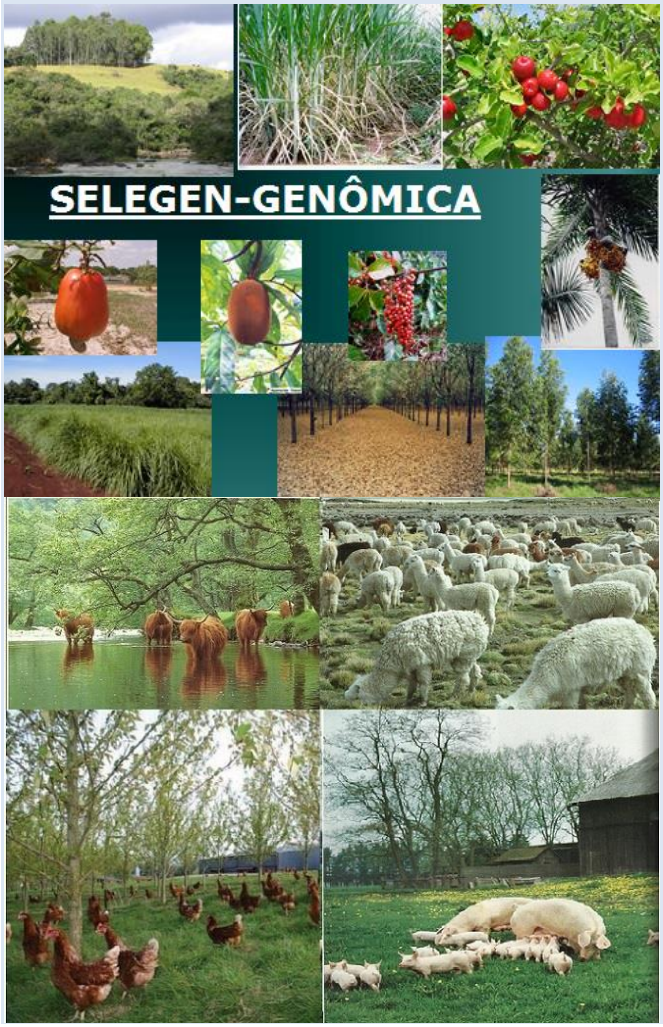
Selegen Genômica – Software para Seleção Genômica Ampla (GWS)

$(\mathrm{M}\alpha\tau\zeta\theta\int\partial\varepsilon\theta\eta$, *Maio – 2014)*

Marcos Deon Vilela de Resende

Universidade Federal de Viçosa
Centro de Ciências Exatas
Departamento de Estatística
Estatística Aplicada e Biometria



Sumário

Selegen Genômica – Software para Seleção Genômica Ampla (GWS)

(Matzθ∫ ∂εθη , Abril – 2014)

Marcos Deon Vilela de Resende

Universidade Federal de Viçosa
Centro de Ciências Exatas
Departamento de Estatística
Estatística Aplicada e Biometria

Resumo

A presente nota reporta sobre o software Selegen Genômica desenvolvido para uso na GWS. O software fundamenta-se no procedimento REML/BLUP e contempla cerca de vinte modelos para GWS e GWAS e destina-se apenas ao ensino e à pesquisa em pequena escala e não ao uso na prática do melhoramento genético associado a uma grande massa de dados. Por não ter o objetivo de uso comercial, não foi investido em capacidade de processamento (velocidade na situação de grande massa de dados) do software. Entretanto, pode ser usado eficientemente nas seguintes situações práticas: (i) início de estudos de seleção genômica, caso em que se tem um número limitado de indivíduos (até 1.000) e de marcadores (até 20.000, sendo que este número pode ser maior à medida que o número de indivíduos torna-se menor que 1000); (ii) fase final de estudos de seleção genômica, caso em que já se tem um número limitado de genes já identificados (gene assisted selection – GAS); (ii) fase de teste no início do uso de outros softwares mais potentes, em que um exemplo pequeno é avaliado no Selegen Genômica e no outro software, visando descobrir que tipo de modelo e parametrização o software alternativo usa e também informar se o usuário está usando adequadamente o referido software. Para fins didáticos o software é bastante completo pois, contempla as principais análises e modelos necessários à prática do melhoramento. Os modelos permitem incluir em nível genômico: efeitos aditivos, dominância, epistasia, interação genótipos x ambientes, repetibilidade, herdabilidades genômicas, delineamentos em blocos incompletos, correlações genômicas, índice de seleção genômica, componentes principais genômicos, divergência genômica, agrupamento genômico, acurácia genômica, análise de deviance genômica, validação cruzada, valores genômicos, heterozigose individual, correlações envolvendo valores aditivos, de dominância, fenotípicos e heterozigoses dos indivíduos. O Selegen Genômica realiza também a análise em Genética de Populações de marcadores fornecendo: frequências alélicas, heterozigoses de marcas, frequência do alelo menos frequente (MAF), *call rate*, coeficiente de endogamia (F) e tamanho efetivo populacional (Ne). A análise estatística em geral tem três grandes objetivos: a estimação de componentes de médias, a estimação de componentes de variância e a realização de testes de hipóteses. Esse software abarca os três, sendo este último realizado via seleção de modelos baseado no critério da deviance. Diferentes valores de deviance são obtidos pelos diferentes softwares e algoritmos, sendo que diferem por uma constante que não afeta a maximização da função de verossimilhança restrita. Entretanto, a diferença entre deviances para realização do LRT permanece inalterada.

1. Introdução

Os estudos de associação genômica ampla (GWAS) e seleção (ou estimação) genômica ampla (GWS) são importantes no melhoramento genético de animais e plantas e também na genética humana. No melhoramento genético a GWS aumenta a eficiência e rapidez do processo seletivo. Em genética humana, as ferramentas da GWS propiciam a medicina personalizada ou medicina genômica, a qual fundamenta-se na predição de fenótipos com base na leitura de genótipos marcadores e uso de métodos preditivos. As predições geradas são usadas na prevenção, diagnose e tratamento das doenças (Resende et al., 2012).

Com a GWS, 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 (Resende et al., 2010).

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, 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) e vem sendo recomendada também para o melhoramento de plantas anuais (Heffner et al., 2009), de espécies florestais (Resende et al. 2008, 2012; Grattapaglia e Resende, 2011; Resende Jr. et al, 2012), de fruteiras (Cavalcanti et al., 2011) e de animais (VanRaden et al., 2009; Silva et al., 2011).

A superioridade da GWS sobre a seleção baseada em fenótipos pode ser atribuída a cinco fatores (Resende et al., 2012): (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.*

O Selegen Genômica contempla alguns procedimentos de GWS, teve seu início em 2007 e tem sido usado para o desenvolvimento de teses e artigos científicos na UFV e na Embrapa, em algumas situações de pequena dimensão de dados (Resende, 2007; Resende et al., 2010, 2012; Fritsche Neto et al., 2012a e b; Rocha, 2011; Cavalcanti et al., 2011; Oliveira et al. 2012). A seguir esse software é descrito brevemente.

2. Selegen Genômica - Modelos

O software Selegen Genômica contempla pelo menos quinze procedimentos, conforme a Tabela 1. O nome deriva do software Selegen-REML/BLUP, desenvolvido a partir de 1993, em que Selegen deriva do nome Seleção Genética. No contexto da GWS, um software com nome similar é o GenSel (Fernando and Garrick, 2009), cujo nome deriva do termo Genomic Selection.

Selegen Genômica

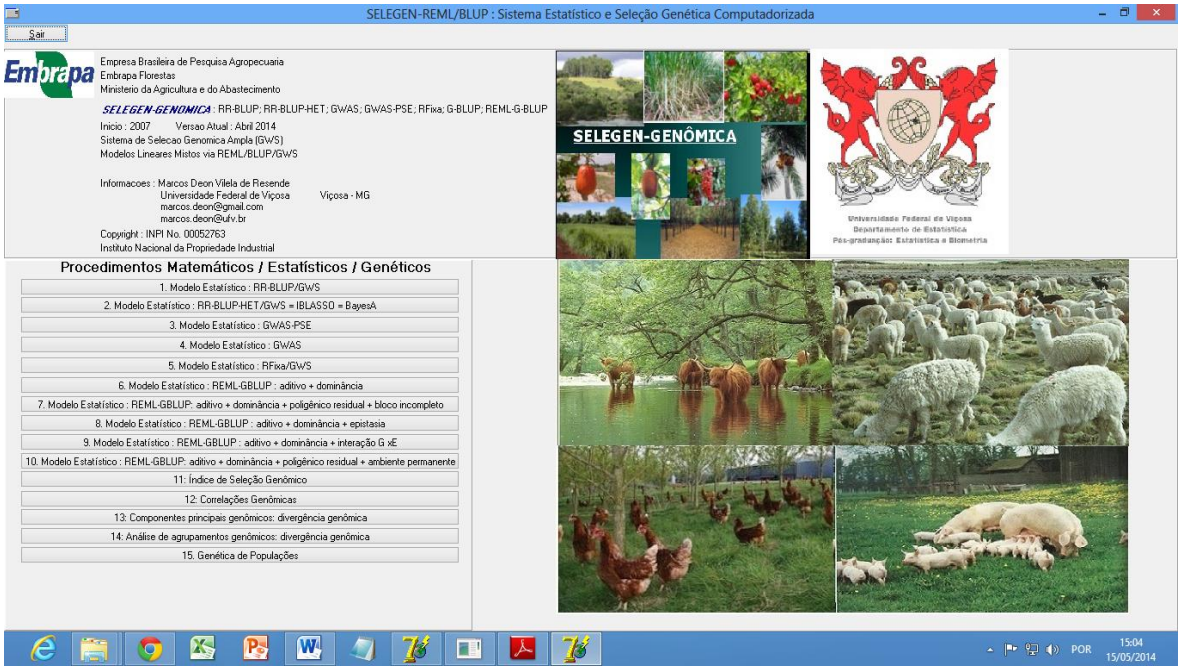


Tabela 1. Modelos do Selegen Genômica para a GWS e GWAS.

