes of indirect genetic effects arising from neighbour disease infection and competition in a forest tree (**New Phytologist**, 2012, in press).

CRUZ, C. D. ; GOOD 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.

CUI, X.; HWANG, J.T.G.; QIU, J. *et al.* Improved statistical tests for differential gene expression by shrinking variance components estimates. **Biostatistics**, v.6, n.1, p.59-75, 2005.

CULLIS, B. R.; GLEESON, A. C. Spatial analysis of field experiments-an extension at two dimensions. **Biometrics**, v. 47, p. 1449-1460, 1991.

CULLIS, B. R.; GOGELL, B.; VERBYLA, A.; THOMPSON, R. Spatial analysis of multi-environment early generation variety trials. **Biometrics**, v. 54, p. 1-18, 1998.

DAETWYLER H.D; VILLANUEVA B; BIJMA P.; WOOLLIAMS JA (2007) Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLoS ONE** 3:e3395.

DAETWYLER H.D; PONG-WONG, R.; VILLANUEVA B; WOOLLIAMS, J.A. The impact of genetic architecture on genome-wide evaluation methods. **Genetics,** v.185, p.1021-1031, 2010.

DAETWYLER, H. D. .; KEMPER, K. E. .; VAN DER WERF, J. H. J.; HAYES, B. J. Components of the accuracy of genomic prediction in a multi-breed sheep population. **J. Anim. Sci.** v. 90, p. 3375–3384, 2012.

DARVASI, A.; SOLLER, M. A simple method to calculate resolving power and confidence interval of QTL map location. **Behavior Genetics**, v. 27, p. 125- 132, 1997.

DEKKERS, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. **Journal of Animal Science**, v.85, p. 2104-2114, 2007.

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

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

DEKKERS, J. C. M. Prediction of response to marker assisted and genomic selection using selection index theory. **Journal of Animal Breeding and Genetics**, v. 124, p. 331-341, 2007.

DEMPFLE, L. Relation entre BLUP (Best linear unbiased prediction) et estimateurs bayesiens. ***Annales de Génétique et Sélection Animale***, v. 9, p. 27-32, 1977.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistic Society**, London, v. 39, p. 1-38, 1977.

DROST, D.R.; NOVAES, E.; BOAVENTURA-NOVAES, C.; BENEDICT, C.I.; BROWN, R.S.; YIN, T.; TUSKAN,, G.A.; KIRST, M. A microarray-based genotyping and genetic mapping approach for highly heterozygous outcrossing species enables localization of a large fraction of the unassembled Populus trichocarpa genome sequence. **The Plant Journal**, 2008. doi: 10.1111/j.1365-313X.2009.03828.x

DUARTE, J. B.; VENCOVSKY, R. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agrícola**, v. 58, n. 1, p. 109-117, 2001.

EFRON, B.; MORRIS, C. Stein's paradox in statistics. **Scientific American**, v. 236, n. 5, p. 119-127, 1977.

ELSTON, R.C.; SATAGOPAN, J.M; SUN, S. **Statistical human genetics: methods and protocols**. Humana Press, 2012. 575 p.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome** , v4, p. 250–255, 2011.

ENDELMAN, J. B.; JANNINK, J.L. Shrinkage estimation of the realized relationship matrix. **PGenes, Genomes, Genetics** , v2, p. 1405–1413, 2012.

EWING, B.; GREEN, P. Analysis of expressed sequence tags indicates 35,000 human genes. **Nature Genetics**, v. 25, p.232-234, 2000.

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

FANG, Y. Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. ***Journal of Data Science***, v.9, p.15-21, 2011.

FERNANDO, R. L., 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. **Proceedings of the 6th World Congress on Genetics Applied to Livestock Production**, Armidale, NSW, Australia, Vol. 26, pp. 329–336.

FERNANDO, R. L.; GIANOLA, D. Optimal properties of the conditional mean as a selection criterion. ***Journal of Animal Science***, v. 59, p. 177, 1984.

FERNANDO, R. L.; NETTLETON, D.; SOUTHEY, B. R.; DEKKERS, J. C. M.; ROTHSCHILD, M. F.; SOLLER, M. Controlling the proportion of false positives in multiple dependent tests. **Genetics**, v. 166, p.611-619, 2004.

FERNANDO, R.L.; GROSSMAN, M. Marker-assisted selection using best linear unbiased prediction. **Genetics Selection Evolution**, v. 21, p. 467-477, 1989.

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

FERNANDO, R.L.; STRICKER, C.; ELSTON, R.L. The finite polygenic mixed model an alternative formulation for the mixed model of inheritance. **Theoretical and Applied Genetics**, v.88, p.573-580, 1994.

FERRAZ, J.B.S. ; REZENDE, F.M. Seleção genômica: o estado da arte. In: Luiz Fernando Aarão Marques. (Org.). A Importância da pecuária bovina na economia brasileira: Coletânea do III Congresso Capixaba de Pecuária Bovina. 1 ed. Alegre (ES): CAUFES, 2012, v. 1, p. 91-122.

FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London, Series A**, n. 222, p. 309-368, 1922.

FISHER, R. A. ***Statistical methods for research workers***. 1. ed. London: Oliver and Boyd, 1925. 314 p.

FISHER, R. A. The arrangement of field experiments. **Journal of the Ministry of Agriculture of Great Britain**, v. 33, p. 503-513, 1926.

FISHER, R. A. The correlation between relatives on the supposition of mendelian inheritance. ***Transaction Royal Society of Edinburgh***, v. 32, p. 399-433, 1918.

FOULLEY, J. L. **Le modéle linéaire mixte**. Paris: INRA, 2003. 139 p.

FOULLEY, J. L.; DYK, D. A. van. The PX-EM algorithm for fast and stable fitting of Henderson's mixed model. **Genetics, Selection, Evolution**, v. 32, p. 143-163, 2000.

FOULLEY, J. L.; QUAAS, R. L. Heterogeneous variances in gaussian linear mixed models. **Genetics Selection Evolution**, v. 27, p. 211–228, 1995.

