GILSON SILVÉRIO DA ROCHA

MÉTODOS ESTATÍSTICOS NA SELEÇÃO GENÔMICA AMPLA PARA CURVAS DE CRESCIMENTO EM ANIMAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2011

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
R672m
2011

Rocha, Gilson Silvério da, 1984-
Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais / Gilson Silvério da Rocha. – Viçosa, MG, 2011.
xi, 46f. : il. (algumas col.) ; 29cm.

Inclui apêndice.
Orientador: Fabyano Fonseca e Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 43-45

1. Suíno - Curvas de crescimento - Métodos estatísticos.
2. Marcadores genéticos - Métodos estatísticos.
3. Polimorfismo (Genética). 4. Genética molecular.
I. Universidade Federal de Viçosa. II. Título.

CDD 22. ed. 519.5

GILSON SILVÉRIO DA ROCHA

MÉTODOS ESTATÍSTICOS NA SELEÇÃO GENÔMICA AMPLA PARA CURVAS DE CRESCIMENTO EM ANIMAIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 20 de junho de 2011

Prof. Carlos Souza do Nascimento

Prof. Cosme Damião Cruz

Dr. Marcos Deon Vilela de Resende
(Co-orientador)

Prof. Fabyano Fonseca e Silva
(Orientador)

Aos meus pais, Sueli e Paulo;
Aos meus irmãos, Gilmar e Júnior;
Aos meus amigos,
por sempre acreditarem em mim.

AGRADECIMENTOS

A Deus, pelo dom da vida e por sempre iluminar meu caminho, dando-me forças para vencer todos os obstáculos e para tornar possível mais este sonho.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade de realizar o curso.

Ao Departamento de Zootecnia da Universidade Federal de Viçosa, pela concessão dos dados utilizados na pesquisa.

Aos meus pais, Sueli e Paulo, pelo amor, pela educação, pela confiança, pelo cuidado e por estarem sempre presentes em todas as etapas da minha vida.

Aos meus irmãos, Gilmar e Júnior, pela amizade, pela confiança e por estarem sempre dispostos a ajudar.

Aos meus familiares, pela ajuda e pelo incentivo na luta pelos meus ideais.

Ao professor e orientador Fabyano Fonseca e Silva, pelo exemplo de pesquisador, pela amizade, pela orientação, pelos sábios ensinamentos, responsáveis pelo meu crescimento profissional, e por estar sempre disponível e atencioso para me atender, mesmo com várias outras obrigações e compromissos.

Ao Doutor e coorientador Marcos Deon Vilela de Resende, pelos ensinamentos, pelos conselhos e pela confiança.

Aos coorientadores Luiz Alexandre Peternelli e Simone Eliza Facioni Guimarães, pelo apoio.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, em especial àqueles que contribuíram para minha formação acadêmica.

Aos membros da banca examinadora, professor Carlos Souza do Nascimento e professor Cosme Damião Cruz, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

Aos funcionários da secretaria do Departamento de Estatística, pelo carisma e por estarem sempre dispostos a ajudar.

A todos os amigos do mestrado, pela companhia nos estudos e no lazer. O apoio de vocês foi essencial.

Ao REUNI, pela concessão da bolsa de estudos.

A todos que, de alguma forma, contribuíram para o meu crescimento profissional e para a realização deste trabalho.

BIOGRAFIA

GILSON SILVÉRIO DA ROCHA, filho de Sueli Aparecida da Silva Rocha e de Paulo Antônio da Rocha, nasceu em Ouro Fino, Minas Gerais, em 24 de julho de 1984.

Em março de 2003, ingressou no curso de Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em julho de 2008.

Em agosto de 2009, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 20 de junho de 2011.

SUMÁRIO

RESUMO

ROCHA, Gilson Silvério da, M. Sc., Universidade Federal de Viçosa, junho de 2011. **Métodos estatísticos na seleção genômica ampla para curvas de crescimento em animais**. Orientador: Fabyano Fonseca e Silva. Coorientadores: Luiz Alexandre Peternelli, Marcos Deon Vilela de Resende e Simone Eliza Facioni Guimarães.