Modelo no Selegen e Método*	Efeitos de Marcas/Indiv	Objetivo	Penalização λ	Arquivo de Lambdas	Arquivo de Marcas
1 RR-BLUP	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_{gc}^2$	Não necessário	No mesmo arquivo de fenótipos
2 RR-BLUP-HET	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	Lambda de cada marca na coluna	No mesmo arquivo de fenótipos
3 GWAS-PSE	Aleatório	GWAS	$\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$	Lambda de cada marca na coluna	No mesmo arquivo de fenótipos
4 GWAS	Fixo	GWAS	$\lambda = 0$	Zeros na coluna 2	No mesmo arquivo de fenótipos
5 FR-LS	Fixo	GWS	$\lambda = 0$	Não necessário	No mesmo arquivo de fenótipos
6 G-REML/G-BLUP: aditivo + dominância	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	Não necessário	Arquivos separados
7 G-REML/G-BLUP: aditivo + dominância + poligênico + látice	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	Não necessário	Arquivos separados
8 G-REML/G-BLUP: aditivo + dominância + epistasia	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	Não necessário	Arquivos separados
9 G-REML/G-BLUP: aditivo + dominância + G x E**	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	Não necessário	Arquivos separados
10 G-REML/G-BLUP: aditivo + dominância + poligênico + permanente (repetibilidade)	Aleatório	GWS	$\lambda = \sigma_e^2 / \hat{\sigma}_g^2$	Não necessário	Arquivos separados
11 Índice de Seleção Genômica	-	GWS	-	-	-
12 Correlação Genômica entre Caracteres	-	GWS	-	-	-

13 Divergência Genômica via Componentes Principais	-	GWS	-	-	-
14 Divergência Genômica e Agrupamento	-	GWS	-	-	-
15. Genética de Populações	-	GWAS e GWS	-	-	-

* RF - Regressao fixa. RR - Regressao Aleatória. ** Demanda um terceiro arquivo.

Trabalhando-se com esses modelos, outros modelos também podem ser ajustados, conforme a Tabela 2, por meio da especificação de valores zero para alguns coeficientes de determinação (c^2) na tela do Selegen.

Tabela 2. Outros Modelos Ajustados pelo Selegen Genômica.

Modelo no Selegen e Método*	Coeficientes a serem Modificados	Modelo a ser Usado
16 G-REML/G-BLUP: aditivo	h2g = 0.10	6
17 G-REML/G-BLUP: aditivo + permanente (repetibilidade)	h2g = 0.10; c21 = 0	10
18 G-REML/G-BLUP: aditivo + poligênico	h2g = 0.10; c2 = 0	7
19 G-REML/G-BLUP: aditivo + epistasia	h2g = 0.10	8
20 G-REML/G-BLUP: aditivo + G x E**	h2g = 0.10	9
21 G-REML/G-BLUP: aditivo + poligênico + permanente (repetibilidade)	h2g = 0.10	10

As seguintes análises são necessárias em seleção genômica:

$$\frac{Seleção}{Genômica} = \begin{cases} 1.\ Correção\ \ de\ Fenótipos \\ 2.\ Controle\ \ de\ Qualidade\ \ das\ Marcas - Genética\ \ de\ Populações \\ 3.\ REML/G-BLUP: \hat{h}^2\ \ e\ \tilde{g}\ -\ Catá\log o\ de\ Valores\ \ Genéticos\ de\ Indivíduos \\ 4.\ RR-BLUP: \tilde{m}\ usando\ \hat{h}^2 - Catá\log o\ de\ Valores\ \ Genéticos\ de\ Marcas \\ 5.\ Validação\ \ Cruzada: Jackknife,\ Capacidade\ \Pr editiva,\ Viés \\ 6.\ GWAS: FR-BLUE: Significância\ de\ \tilde{m} \\ 7.\ ANADEV: Seleção\ \ de\ Modelos\ \ Genômicos\ via\ Deviance \\ 8.\ ACURÁCIA\ GENÔMICA \\ 9.\ ANÁLISE\ \ MULTIVARIADA\ GENÔMICA: Correlaçõe s,\ Índices,\ Divergência,\ Agrupamento \end{cases}$$

Recomenda-se rodar os softwares na seguinte sequência:

$$\frac{Selegen}{REML/BLUP} = \begin{cases} 1.\ Correção\ \ de\ Fenótipos: \\ \tilde{e} = y - X\hat{u} - T\tilde{c} \rightarrow Valores\ \ genéticos\ desregressados \\ \tilde{e} = y - X\hat{u} - T\tilde{c} - Z\tilde{f} \rightarrow Valores\ \ genéticos\ desregressados\ e\ corrigidos\ \ para\ \ famílias \end{cases}$$

$$\frac{Selegen}{Genômica} = \begin{cases} 2.\ Controle\ \ de\ Qualidade\ \ das\ Marcas - Genética\ \ de\ Populações \rightarrow Modelo\ 15 \\ 3.\ REML/G-BLUP: \hat{h}^2\ \ e\ \tilde{g}\ -\ Catá\log o\ de\ Valores\ \ Genéticos\ de\ Indivíduos \rightarrow Modelos\ 6\ a\ 10 \\ 4.\ RR-BLUP: \tilde{m}\ usando\ \hat{h}^2 - Catá\log o\ de\ Valores\ \ Genéticos\ de\ Marcas \rightarrow Modelos\ 1\ a\ 5 \\ 5.\ Validação\ \ Cruzada: Jackknife \rightarrow Modelos\ 1\ a\ 3 \\ 6.\ GWAS: FR-BLUE: Significância\ de\ \tilde{m} \rightarrow Modelos\ 3\ e\ 4 \\ 7.\ ANADEV: Seleção\ \ de\ Modelos\ \ Genômicos\ via\ Deviance: c^2 = 0 \rightarrow Modelos\ 6\ a\ 10 \\ 8.\ ACURÁCIA\ \ GENÔMICA: Arquivos\ .fam \\ 9.\ ANÁLISE\ MULTIVARIADA: GENÔMICA\ Correlaçõe s,\ Índices,\ Divergência,\ Agrupamento \rightarrow Modelos\ 11\ a\ 14 \end{cases}$$

3. Arquivos de Dados

Para os modelos 1 a 5 o arquivo de dados deve conter a seguinte seqüência de colunas: *Indivíduo Família Bloco Planta Fenótipos Marcas*. As colunas Família, Bloco e Planta podem ser preenchidas com o numero 1, pois são ignoradas na análise (isto vale também para os modelos 6, 8 e 9). Alguns modelos exigem adicionalmente um arquivo de lambdas dados por $\lambda = \sigma_e^2 / \hat{\sigma}_{gi}^2$, em que σ_e^2 e a estimativa da variância residual e $\hat{\sigma}_{gi}^2$ 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. (2011a). Para os modelos 1 e 2 é usada a opção BLUP e deve ser fornecida a herdabilidade da característica.

Os modelos 7 a 10 exigem pelo menos dois arquivos: o de dados que deve conter a seguinte seqüência de colunas: *Indivíduo Família Bloco Planta Fenótipos* e o de marcas com as seguintes colunas: *Indivíduo Marcas*. Indivíduos genotipados mas não fenotipados devem ter seus fenótipos imputados como sendo a média geral dos fenótipos. Para o modelo 9, o arquivo com a informação de parentesco da interação G x E (Resende et al., 2012), deve ser fornecido à parte, contendo apenas colunas referentes aos marcadores e uma primeira linha de identificação, a qual será ignorada pelo programa. O card em R (desenvolvido com a ajuda de Camila Ferreira Azevedo) para gerar esse arquivo é apresentado a seguir. O arquivo de marcas a ser lido pelo R deve conter as seguintes colunas: *Local Indivíduo Marcas*.

```
setwd("C:\\teste")

dados=read.table("TESTE-GWS-AD-MARCAS-AE.txt",h=T)

M1=dados[dados[,1]==1,-c(1:2)]
M2=dados[dados[,1]==2,-c(1:2)]

# Para o local 1

M1A=matrix(0,nrow(M1),ncol(M1))
M1A[M1==2]=1
M1A[M1==1]=0
M1A[M1==0]=-1

# Para o local 2

M2A=matrix(0,nrow(M2),ncol(M2))
M2A[M2==2]=1
M2A[M2==1]=0
M2A[M2==0]=-1

G_AE1=(M1A%*%t(M1A))/sum(diag((M1A%*%t(M1A)))/nrow(M1A))
G_AE2=(M2A%*%t(M2A))/sum(diag((M2A%*%t(M2A)))/nrow(M2A))

# Matriz G_AE

G_AE=rbind(cbind(G_AE1,matrix(0,nrow(G_AE1),ncol(G_AE2))),
cbind(matrix(0,nrow(G_AE2),ncol(G_AE1)),G_AE2))

write.table(G_AE,"G_AE.txt",col.names=FALSE,row.names=FALSE,quote=FALSE)
write.table(G_AE,"G_AE1.txt",col.names=FALSE,row.names=FALSE,quote=FALSE,sep=" ") # delimitado por espaço no txt
write.table(G_AE,file="G_AE2.csv",col.names=FALSE,row.names=FALSE,quote=FALSE,sep=" ") # delimitado por espaço
no excel
```