FRITSCHE NETO, R. **Seleção genômica ampla para as eficiências no uso de nitrogênio e fósforo em milho tropical**. 2011. Tese (Genética e Melhoramento) - Universidade Federal de Viçosa.

FRITSCHE-NETO, R.; DOVALE, J. C.; RESENDE, M. D. V.; MIRANDA, G.V. Genome wide selection for root traits in tropical maize under stress conditions of nitrogen and phosphorus. **Acta Scientiarum Agronomy,** v34, p.389-395, 2012.

FRITSCHE-NETO, R ; RESENDE, M. D. V. de ; DOVALE, J. C. ; LANES, ECM ; SEDIYAMA, C. S. ; PEREIRA, F.B.; MIRANDA, G. V. Seleção genômica ampla e novos métodos de melhoramento do milho. **Revista Ceres** (Online), 2013.

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

GAMERMAN, D. ***Simulação estocástica via cadeias de Markov***. Caxambu: Associação Brasileira de Estatística, 1996. 196 p.

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

GARTHWAITE, P.H. An Interpretation of Partial Least Squares. **Journal of the American Statistical Association**, v. 89, p. 122-127, 1994.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. ***Journal of the American Statistical Association***, v. 85, p. 398-409, 1990.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distribution and the bayesian restoration of imagens. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, v. 6, p. 721-741, 1984.

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

GEORGE, A.W.; VISSCHER, P.M.; HALEY, C. S. Mapping quantitative trait loci in complex pedigree: a two step variance component approach. ***Genetics***, v. 156, p.2081-2092, 2000.

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

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.; 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.; FERNANDO, R. L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v. 63, p. 217-244, 1986.

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.

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

2003.

GILMOUR, A. R. Mixed model regression mapping for QTL detection in experimental crosses. **Computational Statistics e Data Analysis**, v.51, n.8, p. 3749-3764, 2007.

GILMOUR, A. R. Mixed model regression mapping for QTL detection in experimental crosses. **Computational Statistics and Data Analysis**, v. 51, n. 8, p. 3749-3764, 2007.

GILMOUR, A. R.; THOMPSON, R. Modelling variance parameters in ASREML for repeated measures. In: WORLD CONGRESS ON GENETIC APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. **Proceedings**... Armidale: AGBU: University of New England, 1998. v. 27, p. 453-454.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average information REML: an efficient algorithm for parameter estimation in linear mixed models. **Biometrics**, v. 51, p. 1440-1450, 1995.

GODDARD, M. E. A mixed model for analysis of data on multiple genetic markers. **Theoretical and Applied Genetics**, v. 83, p. 878-886, 1992.

GODDARD, M. E. Mapping genes for quantitative traits using linkage disequilibrium. **Genetics Selection Evolution**, v.23, p. 131-134, 1991.

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

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

GODDARD, M.E. New technology to enhance genetic improvement of pigs. **Manipulating Pig Production**, v.7, p.44–52, 1999.

GODDARD, M.E.; HAYES, B. J.; MEUWISSEN, T. H. E. Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of Animal Breeding and Genetics**, v. 128, n. 6, p.409-421, 2011.

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**, v.178, n.4, p. 2305 – 2313, 2008.

GRASER, H. U.; SMITH, S. P.; TIER, B. A derivative free approach for estimating variance components in animal models by restricted maximum likelihood. **Journal of Animal Science**, Champaign, v. 64, n. 5, p. 1362-1370, 1987.

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

HAAS, B. J.; ZODY, M. C. Advancing RNA-Seq analysis. Nature Biotechnology, 28: 421–423, 2010.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of Genetic Relationship on Genome-Assisted Breeding Values. **Genetics**, v. 117, p. 2389-2397, 2007.

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

HALDANE, J. B. S. The combination of linkage value and the calculation of distances between the loci of linkage factors. **Journal of Genetics**, v. 8, p. 299-309, 1919.

HALEY, C.S.; KNOTT, S.A. A simple regression method for mapping quantitative loci in line crosses using flanking markers. **Heredity**, v. 69, p.315-324, 1992.

HARTL, L. D.; JONES, E. W. **Essencial genetics: a genomics perspective**. Sudbury: Jones & Bartlet, 2002.

HARTLEY, H. O.; RAO, J. N. K. Maximum likelihood estimation for the mixed analysis of variance model. **Biometrika**, v. 54, p. 93-108, 1967.

HARVILLE, D. A. **Matrix algebra from a statistician perspective**. New York: Springer Verlag, 1997. 630 p.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American Statistical Association**, v. 72, n. 2, p. 320-328, 1977.

HARVILLE, D. A.; CARRIQUIRY, A. L. Classical and Bayesian prediction as applied to unbalanced mixed linear models. **Biometrics**, v. 48, p. 987-1003, 1992.

HASEMAN, J. M.; ELSTON, R. C. The investigation of linkage between a quantitative trait and a marker locus. **Behavioral Genetics**, v.2, p.3-19, 1974.

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

HAYES, B. J. **Course on QTL mapping, MAS and genomic selection**. Ames: Iowa State University, 2008.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges**. Journal of Dairy Science**, 2009. doi:10.3168/jds.2008-1646.

HAYES, B. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Use of markers in linkage disequilibrium with QTL in breeding programs. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

HAYES, B. J.; CHAMBERLAIN, A. J.; McPARTLAN, H.; MACLEOD, I.; SETHURAMAN, L.; GODDARD, M. E. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. **Genetical Research**, v.89, p. 215-220, 2007.

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

HAYES, B.J.; VISSCHER, P. E.; MCPARTLAN, H.; GODDARD, M. E. A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. **Genome Research**, v.13, p. 635–643, 2003.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection for crop improvement**. Crop Science**, v49, n.1, p. 1 – 12, 2009.

HENDERSON, C.R. Selection index and expected genetic advance. In: HANSON, W.D.; ROBINSON, H.F. (Ed.). ***Statistical genetics and plant breeding***. Whashington: National Academy of Sciences, 1963. p. 141-163. (NAS-NCR. Pub., 982).