O principal atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações do DNA na seleção genômica, de modo a permitir alta eficiência seletiva, rapidez na obtenção de ganhos genéticos com a seleção e baixo custo. Uma forma prática e consistente de analisar a eficiência produtiva de animais de corte sujeitos à seleção é por meio dos estudos de curvas de crescimento, pois estas representam uma trajetória longitudinal dos pesos dos animais em função do tempo. Para isso, primeiramente ajustam-se modelos de crescimento (modelos não lineares) aos dados de peso-idade de cada animal submetido à seleção e consideram-se os parâmetros estimados como fenótipos. Este procedimento permite a obtenção de estimativas de parâmetros genéticos para qualquer ponto da trajetória de crescimento e possibilita a compreensão da arquitetura genética de toda a trajetória, uma vez que as informações de todas as pesagens são condensadas por esses poucos parâmetros interpretáveis biologicamente. Em seguida, os parâmetros estimados dos modelos de crescimento são utilizados para predizer os Valores Genéticos Genômicos (*Genomic Breeding Value* – GBV) por meio de métodos estatísticos específicos para a Seleção Genômica

Ampla (*Genome Wide Selection* – GWS). O objetivo geral do presente trabalho foi empregar métodos estatísticos usados na Seleção Genômica Ampla, especificamente o RR-BLUP/GWS e o LASSO Bayesiano, no estudo de curvas de crescimento animal, considerando como variáveis fenotípicas as estimativas dos parâmetros de modelos de regressão não linear. Os objetivos específicos foram: estimar valores genéticos genômicos para cada indivíduo avaliado; estimar efeitos de marcadores SNPs e identificar os de maiores efeitos; selecionar, via técnicas de agrupamento, grupos de indivíduos geneticamente superiores em relação à curva de crescimento; e validar toda metodologia utilizada via estudo de simulação e aplicá-la a dados reais de uma população F_2 de suínos proveniente do cruzamento de dois varrões da raça naturalizada brasileira Piau com 18 fêmeas de linhagem comercial (Landrace × Large White × Pietrain). Os resultados indicaram que os métodos estatísticos na Seleção Genômica Ampla foram eficientes no estudo de curvas de crescimento, considerando dados simulados e dados reais de peso-idade de suínos. A GWS apresentou alta acurácia na seleção para a trajetória das curvas de crescimento e possibilitou a detecção de QTLs (*Quantitative Trait Loci*) para os parâmetros da curva dos indivíduos considerados. Na ausência de genes de grande efeito, os métodos RR-BLUP/GWS e LASSO Bayesiano produziram resultados semelhantes, no entanto o método LASSO Bayesiano apresentou maior eficiência quando o gene halotano, caracterizado como de grande efeito, foi incluído como marcador nas análises.

ABSTRACT

ROCHA, Gilson Silvério da, M. Sc., Universidade Federal de Viçosa, July, 2011. **Statistical methods used in genome wide selection for growth curves in animals.** Adviser: Fabyano Fonseca e Silva. Co-advisers: Luiz Alexandre Peternelli, Marcos Deon Vilela de Resende and Simone Eliza Facioni Guimarães.

The main contribution of molecular genetics to the benefit of applied genetic breeding is the direct use of the DNA data in genomic selection, allowing high selective efficiency and speed in the acquisition of genetic gains in selection and low costs. A practical and consistent way of analyzing the productive efficiency of beef animals subjected to selection is through the study of growth curves, as these represent a longitudinal trajectory of the weights of the animals in function of time. Thus, firstly, growth models (non-linear models) are adjusted to the weight-age data of each animal submitted to selection and the parameters estimated as phenotypes are considered. This procedure permits to determine genetic parameter estimates for any growth trajectory point, and to understand the genetic architecture of the entire trajectory, since all the weighing information is condensed by these few biologically interpretable parameters. The parameters estimated from the growth models are used to predict the Genomic Breeding Value (GBV) by means of specific statistical methods for the Genome Wide Selection (GWS). The general objective of this work was to apply statistical methods used in the Genome Wide Selection, mainly RR-BLUP/GWS and the Bayesian LASSO on the study of animal growth curves, considering as phenotypic variables the estimates of the parameters of non-linear