O modelo 9 pode ser usado também para ajustes usando outras matrizes de correlação, em lugar da matriz de parentesco da interação G_AE. Assim, podem ser usados, por exemplo, o ajuste de efeitos epigenéticos (via uso da matriz T de transmissibilidade epigenética em lugar de G_AE), o ajuste de efeitos poligênicos residuais (via uso da matriz A de parentesco genético aditivo em lugar de G_AE). A matriz é fornecida pelo modelo 184 do Selegen-Reml/Blup, mediante fornecimento do pedigree.

4. Resultados e seus Significados

As estimativas dos componentes de variância e outros parâmetros genéticos associados aos diferentes modelos são simbolizados por:

Va: variância genética aditiva.

Vd: variância genética de dominância.

Vi: variância genética epistática.

Vpedfam: variância poligênica residual entre famílias, ou variância genética entre famílias, capturada pelo pedigree, mas não pelos marcadores.

Vint: variância dos efeitos da interação genético aditivo x ambiente.

Vbloc: variância entre blocos.

Vperm: variância de ambiente permanente entre indivíduos.

Ve: variância ambiental.

Vf: variância fenotípica individual.

h2a = h2: herdabilidade individual no sentido restrito, ou seja, dos efeitos aditivos.

h2g: herdabilidade individual no sentido amplo, ou seja, dos efeitos genotípicos totais. É calculado pelo Selegen como h2a + c2d. Mas no modelo 8 deve-se somar também c2i.

c2d: coeficiente de determinação dos efeitos de dominância.

c2i: coeficiente de determinação dos efeitos epistáticos.

c2int: coeficiente de determinação dos efeitos da interação genético aditivo x ambiente.

c2pedfam = c21: coeficiente de determinação dos efeitos poligênicos residuais entre famílias.

c2bloc = c2: coeficiente de determinação dos efeitos de bloco.

c2perm: coeficiente de determinação dos efeitos de ambiente permanente entre indivíduos.

r = repetibilidade individual.

h = heterozigose individual.

rhf = Correlação entre heterozigose individual e fenótipo.

Para a GWAS é mostrado o resultado ao nível de significância 5% pelo teste F.

Após o controle de qualidade do arquivo de marcas realizado pelo modelo 15, o Selegen Genômica grava automaticamente um arquivo com extensão _cor onde são

eliminadas as marcas com MAF < 0.05 e/ou CALL rate < 0.95 e também realiza-se a imputação de marcas perdidas.

5. Teoria e Modelos

Modelos análogos à maioria daqueles usados na Genética Quantitativa Clássica foram programados no Selegen Genômica. Tais modelos incluem, ao nível genômico, os seguintes efeitos e parâmetros: efeito aditivo, dominância, epistasia, interação genótipo x ambiente, efeito poligênico residual, herdabidade no sentido restrito , herdabidade no sentido amplo, repetibilidade, correlação genômica entre caracteres, índice de seleção genômico, correlação genômica através de locais, blocos incompletos, divergência genômica e agrupamento. A seguir são apresentados os modelos programados no Selegen Genômica.

Modelos Fenotípicos

Um modelo adequado para descrever bem os fenótipos medidos a campo contempla os efeitos ambientais fixos (u), ambientais aleatórios ou de ambiente comum (c, podendo ser parcelas, blocos, ambiente permanente) e os efeitos genéticos aleatórios capturados pelos marcadores (g) e poligênicos residuais (f) associados ao pedigree ou estrutura de famílias e não capturados pelos marcadores. O modelo é especificado por:

$$y = Xu + Zg + Zf + Tc + e$$

Os efeitos genéticos g podem ser decompostos em aditivos (a), de dominância (d), epistáticos (i) e da interação aditivo x ambientes (ae). O modelo torna-se então:

$$y = Xu + Za + Zd + Zi + Zae + Zf + Tc + e$$

, em que X, Z e T são matrizes de incidência. Uma vez que o modelo contém um grande número (6) de fatores de efeitos aleatórios é necessário escolher aqueles de maior relevância em cada situação prática, de forma a evitar a super-parametrização na modelagem. Por exemplo, um modelo relevante na maioria das situações é:

$$y = Xu + Za + Zd + Zf + Tc + e.$$

Esse modelo está implementado nos modelos 7 e 10 do Selegen Genômica e também no software GS3 de Legarra et al. (2011) e é um procedimento single step de ajuste simultâneo dos efeitos genéticos e ambientais. O GS3 usa o enfoque Bayesiano e adota um modelo equivalente em nível de marcas e não de indivíduos.

Se os dados forem pré-ajustados para alguns efeitos, tem-se o ajuste em dois ou mais passos e os modelos tornam-se:

(i) ajuste apenas para os efeitos ambientais $y_c = \tilde{e} = y - X\hat{u} - T\tilde{c}$

$$y_c = Ju^* + Za + Zd + Zi + Zae + Zf + e$$
$$y_c = Ju^* + Za + Zd + Zi + Zf + e$$

$y_c = Ju^* + Za + Zd + Zge + Zf + e$
$y_c = Ju^* + Za + Zd + Zf + e$ *(modelo 7 e 10 do Selegen Genômica com* $c^2 = 0$*)*

Na notação, J refere-se a um vetor de uns, associado ao escalar u* referente à média geral. O interesse nesse tipo de ajuste reside na aplicação da GWS no curto prazo.

(ii) ajuste para os efeitos ambientais e de família $y_c = \tilde{e} = y - X\hat{u} - T\tilde{c} - Z\tilde{f}$

$y_c = Ju^* + Za + Zd + Zi + Zae + e$
$y_c = Ju^* + Za + Zd + Zi + e$ *(modelo 8 do Selegen Genômica)*
$y_c = Ju^* + Za + Zd + Zae + e$ *(modelo 9 do Selegen Genômica)*
$y_c = Ju^* + Za + Zd + e$ *(modelo 6 do Selegen Genômica)*

O interesse nesse tipo de ajuste reside na aplicação da GWS no longo prazo e também na GWAS.

Fenótipos Corrigidos

Para uso dos modelos descritos na situação (i) os fenótipos devem ser corrigidos para os efeitos ambientais ou os valores genéticos preditos pelo BLUP tradicional devem ser desregressados. Em ambos os casos obtém-se $\tilde{e}$, o vetor de fenótipos corrigidos.

A correção a ser efetuada é dada por $\tilde{e} = y - X\hat{u} - T\tilde{c}$. Nesse caso, u e c devem ser estimados sob um modelo em que se ajusta também o g. Uma maneira de se fazer isso no Selegen Reml/Blup é por meio do ajuste do modelo com u, g e c, via REML/BLUP e, posteriormente, usar somente a opção BLUP fixando a herdabilidade h^2 em 1 e fixando o coeficiente de determinação c^2 em seu valor estimado por REML no passo anterior. Assim, os valores genéticos produzidos por esse último BLUP são os valores fenotípicos corrigidos a entrarem na seleção genômica.

Pela alternativa de desregressão do BLUP tradicional, tem-se $\tilde{e} = \tilde{g} / \hat{r}_g^2$, em que $\hat{r}_g^2$ é a confiabilidade (quadrado da acurácia) da predição tradicional de g.

Na situação (i) o interesse reside na aplicação da GWS no curto prazo (uma ou duas gerações).

Na situação (ii) o interesse reside na aplicação da GWS no longo prazo (uma ou duas gerações) e a correção dos fenótipos para os efeitos de genitores ou famílias deve ser dado $\tilde{e} = y - X\hat{u} - T\tilde{c} - Z\tilde{f}$ em que u, c e f devem ser estimados na presença de g. Para fazer isso no Selegen Reml/Blup é por meio do ajuste do modelo com u, g, c e f, via REML/BLUP e posterior uso dos resíduos do arquivo .dev, os quais já equivalem a $\tilde{e} = y - X\hat{u} - T\tilde{c} - Z\tilde{f}$, que é denominado vetor de efeitos da segregação mendeliana desregressada.

Quando não se dispõe do arquivo de fenótipos mas, apenas de um catálogo

com valores genéticos preditos e sua acurácia, o procedimento de desregressão do BLUP tradicional deve ser usado, conforme descrito por Garrick et al. (2009) e Resende et al. (2010). Neste caso, demanda-se uma trinca de informações referentes aos vetores de valores genéticos preditos ($\tilde{g}$) e suas acurácias ($\hat{r}_g$), referentes a três entidades: indivíduos i, pais j e mães k. Esse procedimento propicia $\tilde{g}_i = (y - X\hat{u} - T\tilde{c} - 0{,}5\ \tilde{g}_j - 0{,}5\ \tilde{g}_k)$, que é similar a $\tilde{e} = y - X\hat{u} - T\tilde{c} - Z\tilde{f}$.