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, v. 32, p. 69-83, 1976.

HENDERSON, C. R. **Aplications of linear models in animal breeding**. Guelph: University of Guelph, 1984. 462 p.

HENDERSON, C. R. Best linear estimation and prediction under a selection model. **Biometrics**, v. 31, p. 423-447, 1975.

HENDERSON, C. R. Estimation of changes in herd environment. ***Journal of Dairy Science***, v. 32, p. 709, 1949.

HENDERSON, C. R. ***Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine***. Ames: Iowa State University, 1948. Ph. Thesis.

HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, v. 9, p. 226-252, 1953.

HENDERSON, C. R. Estimation of variances in animal model and reduced animal model for single traits and single records. **Journal of Dairy Science**, v. 69, p. 1394-1402, 1986.

HENDERSON, C. R. **Sire evaluation and genetic trends**. In: ANIMAL BREEDING AND GENETICS SYMPOSIUM IN HONOUR OF J. LUSH, 1973, Champaign. **Proceedings...** Champaign: American Society of Animal Science, 1973. p.10-41.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; VON KROSIGH, C. M. The estimation of environmental and genetic trends from records subject to culling**.** ***Biometrics***, v. 15, p. 192, 1959.

HILL W.G. Estimation of effective population-size from data on linkage disequilibrium. **Genetical Research** , v.38, p.209–216, 1981.

HILL, W. G.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, v. 38, p. 226-231, 1968.

HOGGART, C. J.; WHITTAKER, J. C.; DE IORIO, M.; BALDING, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. **PLoS Genetics**, v.4, n.7, e1000130, 2008.

HOSPITAL, F.; MOREAU, L.; LACOUDRE, F.; CHARCOSSET, A.; GALLAIS, A. More on the efficiency of marker assisted selection. **Theoretical and Applied Genetics**, v. 95, p.1181-1189, 1997.

HYVÄRINEN, A. New approximations of differential entropy for independent component analysis and projection pursuit. **In Advances in Neural Information Processing Systems,** v. 10, p. 273-279, 1998.

JAFFREZIC, F.; MEZA, C.; LAVIELLE, M.; FOULLEY, J. L. Genetic analysis of growth curves using the SAEM algorithm. **Genetics Selection Evolution**, v. 38, n. 6, p. 583-600, 2007.

JAMES, W.; STEIN, C. Estimation with quadratic loss. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 4., 1961, Berkeley: **Proceedings**... Berkeley: University of Berkeley, 1961. p. 361-379.

JAMES, W.; STEIN, C. Estimation with quadratic loss**.** ***Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability***, v. 1, p. 361-379, 1961.

JAMROZIK, J.; SCHAEFFER, L. R. Estimates of genetic parameters for a test day model with random regressions for yield of first lactation Holsteins. ***Journal of Dairy Science***, v. 80, p. 726-770, 1997.

JAMROZIK, J.; SCHAEFFER, L. R.; DEKKERS, J. C. M. Genetic evaluation of dairy cattle using test day yields and random regression model. ***Journal of Dairy Science***, v. 80, p. 1217-1226, 1997.

JANSEN, R. C.; NAP, J. Genetical genomics: the added value from segregation. **Trends in Genetics**, v. 17, p.388-391, 2001.

JANSS, L.; DE LOS CAMPOS, G.; SHEEHAN, N.; SORENSEN, D. Inferences from Genomic Models in Stratified Populations. **Genetics**, 2012. Doi: 10.1534/genetics.112.141143

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. **Journal of Dairy Science**, v. 78, p. 449-456, 1995.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. Englewood : Prentice Hall Inc., 1988. 594 p.

KENNEDY, B. W.; QUINTON, M.; VAN ARENDONK, J. A. M. Estimation of effects of single genes on quantitative traits. **Journal of Animal Science**, *v.* **70, p.** 2000-2012, 1992.

KERR, M.K.; MARTIN, M.; CHURCHILL, G.A. Analysis of variance for gene expression microarray data. **Journal of Computational Biology**, v.7, n. 6, p.819-837, 2000.

KNOTT, S.A.; ELSEN, J.M.; HALEY, C.S. Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. **Theoretical and Applied Genetics**, v. 93, p.71-80, 1996.

KOSAMBI, D. D. The estimation of map distances from recombination values. **Annals of Eugenics**, v. 12, p. 172-175, 1944.

KRUGLYAK, L. Prospect for whole genome linkage disequilibrium mapping of common disease genes. **Nature Genetics**, v. 22, p.139-144, 1999.

LAIRD, N.M.; LANGE, C. **The fundamentals of modern statistical genetics**. Harvard: Springer, 2010. 240 p.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743-756, 1990.

LANDER, E. S.; BOTSTEIN, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**, v. 121, p.185-199, 1989.

LEDOIT, O.; WOLF, M. A well-conditioned estimator for large-dimensional covariance matrices. **J. Multivariate Anal**. v88, p. 365–411, 2004.

LEE, S. H.; van der WERF, J. H. J. The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree. **Genetics**, v. 169, p. 455-466, 2005.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized linear models with random effects: unified analysis via H – likelihood**. London: Chapman & Hall, 2007. 416 p.

LEE, Y.; HA, I. D. Orthodox BLUP versus h-likelihood methods for inference about random effects in Tweedie mixed models. **Statistics and Computing**, v. 20, n. 3, p.295-303 , 2010.

LEEMIS, L. M. Relationships among common univariate distributions. **The American Statistician**, v. 40, n. 2, p. 143–146, 1986.

LEGARRA, A.; MISZTAL, I. Computing strategies in genome-wide selection. **Journal of Dairy Science**, v. 91, n.1, p. 360-366, 2008.

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.
LEITE, H. G.; OLIVEIRA, F. H. T. Statistical procedure to test the identity of analytical methods. **Communications in Soil Science Plant Analysis**, New York, v. 33, n. 7/8, p. 1105-1118, 2002.

LEWONTIN, R. C. The interaction of selection and linkage. II. Optimal models. **Genetics**, v. 50, p. 757-782, 1964.