regression models. The specific objectives were: to estimate the genomic breeding values for each individual evaluated; to estimate the effect of SNP markers and to identify those with the greatest effects; to select, via grouping techniques, groups of individuals genetically superior, in relation to the growth curve; and to validate all the methodology used via simulation study and apply it to real data of an F_2 population of swine originated from the cross of two males from the naturalized Brazilian race Piau with 18 females of a commercial line (Landrace × Large White × Pietrain).The results indicated that the Genome Wide Selection statistical methods were efficient in studying the growth curves, considering simulated and real swine weight-age data. GWS presented high accuracy in the selection of the growth curve trajectory, allowing the detection of the QTLs (*Quantitative Trait Loci*) for the curve parameters of the individuals studied. In the absence of genes of significant effect, the methods RR-BLUP/GWS and Bayesian LASSO showed similar results but the latter showed more efficiency when the halothane gene, characterized as of significant effect, was included as a marker in the analyses.

1 INTRODUÇÃO

A eficiência dos programas de melhoramento genético depende basicamente de duas ações do geneticista: a criação e a identificação de genótipos superiores. Em ambas as ações, a seleção desempenha papel fundamental na definição dos cruzamentos a serem realizados, visando à criação de novos genótipos, e na indicação dos indivíduos superiores a serem usados comercialmente ou em novos ciclos de seleção (RESENDE *et al*., 2008).

O principal atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações de DNA na seleção, de modo a permitir alta eficiência seletiva, rapidez na obtenção de ganhos genéticos com a seleção e baixo custo (RESENDE *et al*., 2008).

A partir do início do século XXI, os avanços biotecnológicos na área de automação do processo de genotipagem permitiram o desenvolvimento de novas classes de marcadores moleculares, dentre os quais se destacam os polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms* – SNPs). Diante da disponibilidade desses marcadores, Meuwissen *et al*. (2001) idealizaram a Seleção Genômica Ampla (*Genome Wide Selection* – GWS), que consiste na análise simultânea de grande número de marcadores amplamente distribuídos no genoma. Esses marcadores podem ser utilizados na estimação de valores genéticos genômicos e na localização de locos de características quantitativas (*Quantitative Trait Loci* – QTLs). Sua aplicação foi vislumbrada para auxiliar os procedimentos de seleção no melhoramento convencional, chamado de seleção genômica.

Tornam-se cada vez mais abundantes informações genômicas individuais massivas para animais de produção, com dezenas ou centenas de milhares de marcadores (GODDARD; HAYES, 2007). A disponibilidade dessas informações incentiva o desenvolvimento de modelos que as utilizam para assistir e melhorar diretamente a predição do mérito genético individual para as características de interesse.

Em programas de melhoramento genético de suínos é de interesse obter animais com bom desempenho para características que envolvam taxa de crescimento, eficiência alimentar e peso de abate. Para o produtor, as características desejáveis são taxa de crescimento e eficiência alimentar; para o mercado processador, quantidade de carne na carcaça e, ou, qualidade da carne; e para o mercado consumidor, qualidade da carne.

Uma classe especial de características quantitativas é aquela representada por avaliações tomadas ao longo do tempo no mesmo indivíduo (dados longitudinais), por exemplo, medidas de comprimento, altura ou peso corporal. Quando se considera este último, o termo curva de crescimento passa a ser utilizado, e representa importante área de pesquisa em estudos zootécnicos. Diferentes metodologias têm sido propostas para analisar curvas de crescimento, sendo o ajuste de modelos de regressão não linear a mais usual (SILVA, 2010).

Entre as alternativas para seleção de indivíduos com base em características longitudinais de crescimento duas se destacam. Na primeira, considera-se o fenótipo como sendo o peso em cada idade e realizam-se análises separadas a cada tempo considerado. Este procedimento impossibilita compreender a dinâmica de todo o processo de crescimento sob o ponto de vista genético, uma vez que os resultados ficam restritos a pontos específicos. Na segunda, ajustam-se modelos de crescimento aos dados de peso-idade de cada animal submetido à seleção e consideram-se os parâmetros estimados como fenótipos. Esta alternativa permite a obtenção de estimativas de parâmetros genéticos para qualquer ponto da trajetória de crescimento e possibilita a compreensão da arquitetura genética de toda a trajetória, uma vez que as informações de todas as pesagens são condensadas por esses poucos parâmetros interpretáveis biologicamente.