Alternativa de Correção para Estrutura de Família

Um procedimento muito usado em GWAS para correção para estrutura de família é o ajuste dos primeiros autovetores (tipicamente 1 a 20, associados aos maiores autovalores) da matriz de parentesco G como covariáveis de efeitos fixos (similar a famílias ajustadas como efeitos fixos, com matriz de incidência X). Nesse caso, o modelo ajustado é $y = Ju + Xf + Za + e$.

6. Parametrização das Matrizes de Incidência dos Efeitos Genéticos de Marcas

Modelo Aditivo-Dominância

A parametrização mais adequada para o modelo aditivo-dominância é:

Efeitos aditivos: W

$$W = \begin{cases} Se\ MM; \quad 2 \quad \rightarrow 2 - 2p = 2q \\ Se\ Mm; \quad 1 \quad \rightarrow 1 - 2p = q - p \\ Se\ mm; \quad 0 \quad \rightarrow 0 - 2p = -2p \end{cases}$$

Os valores de W devem ser centrados em zero para que os efeitos das marcas codominantes sejam efeitos de substituição alélica (α) com média zero na população, e, nesse caso, assumindo equilíbrio de Hardy-Weinberg, a variação genética aditiva do caráter na população equivale a $\sigma_a^2 = 2\sum_i^m p_i(1-p_i)\sigma_{ma}^2$.

Efeitos de dominância: S

$$S = \begin{cases} Se\ MM; \quad 0 \quad \rightarrow -2q^2 \\ Se\ Mm; \quad 1 \quad \rightarrow 2pq \\ Se\ mm; \quad 0 \quad \rightarrow -2p^2 \end{cases}$$

Isto é coerente com as seguintes definições básicas em Genética Quantitativa:

Efeitos no loco gênico b.

Genótipo	Frequência	Efeitos Aditivos	Efeitos de Dominância
BB	p^2	$2q\alpha$	$-2q^2d$
Bb	2pq	$(q-p)\alpha$	$2pqd$
bb	q^2	$-2p\alpha$	$-2p^2d$

Variâncias no loco gênico b.

Genótipo	Frequência	Variância Aditiva	Variância de Dominância
BB	p^2	$p^2(2q\alpha)^2$	$p^2(-2q^2d)^2$
Bb	2pq	$2pq[(q-p)\alpha]^2$	$2pq(2pqd)^2$
Bb	q^2	$q^2(-2p\alpha)^2$	$q^2(-2p^2d)^2$
Soma		$\sigma_a^2 = 2pq\alpha^2$	$\sigma_d^2 = (2pqd)^2$

Outra parametrização usada (inclusive pelo Selegen Genômica) surge por meio da projeção das frequências alélicas para uma população base com frequências p = q = 0.5. Assim, tem-se:

Efeitos aditivos: W

$$W = \begin{cases} Se\ MM; \quad 2 \quad \rightarrow 2-2*0.5=1 \\ Se\ Mm; \quad 1 \quad \rightarrow 1-2*0.5=0 \\ Se\ mm; \quad 0 \quad \rightarrow 0-2*0.5=-1 \end{cases}$$

Efeitos de dominância: S

$$S = \begin{cases} Se\ MM; \quad 0 \quad \rightarrow -2*0.5^2=-0.5 \\ Se\ Mm; \quad 1 \quad \rightarrow 2*0.5*0.5=0.5 \\ Se\ mm; \quad 0 \quad \rightarrow -2*0.5^2=-0.5 \end{cases}$$

7. Modelos Equivalentes

7.1. Modelo de Indivíduos (G-BLUP)

Modelo Aditivo-Dominância

$y_c = Xu + Za + Zd + e$

a: *vetor de efeitos genéticos aditivos dos indivíduos*
d = *vetor de efeitos de dominância dos indivíduos*

$a \sim N(0, G_a\sigma_a^2);\ d \sim N(0, G_d\sigma_d^2);\ e \sim N(0, I\sigma_e^2)$

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \tilde{a} \\ \tilde{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix}$$

em que:

$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_i q_i)}$; $G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_i q_i)^2}$, em que p_i e q_i são as frequências alélicas.

Ajustar um modelo genômico individual é equivalente a ajustar um modelo individual traditional mas, com as matrizes de parentesco A e D baseadas em pedigree substituídas pelas matrizes de parentesco genômico G_a and G_d para os efeitos aditivos e de dominância, respectivamente.

Segundo o modelo de marcas (ver item 7.2), a matriz de covariância para os efeitos aditivos é dada por $G_a\sigma_a^2 = Var(Wm_a) = WW'\sigma_{ma}^2$, a qual conduz a $G_a = WW'/(\sigma_{ma}^2/\sigma_a^2) = WW'/\sum_{i=1}^{n}[2p_i(1-p_i)]$. A matriz de covariância para os efeitos de dominância é dada por $G_d\sigma_d^2 = Var(Sm_d) = SS'\sigma_{md}^2$. Então $G_d = SS'/(\sigma_{md}^2/\sigma_d^2) = SS'/\sum_{i=1}^{n}[2p_i(1-p_i)]^2$. As igualdades $(\sigma_{ma}^2/\sigma_a^2) = \sum_{i=1}^{n}[2p_i(1-p_i)]$ e $(\sigma_{md}^2/\sigma_d^2) = \sum_{i=1}^{n}[2p_i(1-p_i)]^2$ advém de $\sigma_a^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]\,\sigma_{ma}^2$ e $\sigma_d^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]^2\sigma_{md}^2$ (Falconer, 1989), σ_{mai}^2 em que m_a e m_d referem-se aos efeitos aditivos e de dominância para qualquer marca i (segundo o modelo infinitesimal os diferentes locos marcadores explicam iguais quantidades de variação genética, de forma que $\sigma_{mai}^2 = \sigma_{ma}^2$ e $\sigma_{mdi}^2 = \sigma_{md}^2$).

Para evitar problemas de singularidade em G_a e G_d o Selegen Genômica as obtém da seguinte forma: $G_a = (WW')/[tr(WW')/N]$ e $G_d = (SS')/[tr(SS')/N]$ em que tr é o operador traço matricial e N é o número de indivíduos. Dessa forma, G_a e G_d apresentam melhores propriedades numéricas. Verifica-se que nesse procedimento essas matrizes são escaladas por $[tr(WW')/N]$ e $[tr(SS')/N]$ e não por $\sum_{i=1}^{n}[2p_i(1-p_i)]$ e $\sum_{i=1}^{n}[2p_i(1-p_i)]^2$. Com número de indivíduos maior que o número de marcadores as matrizes G_a e G_d são não inversíveis.

Modelo Aditivo-Dominância-Epistasia

$y_c = Xu + Za + Zd + Zi + e$

a: *vetor de efeitos genéticos aditivos dos indivíduos*
d = *vetor de efeitos de dominância dos indivíduos*
i: *vetor de efeitos genéticos epistáticos aditivo x aditivo dos indivíduos*

$a \sim N(0, G_a\sigma_a^2)$; $d \sim N(0, G_d\sigma_d^2)$; $i \sim N(0, G_i\sigma_i^2)$; $e \sim N(0, I\sigma_e^2)$

$$\begin{bmatrix} X'X & X'Z & X'Z & Z'X \\ Z'X & Z'Z+G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z & Z'Z \\ Z'X & Z'Z & Z'Z+G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} & Z'Z \\ Z'X & Z'Z & Z'Z & Z'Z+G_i^{-1}\frac{\sigma_e^2}{\sigma_i^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \tilde{a} \\ \tilde{d} \\ \tilde{i} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \\ Z'y \end{bmatrix}$$

em que $G_i = G_a \# G_a$; # é o operador produto de Hadamard.

Modelo Aditivo-Dominância-Interação g x e

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.