LONG N, GIANOLA D, ROSA GJ, WEIGEL KA. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **J Anim Breed Genet.** 2011.

LOPES, P. S. **Teoria do Melhoramento Animal**. 1. ed. Belo Horizonte: FEPMVZ Editora, 2005. 118 p.

LOPES, P. S.; MARTINS, E. N.; SILVA, M. A. E.; REGAZZI, A. J. **Estimação de componentes de variância**. Viçosa: Imprensa Universitária, 1998. 61 p.

LOPES, P. S. ; MARTINS, E. N. ; SILVA, M. A. E. ; RAGGI, L. A. . **Métodos de resolução de sistemas de equações lineares**. Viçosa - MG: Imprensa Universitária, 1999. 55 p.

LUO, Z. W. Linkage disequilibrium in a two-locus model. **Heredity**, v. 80, p.198-208, 1998.

LUSH, J. L. ***Animal breeding plans***. 3. ed. Ames: Iowa State University Press, 1945. 443p.

LUSH, J. L. ***Animal breeding plans***. Ames: Iowa State University Press. 1937. 433p.

LUSH, J. L. Family merit and individual merit as bases for selection. **American Naturalist**, v. 81, p. 241-261, 1947.

LUSH, J. L. The number of daughters necessary to prove a sire. **Journal of Dairy Science**, v. 14, p. 209-220, 1931.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, Inc., 1997. 980 p.

MACLEOD, I. M.; HAYES, B. J.; SAVIN, K.; CHAMBERLAIN, A. J.; MCPARTLAN, H.; GODDARD, M. E. Power of dense bovine single nucleotide polymorphisms (SNPs) for genome scans to detect and position quantitative trait loci (QTL). **Genetics**, 2008 (in press).

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.

MARTINS, E. N. ; LOPES, P. S. ; SILVA, M. A. E. ; REGAZZI, A. J. **Modelo linear misto**. Viçosa: Imprensa Universitária, 1998. 46 p.

MARTINS, E. N. ; LOPES, P. S. ; SILVA, M. A. E. ; TORRES JUNIOR, R. A. **Uso de modelos mistos na avaliação genética animal**. Viçosa: Imprensa Universitária, 1997. v. 1. 121 p.

MATHERON, G. **La Théorie des variables régionalisées et ses applications**. École Nationale Supérieure des Mines de Paris, 1970.

McRAE, A. F.; McEVAN, J. C.; DODDS, K. G.; WILSON, T.; CRAWFORD, A. M.; SLATE, J. Linkage disequilibrium in domestic sheep. **Genetics**, v. 160, p.1113-1122, 2002.

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

MEUWISSEN, T.H. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. **Genetics Selection Evolution** v.41, p.35, 2009.

MEUWISSEN, T. H. E.; GODDARD, M. E. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. **Genetics**, v. 155, p.421-430, 2000.

MEUWISSEN, T. H. E.; GODDARD, M. E. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. **Genetics Selection Evolution**, v. 36, p. 261-279, 2004.

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

MEUWISSEN, T. H. E.; KARLSEN, A.; LIEN, S; OLSAKER, I; GODDARD, M. E. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. **Genetics**, v.161, p. 373-379, 2002.

MEUWISSEN, T. H. E. ; LUAN, T.; WOOLLIAMS, J. A. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. **Journal of Animal Breeding and Genetics**, v. 128, n. 6, p.429-39, 2011.

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.

MEYER, K. DFREML – a set of programs to estimate variance components under an individual animal model. ***Journal of Dairy Science***, Champaign, n. 2, Suppl., p. 33-34, 1988.

MEYER, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. **Genetique, Selection, Evolution**, v. 23, p. 67-83, 1991.

MEYER, K. Random regression analyses using B-splines to model growth of Australian Angus cattle. **Genetics, Selection, Evolution**, v. 37, p. 473-500, 2005.

MEYER, K. WOMBAT – digging deep for quantitative genetic analysis by restricted maximum likelihood. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

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.

MISZTAL, I.; PEREZ-ENCISO, M. Sparse matrix inversion for restricted likelihood estimation of variance components by expectation-maximization. **Journal of Dairy Science**, v. 76, p. 1479-1483, 1993.

MOREAU, L.; MONOD, H.; CHARCOSSET, A.; GALLAIS, A. Marker assisted selection with spatial analysis of unreplicated field trials. **Theoretical and Applied Genetics**, v. 98, p.234-242, 1999.

MORGAN, T.H. The theory of genes. New Heaven: Yale University Press, 1928.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L., WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. **Nat Methods**, 5(7):621-628.

MORTON, N.E. Sequential tests for the detection of linkage. **American Journal of Human Genetics**, v.7, p. 277-318, 1955.

MRODE, R. A. ***Linear models for the prediction of animal breeding values***. Wallingford: CAB International, 2005. 2 Edition.

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

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v. 124, p. 342-355, 2007.

MUIR, W.M. Incorporation of competitive effects in forest trees or animal breeding programs. **Genetics**, v. 170, p. 1247-1259, 2005.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J.P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, v. 75, p. 1738 - 1745,1997.

ØDEGÅRD J.; MEUWISSEN T.H. Estimation of heritability from limited family data using genome-wide identity-by-descent sharing. **Genet Sel Evol**, v.44, n.1, p.16, 2012.

OLIVEIRA, E. J.; RESENDE, M.D.V; SANTOS, V.S. et al. Genome-wide selection in cassava. **Euphytica**, v. 187, p.263-276, 2012.

OTTO, M. **Chemometrics: Statistics and Computing Aplication in Analytical Chemistry**. Wiley, 2007. 321p.

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

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, p. 545-554, 1971.

PEARSON, K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. **Philosophical Transactions of the Royal Society of London, Section A**, v. 200, p. 1-66, 1903.

PETERNELLI L.A.; RESENDE M.D.V.; MENDES T.O.P. (2011) Experimentação e análise estatística em cana-de-açúcar. In Santos FA, Borém A and Caldas C (eds). **Cana-de-açúcar**. Editora UFV, Viçosa, p. 333-353.

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.

PÉREZ-ENCISO, M.; MISZTAL, I. Qxpak: A versatile mixed model application for genetical genomics and QTL analyses. **Bioinformatics**, v.20, p. 2792-2798, 2004.

PEREZ-ENCISO, M.; TORO, M. A.; TENENHAUS, M; GIANOLA, D. Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. **Genetics**, v. 164, p.1597-1606, 2003.

PEREZ-ENCISO, M.; VARONA, L. Quantitative trait loci mapping in F_2 crosses between outbred lines. **Genetics**, v.155, p.391–405, 2000.

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.

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage disequilibrium in humans: models and data. **American Journal of Human Genetics**, v. 69, p.1-14, 2001.

RAMALHO, M.A.P.; SANTOS, J.B. dos; PINTO, C.A.B.P. **Genética na agropecuária.** Lavras: UFLA, 2008. 463p.

RESENDE, M. D. V. de; OLIVEIRA, E. B.; HIGA, A. R. Utilização de índices de seleção no melhoramento do *Eucalyptus*. ***Pesquisa Florestal Brasileira***, Colombo, n. 21, p. 1-13, 1990.

RESENDE, M. D. V. de; HIGA, A. R.; LAVORANTI, O. J. Predição de valores genéticos no melhoramento de *Eucalyptus* – melhor predição linear (BLP). In: Congresso Florestal Brasileiro, 7, 1993, Curitiba. **Anais...**, Curitiba: SBS, 1993. p. 144-147.

RESENDE, M. D. V. de; HIGA, A. R. Maximização da eficiência da seleção em testes de progênies de *Eucalyptus* através da utilização de todos os efeitos do modelo matemático. ***Pesquisa Florestal Brasileira***, Colombo, v. 28/29, p. 37-55, 1994.

RESENDE, M. D. V. de; OLIVEIRA, E. B. DE; MELINSKI, L. C.; GOULART, F. S.; OAIDA, G. R. SELEGEN - ***Seleção Genética Computadorizada***: manual do usuário. Colombo: Embrapa Florestas, 1994. 31 p.

RESENDE, M. D. V. de; PRATES, D. F.; JESUS, A.; YAMADA, C. K. Estimação de componentes de variância e predição de valores genéticos pelo método da máxima verossimilhança restrita (REML) e melhor predição linear não viciada (BLUP) em *Pinus*. ***Pesquisa Florestal Brasileira***, Colombo, n.32/33, p.18-45, 1996.

RESENDE, M. D. V. de. Avanços da genética biométrica florestal. In: Bandel, G.; Vello, N. A.; Miranda Filho, J. B. (Ed.). **Encontro sobre temas de genética e melhoramento**: genética biometrica vegetal. ***Anais***. Piracicaba: Esalq/Usp, 1997. p.20-46.

RESENDE, M. D. V., FERNANDES, J. S. C., SIMEÃO, R.M. BLUP individual multivariado em presença de interação genótipos x ambientes para delineamentos experimentais repetidos em vários ambientes. **Revista de Matemática e Estatística**, v.17, p.209-228, 1999.

RESENDE, M. D. V. de; ROSA-PEREZ, J. R. H. **Genética Quantitativa e Estatística no Melhoramento Animal**. Curitiba: Imprensa Universitária - UFPR, 1999. 496 p.

RESENDE, M. D. V. **Inferência bayesiana e simulação estocástica (amostragem de Gibbs) na estimação de componentes de variância e valores genéticos em plantas perenes**. Colombo: Embrapa Florestas, 2000. 68 p.

RESENDE, M. D. V. **Análise estatística de modelos mistos via REML/BLUP no melhoramento de plantas perenes**. Colombo : Embrapa Florestas, 2000. 101 p.

RESENDE, M. D. V., STURION, J. A. **Análise genética de dados com dependência espacial e temporal via modelos geoestatísticos e de séries temporais via REML/BLUP**. Colombo : Embrapa Florestas, 2001. 79p.

RESENDE, M. D. V., DUDA, L. L., GUIMARÃES, P. R. B., FERNANDES, J. S. C. Análise de modelos lineares mistos via Inferência Bayesiana. **Revista de Matemática e Estatística**. , v.21, p.41-70, 2001.

RESENDE, M. D. V., REZENDE, G. D. S. P., FERNANDES, J. S. C. Regressão aleatória e funções de covariância na análise de medidas repetidas. **Revista de Matemática e Estatística**. , v.19, p.21-40, 2001.

RESENDE, M.D.V. **Genética Biométrica e Estatística no Melhoramento de Plantas Perenes**. Brasília: Embrapa Informação Tecnológica, 2002. 975p.

RESENDE, M. D. V., BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística**, v.20, p.30-65, 2002.

RESENDE, M. D. V., STURION, J. A. Análise estatística espacial de experimentos via modelos mistos individuais com erros modelados por processos ARIMA em duas dimensões. **Revista de Matemática e Estatística**, v.21, p.7-33, 2003.

RESENDE, M. D. V. de; THOMPSON, R. **Multivariate spatial statistical analysis of multiple experiments and longitudinal data**. Colombo: Embrapa Florestas, 2003. 126 p. (Embrapa Florestas. Documentos, 90).

RESENDE, M. D. V., THOMPSON, R. Factor analytic multiplicative mixed models in the analysis of multiple experiments. **Revista de Matemática e Estatística** , v.22, p.1-22, 2004.

RESENDE, M. D. V., STRINGER, J. K.; CULLIS, B. R.; THOMPSON, R. Joint modelling of competition and spatial variability in forest field trials. **Revista de Matemática e Estatística**, v.22, p.7 - 22, 2005.

RESENDE, M. D. V. de; THOMPSON, R.; WELHAM, S. Multivariate spatial statistical analysis of longitudinal data in perennial crops. **Revista de Matemática e Estatística** , v.24, p.147-169, 2006.

RESENDE, M. D. V. de. **Selegen–Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Colombo: Embrapa Florestas, 2007. 360 p.