$y_c = Xu + Za + Zd + Zae + e$

a: *vetor de efeitos genéticos aditivos dos indivíduos*
d = *vetor de efeitos de dominância dos indivíduos*
ae: *vetor de efeitos da interação genético aditivo x ambiente dos indivíduos*

$a \sim N(0, G_a\sigma_a^2)$; $d \sim N(0, G_d\sigma_d^2)$; $ae \sim N(0, G_{ae}\sigma_{ae}^2)$; $e \sim N(0, I\sigma_e^2)$

$$\begin{bmatrix} X'X & X'Z & X'Z & Z'X \\ Z'X & Z'Z+G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z & Z'Z \\ Z'X & Z'Z & Z'Z+G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} & Z'Z \\ Z'X & Z'Z & Z'Z & Z'Z+G_{ae}^{-1}\frac{\sigma_e^2}{\sigma_{ae}^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \tilde{a} \\ \tilde{d} \\ \tilde{a}e \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \\ Z'y \end{bmatrix}$$

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

7.2. Modelo de Marcas (RR-BLUP)

$y_c = Xu + Wm_a + Sm_d + e$

m_a: *vetor de efeitos genéticos aditivos das marcas*
$Var(Wm_a) = WW'\sigma_{m_a}^2$
m_d = *vetor de efeitos de dominância das marcas*
$Var(Sm_d) = SS'\sigma_{m_d}^2$.

$$\begin{bmatrix} X'X & X'W & X'S \\ W'X & W'W + I\frac{\sigma_e^2}{\sigma_{m_a}^2} & W'S \\ S'X & S'W & S'S + I\frac{\sigma_e^2}{\sigma_{m_d}^2} \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}_a \\ \hat{m}_d \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \\ S\acute{y} \end{bmatrix}$$

em que:

$$\sigma_{m_a}^2 = \frac{\sigma_a^2}{\sum_{i=1}^{n}(2p_iq_i)}; \qquad \sigma_{m_d}^2 = \frac{\sigma_d^2}{\sum_{i=1}^{n}(2p_iq_i)^2}.$$

7.3. Modelos Equivalentes

$y_c = Xu + Za + Zd + e$ *(Modelo de indivíduos)*
$y_c = Xu + ZWm_a + ZSm_d + e$ *(Modelo de marcas)*

7.3.1 Conversão do Modelo de Marcas para Indivíduo

$$a = Wm_a = W\alpha: \textit{vetor de efeitos genéticos aditivos dos indivíduos}$$
$$d = Sm_d: \textit{vetor de efeitos genéticos de dominância dos indivíduos}$$

$$\sigma_a^2 = \sum_{i=1}^{n}(2p_iq_i)\sigma_{m_a}^2; \qquad \sigma_d^2 = \sum_{i=1}^{n}(2p_iq_i)^2\sigma_{m_d}^2$$

Os componentes de variância associados aos efeitos são $\sigma_{m_a}^2$ e $\sigma_{m_d}^2$. A quantidade m_a em um loco é o efeito de substituição alélica dado por $m_a = \alpha_i = a_i + (q_i - p_i)d_i$, em que p_i e q_i são as frequências alélicas e a_i e d_i são os valores genotípicos do homozigoto e heterozigoto, respectivamente, no loco i. A quantidade m_d equivale a $m_d = d_i$.

7.3.2 Conversão do Modelo de Indivíduo para Marcas

$$W'a = W'Wm_a; \qquad m_a = (W'W)^{-1}W'a$$
$$S'd = S'Sm_d; \qquad m_d = (S'S)^{-1}S'd$$

Número de marcas para precisa obtenção da matriz de parentesco genômico

Segundo Rolf et al. (2010), cerca de 2.500 SNP são suficientes para a estimação de G de forma precisa.

8. Genética de Populações

Imputação de Dados Perdidos

A matriz precursora de parametrização $\vec{W}$ é dada por:

$$\vec{W} = \begin{cases} Se\ MM; & 2 \\ Se\ Mm; & 1 \\ Se\ mm; & 0 \end{cases}$$

A variável indicadora $\vec{W}$ tem a seguinte distribuição:

$$\vec{W} \sim \begin{bmatrix} Média: E(\vec{W}) = 2p \\ Variância: Var(\vec{W}) = 2pq \end{bmatrix}$$

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 (W) = Var (W_i) = $(0 - 2p)^2$ x p^2 + $(1 - 2p)^2$ x 2p(1-p) + $(2 - 2p)^2$ x $(1-p)^2$= 2p(1-p)

O valor médio ou esperado da variável indicadora é 2p. Esse valor é usado pelo Selegen Genômica para imputação de valores perdidos (NA) de marcadores nos indivíduos. A seguinte regra é adotada:

$$\vec{W}_i = \begin{cases} Se\ 2p \le 0.5 \ \rightarrow \vec{W}_i = 0; \\ Se\ 0.5 < 2p < 1.5 \ \rightarrow \vec{W}_i = 1; \\ Se\ 1.5 \le 2p \le 2.0 \ \rightarrow \vec{W}_i = 2; \end{cases}$$

Após a imputação aplica-se a parametrização em W e S.

Frequências alélicas

O Selegen Genômica fornece, pelo modelo 15, as seguintes estimativas em Genética de Populações: frequências alélicas (p e q), variância ou heterozigose (2pq), MAF (frequência do alelo menos frequente), *call rate*, coeficiente de endogamia e tamanho efetivo populacional. Em seguida realiza o controle de qualidade do arquivo de marcas, eliminando locos não polimórficos (MAF < 5%) e/ou locos com baixa taxa de atendimento ou presença nos indivíduos (*call rate* < 95%). Um arquivo de marcas _cor corrigido é salvo e está pronto realizar as análises genômicas pelos demais modelos.

Assumindo os alelos de cada marca como em equilíbrio de Hardy-Weinberg na população, as frequências alélicas são obtidas por $p = \frac{n_{MM} + 0.5 n_{Mm}}{N} = \frac{2n_{MM} + n_{Mm}}{2N}$ e

$q = \frac{n_{mm} + 0.5 n_{Mm}}{N} = \frac{2n_{mm} + n_{Mm}}{2N} = 1 - p$, em que n_{MM} é o número de indivíduos com genótipo MM. A MAF é dada por $MAF = \min[p, (1-p)]$. O quadro a seguir ilustra a obtenção dessas fórmulas.

Genótipos	Código	Contagem	Frequencia	Cálculo da Frequencia de M
MM	2	n_{MM}	p^2	$n_{MM}/N = p^2$
Mm	1	n_{Mm}	$2p(1-p)$	$(1/2)\ n_{Mm}/N = p(1-p)$
mm	0	n_{mm}	$(1-p)^2$	0
Soma	-	**N**	1	$\mathbf{p = n_{Mm}/N + (1/2)\ n_{Mm}/N}$

Coeficiente de Endogamia (F) e Tamanho Efetivo (Ne)

O coeficiente de endogamia é calculado por $F = \frac{tr(WW')}{N} - 1$ e o Ne por $Ne = \frac{1}{2F}$, em que N é o número de indivíduos.

Call Rate

A *call rate* (CR) é calculada por $CR = \frac{P - A}{P} = 1 - \frac{A}{P}$, em que dentre os N indivíduos, A apresentam genótipos ausentes (perdidos ou missing) e P (presentes ou not missing) retornam genótipos no processo de genotipagem.

Call Rate ou taxa de atendimento na genotipagem de T indivíduos refere-se à proporção de sucesso no retorno de genótipos. Dos T indivíduos, A apresentam genótipos perdidos (missing) e P (not missing) retornam genótipos no processo de genotipagem. Em termos proporcionais simples, segundo o conceito estatístico de frequência, um estimador trivial para a CR seria $CR^* = \frac{P}{N} = \frac{N - A}{N} = 1 - \frac{A}{N}$ ou $CR^* = \frac{P}{N} = \frac{P}{(P + A)} = \frac{(P + A) - A}{(P + A)} = 1 - \frac{A}{(P + A)}$.

No entanto, segundo o conceito estatístico de correlação entre medidas repetidas ou repetibilidade, a obtenção de um estimador adequado remete à necessidade de observações sequenciais em um mesmo indivíduo. E isto não é possível na fração A, pois estes falharam na primeira observação. Automaticamente, o denominador da fração $CR^* = 1 - \frac{A}{(P + A)}$ muda de $(P + A)$ para P e a CR transforma-se em $CR = 1 - \frac{A}{P}$, que é o estimador recomendado. Essa expressão é também dada por $CR = \frac{P - A}{P}$, a qual contrasta com $CR^* = \frac{(P + A) - A}{(P + A)}$. Uma derivação formal de $CR = 1 - \frac{A}{P}$ é apresentada a seguir.

Call Rate como Medida de Repetibilidade

Para variáveis binárias ou com distribuição binomial (0 = fracasso na genotipagem; 1 = sucesso na genotipagem) e tendo-se duas avaliações por indivíduo, uma tabela de contingência 2 x 2 sumaria toda a informação amostral. Tal tabela é estruturada da seguinte forma:

		Avaliação 1		
		0	1	Total
Avaliação 2	0	n_{00}	n_{10}	$n_{.0}$
	1	n_{01}	n_{11}	$n_{.1}$
	Total	$n_{0.}$	$n_{1.}$	$n_{..}$

Com base nessa tabela de contingência, a estatística $\rho = n_{11}/n_{1.} - n_{01}/n_{0.}$ é o estimador para a repetibilidade na escala binomial ou correlação entre avaliações repetidas em um mesmo indivíduo.

Para o caso da variável sucesso/fracasso na genotipagem, podem ser especificadas as quantidades:

n_{11}: número de indivíduos que tiveram sucesso na genotipagem na primeira avaliação e também na segunda avaliação, ou seja, é o próprio número de indivíduos bem sucedidos na segunda avaliação, visto que aqueles que fracassaram na primeira avaliação não participam da segunda genotipagem, uma vez que são inadequados para o estudo de repetibilidade.

$n_{1.}$: soma do número de indivíduos bem sucedidos na primeira avaliação e mal sucedidos na segunda avaliação mais o número de indivíduos bem sucedidos na segunda avaliação, ou seja, essa soma é o próprio número de indivíduos bem sucedidos na primeira avaliação.

n_{01}: número de indivíduos mal sucedidos na primeira avaliação e que se tornaram bem sucedidos na segunda avaliação, ou seja, zero, pois esses não participam da segunda genotipagem.