RESENDE, M. D. V. de; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182-194, 2007.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos lineares mistos. In: **Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético**. Colombo: Embrapa Florestas, 2007. p. 517-534.

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

RESENDE, M. D. V. **Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais**. Colombo: Embrapa Florestas, 2008. 330 p.

RESENDE M. D. V.; 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 M. D. V.; SILVA, F. F.; VIANA, J. M. S.; PETERNELLI, L. A.; RESENDE JUNIOR, M. F. R.; VALLE, P.R.M. **Métodos estatísticos na seleção genômica ampla**. Colombo: Embrapa Florestas. 2011. 106 p.

RESENDE M. D. V.; SILVA, F. F.; VIANA, J. M. S. **I-BAYES-BLUP: Improved Bayesian BLUP method for estimating variance components and breeding values**. Viçosa, 2012 (a ser publicado).

RESENDE, M. D. V., RESENDE JR., M.F.R., SANSALONI, C.; PETROLI, C.; MISSIAGGIA, A. A.; AGUIAR, A. M.; ABAD, J.I.M.; TAKAHASHI, E.; ROSADO, A. M.; FARIA, D.; PAPPAS, G.; KILIAN, A.; GRATTAPAGLIA, D. Genomic Selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, v.194, p.116-128, 2012.

RESENDE JR., M. F. R. **Seleção genômica ampla no melhoramento vegetal**. UFV, 2010. 67 p. (Tese Mestrado).

RESENDE JR., M. F. R. Tecnologia RNA-Seq. **Comunicação pessoal**. 2012.

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**, v.190, p.1503 - 1510, 2012a.

RESENDE JR., M.F.R.; VALLE, P.R.M.; ACOSTA, J. J.; PETER, G. F.; DAVIS, J.M;, GRATTAPAGLIA, D.; RESENDE, M. D. V.; KIRST, M. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytologist**, v.193, p.617 - 624, 2012b.

RESENDE JR., M.F.R. ; ALVES, A.A.; SANCHES, C.F.B; RESENDE, M. D. V.; CRUZ, C.D. Seleção genômica ampla. In: CRUZ, C.D. et al. **Genômica Aplicada**. Viçosa: Editora Universitária, 2012c.

ROBERTSON, A. Prediction equations in quantitative genetics. ***Biometrics***, Washington, v. 11, p. 95-98, 1955.

ROBINSON, D. L. That BLUP is a good thing: the estimation of random effects. ***Statistical Science***, Hayward, v. 6, p. 15-32, 1991.

ROCHA, G. S. **Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais**. 2011. Dissertação (Estatística Aplicada e Biometria) - Universidade Federal de Viçosa.

RONNINGEN, K. Some properties of the selection index derived by "Henderson's mixed model method". ***Z. Tierz Zuchtungsbiol***, v. 88, p. 186, 1971.

ROSA, G.J.M.; ROCHA, L.B.; FURLAN, L.R. Estudos de expressão gênica utilizando-se *microarrays*: delineamento, análise, e aplicações na pesquisa zootécnica. **Revista Brasileria de Zootecnia**, v. 36, p.185-209, 2007.

ROSSI, P.E.; ALLENBY, G.M.; MCCULLOCH, R. **Bayesian Statistics and Marketing**. New York: Wiley Series in Probability and Statistics. 2005.

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

SCHAEFFER, L. R. Linear models. 1999. Disponível em: <www.http://cgil.uoguelph.ca/people/faculty/lschaeffer.html>. Acesso em: 15 jan. 2007.

SCHAEFFER, L. R.; DEKKERS, J. C. M. Random regressions in animal models for test-day production in dairy cattle. WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 5., 1994, Guelph. ***Proceedings...*** Guelph: University of Guelph, 1994, v.18, p.443.

SCHÄFER, J.; STRIMMER, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. **Stat. Appl. Genet. Mol. Biol.** 4: 32, 2005.

SCHUSTER, I.; CRUZ, C. D. **Estatística genômica** ***aplicada a populações derivadas de cruzamentos controlados***. Viçosa: Editora UFV, 2004. 568 p.

SILVA, F. F. e.; ROSA, G. J. M.; GUIMARÃES, S. E.F.; LOPES, P. S.; de los CAMPOS, G. Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. **Livestock Science**, v. 4, p. 1, 2011.

SILVA, F. F. e.; SÁFADI, T. ; MUNIZ, J. A.; ROSA, G. J. M.; AQUINO, L. H.; MOURÃO, G. B. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. **Pesquisa Agropecuária Brasileira**, v. 43, p. 37, 2008.

SILVA, F. F. e.; SÁFADI, T. ; MUNIZ, J. A.; ROSA, G. J. M.; AQUINO, L. H.; MOURÃO, G. B. ; SILVA, C. H. O. Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. **Scientia Agrícola**, v. 68, p. 237-245, 2011.

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**, v. 141, n. 1-3, p. 310-314, 2011.

SILVA, M. A. ; THIEBAUT, J. T. L.; VALENTE, B. D.; TORRES, R. A.; FARIA, F. J. C. **Modelos lineares aplicados ao melhoramento genético animal**. 1. ed. Belo Horizonte MG: PEPMVZ-Editora, 2008. 378 p.

SIMEAO, R. M.; CASLER, M.D.; RESENDE, M. D. V. Genomic selection in forage breeding: designing an estimation population. In: Plant and Animal Genome Conference XXI - 2013, San Diego. Abstracts of the Plant and Animal Genome Conference XXI, 2013.

SINGER, J. M.; STANEK, E. J.; LENCINA, V. B.; GONZÁLEZ, 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. Disponível em: <http://www.gsejournal.org/content/41/1/29>. Acesso em 30/10/2010.

SOLBERG, T. R.; SONESSON, A.; WOOLIAMS, J.; MEUWISSEN, T. H. E. Genomic selection using different marker types and density. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006. **Proceedings**. Belo Horizonte: Ed. da UFMG, 2006. 1 CD-ROM.