Assim, neste caso, o estimador para a repetibilidade eqüivale a $\rho = N_2/N_1$, em que N_2 e N_1 são os números de indivíduos bem sucedidos na segunda (P-A) e primeira (P) avaliações, respectivamente. Assim, $\rho = CR = \frac{P-A}{P}$. Essa expressão assume que o número de perdidos na segunda avaliação é o mesmo ocorrido na primeira avaliação. Portanto, pelo conceito de repetibilidade, $\rho = CR = \frac{P-A}{P} = 1 - \frac{A}{P}$.

9. Validação Cruzada: Capacidade preditiva, Viés e Acurácia

A metodologia generalizada de validação cruzada via método 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.

Outra forma de estimar a acurácia da GWS é via o G-BLUP do Selegen Genômica. Para isso, alguns indivíduos devem entrar na avaliação apenas com a informação das marcas, mas, não dos fenótipos. A acurácia estimada para esses indivíduos, pelos modelos 7 a 10 do Selegen, equivalerá a uma acurácia validada.

A. Correlação: Capacidade Preditiva

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

B. Regressão de y_c em $\hat{g}_V$: Viés

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

C. Acurácia

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

10. Desequilíbrio de Ligação

A definição de desequilíbrio de ligação (r^2) refere-se à associação não aleatória de alelos de diferentes locos. 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.

O r^2 é o quadrado da correlação (r) entre alelos ou genótipos presentes no loco marcador e no loco do QTL, conforme na Tabela 3 abaixo.

Tabela 3. 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 três interpretações: (i) desvio da frequência observada de haplótipos em relação à esperada segundo segregação independente ($D = \mathrm{Pr}ob(ab) - \mathrm{Pr}ob(a)\,\mathrm{Pr}ob(b)$); (ii) quadrado da correlação (r) entre alelos (Tabela 3); (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: $r = \frac{Cov(a,b)}{[Var(a)Var(b)]^{1/2}} = \frac{\sum ab - \sum a \sum b}{[Var(a)]^{1/2}[Var(b)]^{1/2}} = \frac{\mathrm{Pr}ob(ab) - \mathrm{Pr}ob(a)\,\mathrm{Pr}ob(b)}{[pq]^{1/2}[rs]^{1/2}} = \frac{D}{[pq\,rs]^{1/2}}$. 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 $D = \mathrm{Pr}ob(ab) - \mathrm{Pr}ob(a)\mathrm{Pr}ob(b)$, 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.

- Cálculo do desequilíbrio de ligação entre pares de locos

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

De maneira similar ao r² acima, porém envolvendo todos os pares de locos, pode ser obtido o desequilíbrio médio entre todos os pares de marcadores ($\bar{r}^2_{tpl}$). E o tamanho efetivo (Ne) pode ser obtido via $\bar{r}^2_{tpl}$. A partir da expressão $r^2_{mq} = \frac{n_m\ \bar{r}^2_{tpl}}{n_m\ \bar{r}^2_{tpl} + 1}$ estima-se r^2_{mq} e então $Ne = \frac{n_m(1+r^2_{mq})}{2r^2_{mq}\ L}$, em que n_m é o número de marcadores e L é o tamanho do genoma (Resende et al., 2012).

11. GWAS – Identificação de Variantes Causais

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 $F = \frac{QM\ \text{Re}\,gressão}{\hat{\sigma}_e^2} = \frac{\hat{m}W'y + \hat{u}\ J'y - (1/n)\ (J'y)^2}{(y'y - \hat{m}W'y - \hat{u}J'y)/(n-2)}$.

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} (F 18.80).

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 4.

Tabela 4. 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%
2.36	**229.09**	**10.90**	**0.001 (= 10^{-3})**
3	**1000.00**	**13.83**	**0.02%**
3.27	**1862.09**	**15.10**	**0.0001 (= 10^{-4})**
4.08	**12022.5**	**18.80**	**0.00001 (= 10^{-5})**

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 MAS 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 (equivalente ao F para grande número de graus de liberdade do resíduo), o

valor de corte muda de 3.84 para 13.83 visando alterar a significância de 5% para 0.02% (Tabela 4). Para se obter significância de 10^{-5} o F a ser atingido é de 18.80.

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 % ($\alpha = 0.05$), 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 (adotar $\alpha^* = \alpha / n_m$, em que n_m é o número de marcadores) 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 % dos SNPs declarados como significativos na verdade não o são) de taxa de falsa descoberta pode ser considerada aceitável.

12. GWS – Exemplo

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

As frequências alélicas p e suas médias 2p bem como as heterozigoses 2pq são:

Genética	Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
p	0.50	0.30	0.10	0.10	0.50	0.20	0.20
2p	1.00	0.60	0.20	0.20	1.00	0.40	0.40
2pq	0.50	0.42	0.18	0.18	0.50	0.32	0.32

12.1 Parametrização mais Indicada

A matriz de incidência dos valores genéticos aditivos W é dada por:

$$W = \begin{cases} Se\ MM; \quad 2 \quad \rightarrow 2-2p = 2q \\ Se\ Mm; \quad 1 \quad \rightarrow 1-2p = q-p \\ Se\ mm; \quad 0 \quad \rightarrow 0-2p = -2p \end{cases}$$

Matriz de Incidência W

Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
2 – 1	0 – 0.6	0 – 0.2	0 – 0.2	2 - 1	0 – 0.4	0 – 0.4
1 – 1	1 – 0.6	0 – 0.2	0 – 0.2	1 - 1	1 – 0.4	0 – 0.4
0 – 1	2 – 0.6	0 – 0.2	0 – 0.2	0 - 1	0 – 0.4	2 – 0.4
1 – 1	0 – 0.6	1 – 0.2	0 – 0.2	1 - 1	0 – 0.4	0 – 0.4
1 - 1	0 – 0.6	0 – 0.2	1 – 0.2	1 - 1	1 – 0.4	0 – 0.4

Matriz de Incidência W

Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6	Marca 7
1	-0.6	– 0.2	– 0.2	1	– 0.4	– 0.4
0	0.4	– 0.2	– 0.2	0	0.6	– 0.4
-1	1.4	– 0.2	– 0.2	-1	– 0.4	1.6
0	-0.6	0.8	– 0.2	0	– 0.4	– 0.4
0	-0.6	– 0.2	0.8	0	0.6	– 0.4

Os efeitos genéticos aditivos dos marcadores são obtidos resolvendo-se

$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + I\dfrac{\sigma_e^2}{(\sigma_a^2 / \sum_{i=1}^{7} 2p_iq_i)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix} \text{ ou } \begin{bmatrix} X'X & X'W \\ W'X & W'W + I\dfrac{1-h_a^i}{(h_a^i / \sum_{i=1}^{7} 2p_iq_i)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix}.$$

No presente exemplo $\sum_{i=1}^{7} 2p_iq_i = 2.42$. Assumindo uma herdabilidade individual no sentido restrito igual a 0.71 tem-se $\dfrac{1-h_a^2}{(h_a^2 / \sum_{i=1}^{7} 2p_iq_i)} = 1$.

Resolvendo-se o sistema tem-se os resultados $\begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} 11.4900 \\ -0.3526 \\ 0.2761 \\ 1.4467 \\ -1.3701 \\ -0.3526 \\ 0.5436 \\ -1.63765 \end{bmatrix}$, em que 11,4900 é a média geral e os demais valores são as estimativas dos efeitos genéticos aditivos 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 $VGG = \begin{bmatrix} -0.4486 \\ 1.0763 \\ -1.7612 \\ 1.7033 \\ -0.5699 \end{bmatrix}$.

12.2 Parametrização Alternativa

Considerando-se $W = \vec{W}$ 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 tendo-se $\dfrac{1-h_a^2}{(h_a^2 / \sum_{i=1}^{7} 2p_iq_i)} = 1$, 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 $VGG = \hat{y}_j = \sum_i \vec{w}_{ij}\hat{m}_i$. No caso, as predições para os 5 indivíduos são $VGG=\begin{bmatrix}-1.4104\\0.1145\\-2.7230\\0.7415\\-1.5317\end{bmatrix}$.

Verificam-se que os valores genéticos aditivos de marcas (m) foram iguais para as duas parametrizações (ítens 11.1 e 11.2). Isto ocorre devido ao fato de as duas parametrizações manter a mesma proporcionalidade entre os elementos regressores de m. No entanto, a estimativa da média geral diferiu nas duas abordagens, produzindo interceptos diferentes. A diferença entre as duas estimativas foi de 12,4519 – 11,4900 = 0,9619.

Os valores genéticos aditivos de indivíduos (VGG) diferiram entre as duas parametrizações (ítens 11.1 e 11.2). Isto ocorre devido ao fato da mudança de intercepto. Os VGGs produzidos no ítem 11.2 equivale àqueles produzidos no ítem 11.1 menos o valor (0,9619) da diferença entre interceptos. Assim, as diferentes parametrizações afetam o viés de predição, sendo aquela do ítem 11.1 a preferida.

13. Coeficiente de Endogamia Genômicos

Os coeficientes de endogamia genômicos podem ser obtidos de três maneiras, conforme a seguir.

(a) coeficientes de endogamia genômicos via matriz de parentesco G

- os coeficientes de endogamia por indivíduo são dados por $F_{jj} = G_{jj}-1$
- o coeficiente de endogamia médio F da população é dado pela media de F_{jj}
- G_{jj}: são os elementos da diagonal de G.
- A matriz G é dada por $G = \frac{WW'}{\sum_{i=1}^{n} 2p_i q_i}$.
- O tamanho efetivo populacional na geração anterior é dado por $N_e = \frac{1}{2F}$.

(b) coeficientes de endogamia genômicos elemento por elemento

- os coeficientes de endogamia por indivíduo em um dado loco são dados por $\hat{F}_{ij} = \frac{(w_{ij} - 2p_i)^2}{2p_i(1-p_i)} - 1$.
- os coeficientes de endogamia por indivíduo, médio em todos os locos, são dados por $\hat{F}_{jj} = (1/n)\sum_{i=1}^{n} \hat{F}_{ij}$; a média desses fornece F.

Os elementos diagonais de G são dados $G_{jj} = (1/n)\sum_{i=1}^{n} \frac{(w_{ij} - 2p_i)(w_{ik} - 2p_i)}{2p_i(1-p_i)}$ e equivalem a $G_{jj} = 1 + \hat{F}_{jj}$.

(c) coeficientes de endogamia genômicos via frequência de heterozigotos

As frequências genotípicas associadas aos três genótipos de um SNP em uma espécie diploide são dados 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. Na matriz G o parentesco não é uma probabilidade (conforme definição clássica de IBD), mas sim uma correlação entre valores genéticos aditivos. E F também não é uma probabilidade, mas uma correlação entre gametas que se unem.

Com base na estrutura populacional o cálculo de F pode ser realizado via frequência de heterozigotos, conforme derivação a seguir. Nessa estrutura a frequência de heterozigotos equivale a $[2p(1-p)](1-F)$ e, portanto, a heterozigose (het) observada é dada por $het = [2p(1-p)](1-F)$. Assim, o coeficiente de endogamia F

é dado por $F = \frac{[2p(1-p)]-het}{2p(1-p)} = 1 - \frac{het}{2p(1-p)}$. Se $het = 2p(1-p)$, F = 0, ou seja, se as heterozigoses observada e esperada forem iguais a endogamia é nula. Se computada para cada indivíduo, tem-se $F_{jj} = \frac{[2p(1-p)]-het_{jj}}{2p(1-p)}$ e a endogamia média F é obtida como a média de F_{jj}. O F médio atual permite calcular tamanho efetivo (Ne) da geração reprodutora anterior via $Ne = 1/(2F)$.

Ne Atual e Endogamia Futura

Conforme Resende (2002), em uma população estruturada com N_f famílias e k_f indivíduos por família a expressão para o Ne é dada por $N_e = \frac{N_f\ k_f}{(1-r)+k_f r} = 1/(2F)$, em que $F = \frac{1}{2N}[(1-r)+k_f r]$, sendo o coeficiente de parentesco de Wright entre os indivíduos de uma família. Com desconhecida estruturação em famílias, pode-se tomar um coeficiente r médio válido para todos os N = $N_f k_f$ = k_f indivíduos e, portanto, tem-se $N_e = \frac{N}{(1-r)+Nr}$. Sendo o parentesco entre os genitores, determinante da endogamia dos filhos, tem-se o F da geração seguinte dado por F = 1/(2 Ne). A endogamia da geração corrente em relação à população base é dada por Fst = (tr(G)/N) – 1. O valor de r é dado pela média dos elementos fora da diagonal de G, divididos pelos produtos das raízes quadradas de seus correspondentes elementos da diagonal de G. G é dada por:

$$G = \frac{WW'}{\sum_{i=1}^{n}(2p_i q_i)}.$$

O parentesco entre os indivíduos j e k, é dado por $r_{jk} = \frac{G_{jk}}{(1+F_j)^{1/2}(1+F_k)^{1/2}}$.

Com N grande, a quantidade $N_e = \frac{N}{(1-r)+Nr}$ tende a $N_e = \frac{N}{Nr}$ e, portanto, a $N_e = \frac{1}{r}$. Sendo $N_e = \frac{1}{2F}$, vê-se que 2F = r e, portanto, F = (1/2) r, ou seja, a endogamia dos filhos é igual à metade do parentesco de Wright entre os genitores, ou seja, é igual ao coeficiente de coancestria (kinship) ou coeficiente de parentesco de Malecot entre os genitores.

14. Eficiência Comparativa no Uso de G Genômica *vs* A Genealógica

A matriz G calculada utilizando marcadores moleculares pode ser mais eficiente do que a matriz de parentesco A baseada em pedigree, uma vez que considera a amostragem mendeliana dentro de família e também a distorção de segregação. A proporção da variação fenotípica explicada (R^2_{LR}), utilizando a matriz G (nas equações do modelo misto) pode ser maior do que usando a matriz A.

A comparação de modelos usando A ou G pode ser baseada nos valores de R^2_{LR}, o qual representa uma razão de verossimilhança associada à variação fenotípica explicada. O R^2_{LR} é dado por $R^2_{LR} = 1 - \exp[(-2/N)(\log L_m - \log L_0)]$, em que $\log L_m$ é o

máximo da log-verossimilhança pelo ajuste do modelo via as equações de modelo misto; $\log L_0$ é o máximo da log-verossimilhança pelo ajuste de um modelo com apenas o intercepto (a obtenção de log L_o pelo Selegen Genômica é realizada via opção BLUP com h^2 e todos os demais coeficientes de determinação c^2 fixados em zero) e N é o número de indivíduos. Os fundamentos estatísticos associados à essa definição podem ser encontrados em Magee (1990) e Sun et al. (2010). A quantidade R^2_{LR} é calculada para o ajuste com G (R^2_{LR-G}) e com A (R^2_{LR-A}). Se $R^2_{LR-G} > R^2_{LR-A}$, o modelo com G é selecionado. No Selegen genômica é obtido $\log L_{m-G}$ e no Selegen Reml/Blup é obtido $\log L_{m-A}$. Assim, obtém-se R^2_{LR-G} e R^2_{LR-A} para comparação e também para inferência sobre a proporção da variação fenotípica explicada pelos QTL marcados por meio de um coeficiente de determinação baseado em LR.

A quantidade -2 log L_o refere-se à deviance do modelo com o ajuste apenas do intercepto e pode ser comparada com -2 $\log L_{m-G}$ de um modelo com apenas efeitos aditivos visando verificar a significância dos efeitos aditivos.

15. Índice de Seleção via BLUP fenotípico + BLUP GWS

Essa situação ocorre quando estão disponíveis os valores genéticos preditos para o caráter com base em dados fenotípicos (a, usando a matriz A) e genotípicos de marcas (g, usando a matriz 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 (Resende et al., 2014):
$I = b_1\hat{g} + b_2\hat{a}$

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 confiabilidade (quadrado da acurácia) pela inclusão da informação molecular é dado por $r^2_{aum} = \dfrac{r^2_{\hat{g}g}(1 - r^2_{\hat{a}a})^2}{(1 - r^2_{\hat{g}g} r^2_{\hat{a}a})}$. Assim, usando os valores genéticos e as acurácias pelos dois BLUPs, preditos pelo Selegen, esse índice A-BLUP+G-BLUP pode ser obtido e usado na prática.

Agradecimento

O autor agradece a Camila Ferreira Azevedo, pelas pesquisas desenvolvidas em conjunto, na área de Métodos Estatísticos em Seleção Genômica, as quais contribuíram para os procedimentos incluídos nesse software.

16. Referências Bibliográficas

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. de ; SILVA, F. F. E. ; LOPES, P. S. ; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, p. 619-626, 2013.

AZEVEDO, C.F. ; F.F. SILVA, M.D.V. DE RESENDE, M.S. LOPES, N. DUIJVESTEIJN, S.E.F. GUIMARÃES, P.S. LOPES, M.J. KELLY, J.M.S. VIANA, E.F. Knol (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal of Animal Breeding and Genetics**, 2014.

AZEVEDO, C. F. ; SILVA, F.F. ; RESENDE, M. D. V.; PETERNELLI, L. A.; GUIMARÃES, S. E. F.; LOPES, P. S. Quadrados mínimos parciais uni e multivariado aplicados na seleção genômica para características de carcaça em suínos. **Ciência Rural**, v. 43, p. 1642-1649, 2013.

AZEVEDO, C. F. ; ABAD, J.I.M. ; MISSIAGGIA, A. A. ; AGUIAR, A. M. ; RESENDE, M.D.V. de . Parametrizações em marcadores dominantes DArTs para características de crescimento em eucalipto. **Revista Matemática e Estatística em Foco**, v. 1, p. 1-2, 2013.

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

CAVALCANTI, J.J. ; RESENDE, M.D.V. Predição simultânea de efeitos de marcadores e seleção Genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, 2011.

CHURCHILL, G. A.; DOERGE, R. W. Empirical threshold values for quantitative trait mapping. **Genetics**, v. 138, p. 963-971, 1994.