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

STEIN, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 3., 1955, Berkeley. **Proceedings**... Berkeley: University of California Press, 1955. p. 197-206, 1955.

STONE, M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39**, 44-47, 1977.

STRAM, D. O.; LEE, J. W. Variance components testing in longitudinal mixed effects setting. **Biometrics**, v. 50, p. 1171-1177, 1994.

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

TAKAHASHI, K.; FAGAN, J.; CHIN, M. S. Formation of a sparse bus impedance matrix and its application to short circuit study. In: Institutional Pica Conference, 8, 1973. **Proceedings** Minneapolis: IEEE Power Engineering Society, 1973. p.63.

TAL, O.; KISDI, E.; JABLONKA, E. Epigenetic contribution to covariance between relatives. **Genetics**, v. 184, p. 1037-1050, 2010.

TEMPELMAN, R.J. Assessing statistical precision, power and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. **Veterinary Immunology and Immunopathology**, v.105, p. 175-186, 2005.

TENESA, A.; NAVARRO, T.; HAYES, B. J.; DUFFY, D. L.; CLARKE, G. M.; GODDARD, M. E.; VISSCHER, P. M. Recent human effective population size estimated from linkage disequilibrium. **Genome Research**, v. 17, p.520-526, 2007.

THOMPSON, R. Iterative estimation of variance components for non-orthogonal data. **Biometrics**, v. 25, p. 767-773, 1969.

THOMPSON, R. Relationship between the cumulative difference and best linear unbiased predictor methods of evaluating bulls. **Animal Production**, v. 23, p. 15-24, 1976.

THOMPSON, R. Sire evaluation. **Biometrics**, v. 35, p. 339-353, 1979.

THOMPSON, R. The estimation of heritability with unbalanced data. **Biometrics**, v. 33, p. 485-504, 1977.

THOMPSON, R. The estimation of variance and covariance components when records are subject to culling. **Biometrics**, v. 29, p. 527-550, 1973.

THOMPSON, R. A review of genetic parameter estimation. In: WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**. Paris: INRA, 2002. p. 19-23.

THOMPSON, R.; CULLIS, B. R.; SMITH, A. B.; GILMOUR, A. R. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. **Australian and New Zealand Journal of Statistics**, v. 45, n. 4, p. 445-459, 2003.

THOMPSON, R; WRAY, N. R.; CRUMP, R. E. Calculation of prediction error variances using sparse matrix methods. **Journal of Animal Breeding and Genetics**, v. 111, p. 102-109, 1994.

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

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

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

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

VARONA, L. ***Aplicaciones del muestreo de Gibbs en modelos de genética cuantitativa*: analisis de** un caso de heterogeneidad de varianzas. Zaragoza: Universidad de Zaragoza, 1994. PhD. Thesis.

VARONA, L.; MORENO, C.; GARCIA-CORTES, L. A.; ALTARRIBA, J. Estimación multicarácter de componentes de varianza y covarianza en vacuno lechero mediante muestreo de Gibbs. ***Revista Portuguesa de Zootecnia***, v. 1, p. 185-195, 1994.

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 of parent average for several traits in US Holsteins. **Journal of Dairy Science**, Champaign, v. 93, n. 1, p. 5942-5949. DOI: 10.3168/jds.2010-3335.

VENCOVSKY, R.; BARRIGA, P. ***Genética biométrica no fitomelhoramento***. Ribeirão Preto: Sociedade Brasileira de Genética, 1992. 486 p.

VENCOVSKY, R.; CROSSA, J. Variance effective population size under mixed self and random mating with applications to genetic conservation of species. ***Crop Science***, v. 39, p. 1282-1294, 1999.

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

VIANA, J. M. S.; FARIA, V. R.; SILVA, F. F. ; RESENDE, M. D. V. Combined selection of progeny in crop breeding using best linear unbiased prediction. **Canadian Journal of Plant Science**, v. 92, p. On line 1st-doi:10.4141/CJP, 2012.

VIANA, J. M. S. ; LIMA, R. O.; FARIA, V. R.; MUNDIM, G. B. ; RESENDE, M. D. V. ; SILVA, F. F. . Relevance of pedigree, historical data, dominance, and data unbalance for selection efficiency. **Agronomy Journal** (Print), v. 104, p. 722-728, 2012.

VIANA, J. M. S. ; FARIA, V.; SILVA, F. F. ; RESENDE, M. D. V. Best linear unbiased prediction and family selection in crop species. **Crop Science**, v. 51, p. 2371-2381, 2011.

VIANA, J. M. S.; ALMEIDA, R. V.; FARIA, V. R.; RESENDE, M. D. V.; SILVA, F. F. Genetic evaluation of inbred plants based on BLUP of breeding value and general combining ability. **Crop & Pasture Science**, v. 62, p. 515-522, 2011.

VIANA, J. M. S.; VALENTE, M. S. F.; SCAPIM, C. A.; RESENDE, M. D. V. ; SILVA, F. F. Genetic evaluation of tropical popcorn inbred lines using BLUP. **Maydica** (Bergamo), v. 56, p. 273-281, 2011.

VIANA, J. M. S.; ALMEIDA, Í. F.; RESENDE, M. D. V.; FARIA, V. R.; SILVA, F. F. BLUP for genetic evaluation of plants in non-inbred families of annual crops. **Euphytica** (Wageningen), v. 174, p. 31-39, 2010.

VIANA, J. M. S. ; SOBREIRA, F. M. ; DE RESENDE, M. D. V. ; FARIA, V. R. Multi-trait BLUP in half-sib selection of annual crops. **Plant Breeding**, v. 129, p. 599-604, 2010.

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, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. **PLoS Genetics**, v.2, n.3, e41, 2006.

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.

VLECK, L. D; CASSADY, J. P. Unexpected estimates of variance components with a true model containing genetic competition effects. **Journal of Animal Science**, v. 83, p. 68-74, 2005.

WANG, T.; FERNANDO, R. L.; GROSSMAN, M. Genetic evaluation by best linear unbiased prediction using marker and trait information in a multibreed population. **Genetics**, v. 148, p. 507-515, 1998.