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

FERNANDO, R. L., GARRICK D. 2009. **GenSel – User manual for a portfolio of genomic selection related analyses, 2nd ed.** ***for version 2.12***. Animal Breeding and Genetics, Iowa State University, Ames, IWOA.

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.

FRITSCHE NETO, R. **Seleção genômica ampla e novos métodos de melhoramento do milho.** Viçosa: Universidade Federal de Viçosa, 2011. 28 p. (Tese Mestrado em Genética e Melhoramento).

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**, v. 59, p. 794-802, 2012.

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.

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

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

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

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. A sire evaluation method which accounts for unknown genetic and environmental trends, herd differences, season, age effects and differential culling. Proc Symp Estimating Breeding Values of Dairy Sires and Cows, Washington DC. 1966.

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.

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

MAGEE L (1990). R2 measures based on Wald and likelihood ratio joint significance tests. **Am Stat** 44: 250–253.

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

MUÑOZ, P.R.; RESENDE JR., M.F.R. ; HUBER, D. A. ; QUESADA, T. ; RESENDE, M.D.V. de ; NEALE, D. B. ; WEGRZYN, J. L. ; KIRST, M. ; PETER, G. F. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. **Crop Science**, v. 53, p.1115-1123, 2013.

OLIVEIRA, E. J.; RESENDE, M.D.V; SANTOS, V.S. et al. Genome-wide selection in cassava. **Euphytica**, v. 187, p.263-276, 2012.

PINHEIRO, V. R.; SILVA, F. F. E.; GUIMARÃES, S. E. F. ; RESENDE, Marcos Deon Vilela ; LOPES, P. S. ; CRUZ, C. D. ; AZEVEDO, C. F. . Mapeamento de QTL para características de crescimento de suínos por meio de modelos de regressão aleatória. **Pesquisa Agropecuária Brasileira**, v. 48, p. 190-196, 2013.

PINHEIRO, V. R. **Uso de regressão aleatória na detecção de QTL em suínos**. 2012. UFV - Universidade Federal de Viçosa (Mestrado em Estatística Aplicada e Biometria).

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos lineares mistos. In: RESENDE, M. D. V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. 1. ed. Colombo: Embrapa Florestas, 2007. p. 517-534.

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

RESENDE M.D.V.; RESENDE JR., 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. 79p.

RESENDE M.D.V.; SILVA, F.F.; VIANA, J.M.S.; PETERNELLI, L.A. **Métodos estatísticos na seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas. 2011. 106p.

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, 2012a.

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. 2012b. 291 p. http://www.det.ufv.br/ppestbio/corpo_docente.php

RESENDE, M.D.V. de; SILVA, F. F. E. ; Resende Jr., M.F.R. **Genética de Associação (GWAS)**. In: Aluízio Borém; Roberto FRITSCHE-NETO. (Org.). Biotecnologia Aplicada ao Melhoramento de Plantas. 1ed.Visconde do Rio Branco: Suprema, 2013, v. 1, p. 119-150.

RESENDE, M.D.V. de; SILVA, F. F. E. ; Resende Jr., M.F.R. . **Seleção Genômica Ampla (GWS)**. In: Aluízio Borém; Roberto FRITSCHE-NETO. (Org.). Biotecnologia Aplicada ao Melhoramento de Plantas. 1ed.Visconde do Rio Branco: Suprema, 2013, v. 1, p. 151-188.

RESENDE, M.D.V. de ; SILVA, F. F. E. ; Resende Jr., M.F.R. ; AZEVEDO, C. F. **Genome-wide association studies (GWAS)**. In: Aluizio Borem; Roberto Fritsche-Neto. (Org.). Biotechnology and Plant Breeding. 1 ed.Dordrecht: Elsevier, 2014, v. 1, p. 83-104.

RESENDE, M.D.V. de ; SILVA, F. F. E. ; Resende Jr., M.F.R. ; AZEVEDO, C. F. **Genome-wide selection (GWS)**. In: Aluizio Borem; Roberto Fritsche-Neto. (Org.). Biotechnology and Plant Breeding. 1 ed.Dordrecht: Elsevier, 2014, v. 1, p. 105-134.

RESENDE JR., M. F. R. **Seleção genômica ampla no melhoramento vegetal**. UFV, 2010. 67 p. (Tese Mestrado em Genética e Melhoramento).

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.

RESENDE, R. M. S.; CASLER, M. ; RESENDE, M.D.V. de . Genomic Selection in Forage Breeding: Accuracy and Methods. **Crop Science**, v. 54, p. 143-156, 2014.

ROCHA, G.S. **Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais**. Viçosa: Universidade Federal de Viçosa, 2011. 46 p. (Tese Mestrado em Estatística Aplicada e Biometria).

RODRIGUES, D.T. **Interação genótipos ambientes em animais via modelos de normas de reação**. UFV- Universidade Federal de Viçosa, 2012 (Mestrado em Estatística Aplicada e Biometria).

ROLF MM, TAYLOR JF, SCHNABEL RD, MCKAY SD, MCCLURE MC, NORTHCUTT SL, KERLEY MS, WEABER RL. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. **BMC Genetics**. 2010; 11:24.

SANTOS, V. S.; MARTINS FILHO, S.; SILVA, F. F.; RESENDE, M.D.V. de . Uma aplicação do modelo de sobrevivênciade Cox na seleção genômica ampla de suínos. **Revista Matemática e Estatística em Foco**, v. 1, p. 2-2, 2013.

SANTOS, V. S. **Seleção genômica ampla em suínos usando o modelo de sobrevivência de Cox**. 2013. Viçosa: Universidade Federal de Viçosa, 2013. (Tese Mestrado em Estatística Aplicada e Biometria).

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, p. 310-314, 2011.

SILVA, F.F. ; ROCHA, G.S. ; RESENDE, M.D.V. ; GUIMARÃES, S.E.F. ; PETERNELLI, L.A. ; DUARTE, D.A.S. ; AZEVEDO, C. . Seleção genômica ampla para curvas de crescimento. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 65, p. 1519-1526, 2013.

SILVA, F. F. E. ; RESENDE, M.D.V. de ; ROCHA, G. S. ; DUARTE, D. A. S. ; LOPES, P. S. ; BRUSTOLINI, O. J. ; THUS, S. ; VIANA, J. M. S. ; GUIMARÃES, S. E. F. . Genomic growth curves of an outbred pig population. **Genetics and Molecular Biology** (online version), v. 36, 2013.

SILVA, F. F. ; VIANA, J. M. S.; FARIA, V. R.; RESENDE, M. D. V. Bayesian inference of mixed models in quantitative genetics of crop species. **Theoretical and Applied Genetics**, v. ?, p. On line first-?, 2013.

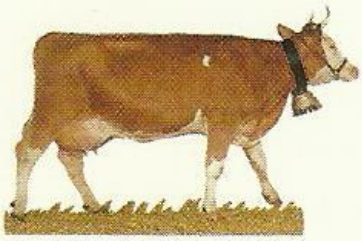
SUN G, ZHU C, KRAMER MH, YANG S-S, SONG W, PIEPHO H-P, YU J: Variation explained in mixed-model association Mapping. **Heredity** 2010, 105:333–340.

LES VACHES / COWS

Lourdaise

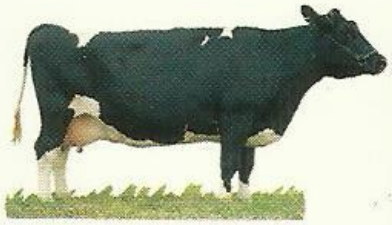
Froment du Léon

Salers

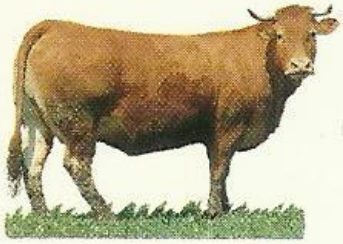


Simental française

Charolaise



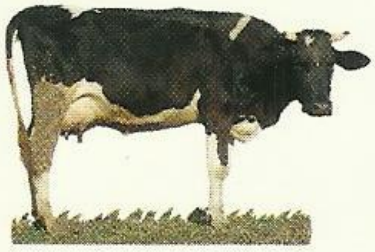
Holstein



Limousine

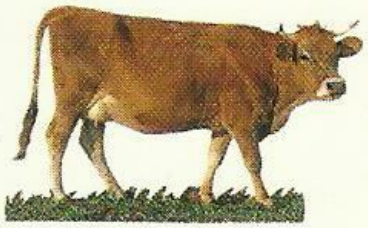
Ferrandaise

Normande

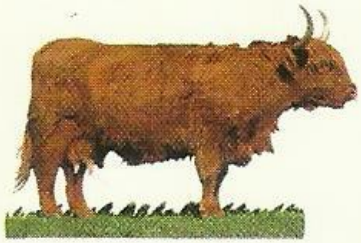


Bretonne pie noire

Corse

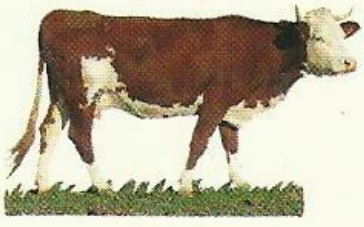


Tarine



Highlands

Brune



Abondance