WELLER, J. I. **Quantitative trait loci analysis in animals**. London: CABI Publishing, 2001. 287 p.

WELLER, J. I.; SHLEZINGER, M.; RON, M. Correcting for bias in estimation of quantitative trait loci effects. **Genetics Selection Evolution**, v. 37, p. 501-522, 2005.

WELLER, J.L. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. **Biometrics**, v.42, p.627-640, 1986.

WHITE, I. M. S.; THOMPSON, R.; BROTHERSTONE, S. Genetic and environmental smoothing of lactation curves with cubic splines. **Journal of Dairy Science**, v. 82, p. 632-638, 1999.

WHITTAKER, J.C.; THOMPSON, R.; DENHAM, M.C. Marker assisted selection using ridge regression. **Genetical Research**, 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.

WOLFINGER, R.D.; GIBSON, G;. WOLDINGER, E.D. *et al.* Assessing gene significance form cDNA microarrayexpression data via mixed models. **Journal of Computational Biology**, v.8, n.6, p. 625-637, 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**, v. 88, p. 76-82, 2011.

YANG, W; TEMPELMAN, R.J. A Bayesian antedependence model for whole genome prediction. **Genetics, 2012.**

YATES, F. A new method of arranging variety trials involving a large number of varieties. **Journal of Agricultural Sciences**, v. 26, p. 424-455, 1936.

YATES, F. The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association**, v. 29, p. 51-66, 1934.

ZEGER, S. L.; LIANG, K. Y.; ALBERT, P. S. Models for longitudinal data: a generalized estimation approach. **Biometrics**, v. 44, p. 1049-1060, 1988.

ZENG, Z. Precision mapping of quantitative loci. **Genetics**, v.136, p.1457-1468, 1994.

ZENGER, K.R.; KHATKAR, M. S.; CAVANAGH, J. A.; HAWKEN, R. J.; RAADSMA, H. W. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian global population variability, including impact of selection. **Animal Genetics**. v.38, p.7-14, 2007.

ZHAO, H.; NUTTLETON, D.; SOLLER, M.; DEKKERS, J. C. M. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. **Genetical Research**, v. 80, p. 77-97, 2005.

ZOLLENKOPF, K. Bi-Factorisation - Basic computational algorithm and programming techniques. In: REID, J. K. (Ed.). **Large sparse sets of linear equations**. London: Academic Press, 1971. p. 75-96.

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.

9 Fotos de Pesquisadores com Participação Relevante na Evolução dos Métodos Estatísticos de Avaliação Genética

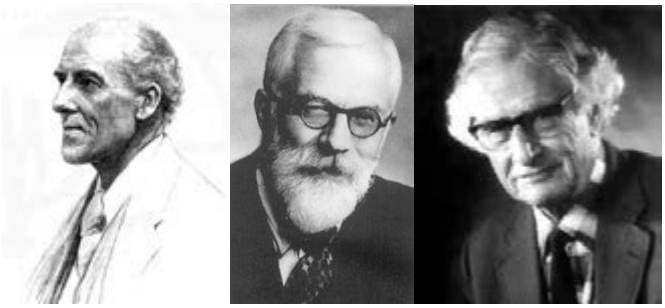


Foto 1: Criadores da Estatística e Biometria na Inglaterra (Rothamsted): Karl Pearson à esquerda; Ronald Fisher, ao centro e Frank Yates à direita.

Foto 2: Criador e especialista no uso do método do índice de seleção na estimação de valores genéticos (Jay Lush, à esquerda; Dale Van Vleck, à direita).

Foto 3: Criadores e difusores do uso dos métodos BLUP (Charles Henderson, à esquerda) e Bayesianos (Daniel Gianola, ao centro) na estimação de valores genéticos; à direita, Richard Quaas, especialista em matriz de parentesco e modelo animal.

Foto 4: Criadores do procedimento de seleção genômica ampla (GWS) (Theo Meuwissen, à esquerda, Mike Goddard, ao centro e Ben Hayes à direita).

Foto 5: Criador do método REML e método numérico de Informação Média (AI) (Robin Thompson, à esquerda) e autor do software ASREML (Arthur Gilmour, ao centro); John Nelder, criador da técnica GLMM e do software Genstat em Rothamsted (à direita).

Foto 6: Autores do software ASREML: Arthur Gilmour, ao centro, Brian Cullis, à esquerda e Robin Thompson, à direita.

Foto 7: Criadores do método de análise espacial via processos ARIMA separáveis em duas direções (AR1⊗AR1) e autores do software ASREML: Arthur Gilmour, à esquerda e Brian Cullis, à direita.

Foto 8: Larry Schaeffer, um dos pioneiros dos modelos de regressão aleatória multivariada e Raphael Mrode, autor de compreensivo livro sobre modelos mistos e regressão aleatória.

Foto 9: Rohan Fernando, pioneiro na análise de ligação gênica via BLUP e um dos pioneiros (juntamente com Daniel Gianola) no uso da Inferência Bayesiana no melhoramento genético; Ignacy Misztal, ícone dos métodos computacionais no melhoramento genético; Miguel Pérez-Enciso, autor do software QxPack.

Foto 10: Ícones dos Modelos Lineares e Componentes de Variância: Shayle Searle, David Harville e Jean Louis Foulley, respectivamente.

Foto 11: Robert Tibshirani, Gustavo de los Campos e Andrés Legarra: autores dos métodos LASSO, BLASSO e IBLASSO, respectivamente.

Foto 12: Pioneiros e Ícones da Genética de Populações: Sewall Wright , Alan Robertson, Gustave Malecot e William Hill, respectivamente.

Foto 13: Ícones da Genética Quantitativa Humana: Robert Elston, Kenneth Lange, Peter Visscher e Naomi Wray, respectivamente.

Foto 14: Roland Vencovsky e Martinho Almeida e Silva: pioneiros e expoentes da Genética Quantitativa Vegetal e Animal no Brasil, respectivamente; Newton Freire-Maia, pioneiro da Genética de Populações Humanas no Brasil.