Considerando a segunda alternativa, predições de valores genéticos genômicos são obtidas para parâmetros estimados dos modelos de crescimento utilizando métodos estatísticos específicos para Seleção Genômica Ampla. De

acordo com Gianola *et al*. (2003), informações genômicas com base em marcadores SNPs podem se constituir de dezenas ou centenas de milhares de covariáveis, possivelmente com alta colinearidade, o que demanda a utilização de métodos estatísticos que considerem a seleção de covariáveis e a regularização do processo de estimação, como o Bayes B e o LASSO Bayesiano, ou métodos que considerem apenas a regularização, como o Bayes A e o RR-BLUP/GWS (*Random Regression-Best Linear Unbiased Prediction/ Genome Wide Selection*).

O principal objetivo deste estudo foi empregar métodos estatísticos usados na Seleção Genômica Ampla, especificamente o RR-BLUP/GWS e o LASSO Bayesiano, no estudo de curvas de crescimento animal, considerando como variáveis fenotípicas as estimativas dos parâmetros de modelos de regressão não linear. Os objetivos específicos foram: estimar valores genéticos genômicos para cada indivíduo avaliado; estimar efeitos de marcadores SNPs e identificar os de maiores efeitos em módulo; selecionar, via técnicas de agrupamento, grupos de indivíduos geneticamente superiores em relação à curva de crescimento; e validar toda metodologia utilizada via estudo de simulação e aplicá-la a dados reais de uma população F_2 de suínos proveniente do cruzamento de dois varrões da raça naturalizada brasileira Piau com 18 fêmeas de linhagem comercial (Landrace × Large White × Pietrain).

2 REVISÃO DE LITERATURA

2.1 Curvas de crescimento

O processo de crescimento dos animais é um fenômeno complexo, sendo de grande importância para a área de Zootecnia. De acordo com Tedeschi *et al*. (2000), o processo de crescimento e desenvolvimento dos animais é assunto de bastante interesse para os pesquisadores, pois o seu domínio permite que o manejo nutricional seja conduzido eficientemente, além de permitir que programas de seleção sejam elaborados para as características de crescimento inerentes a cada raça.

O crescimento dos animais pode ser avaliado com base em algumas características, dentre as quais se destaca o peso corporal, que tem sido empregado como importante ferramenta em programas de seleção. A relação entre o peso do animal e o tempo geralmente é descrita por meio de modelos de regressão não linear, que condensam informações de peso-idade de todo o período de vida de um indivíduo em um conjunto de parâmetros biologicamente interpretáveis (FITZHUGH Jr., 1976). Dentre esses parâmetros, geralmente destacam-se o peso à maturidade, ou peso adulto, e a taxa de maturidade, que representa a velocidade de crescimento, de modo que quanto mais alto for o seu valor, mais precoce é o animal (SILVA *et al*., 2004). Segundo Silva (2010), as estimativas desses parâmetros constituem características fenotípicas alternativas para programas de seleção que visam à obtenção de animais mais precoces, com maior peso e melhor qualidade de carcaça.

Segundo Fitzhugh Jr. (1976), os seguintes requisitos devem ser atendidos para que um modelo de regressão não linear descreva adequadamente a relação peso-idade: interpretação biológica dos parâmetros, "alta qualidade" de ajuste e facilidade de convergência. Assim, nos estudos de curvas de crescimento é imprescindível que se tenha bom conhecimento sobre a teoria de modelos de regressão não linear, a fim de facilitar a compreensão do fenômeno abordado.

2.2 Modelos de Regressão não Linear

O principal objetivo dos modelos de regressão é modelar o relacionamento entre variáveis preditoras e variáveis repostas. Este relacionamento pode ser por meio de uma função linear ou não linear. Nos modelos de regressão linear a variável resposta é expressa como função linear dos coeficientes de regressão. Porém, muitas vezes surge a necessidade de utilizar modelos mais flexíveis, em que pelo menos uma das derivadas parciais em relação aos parâmetros está em função de algum parâmetro. Esses modelos são denominados de modelos de regressão não linear.

Segundo Souza (1998), os modelos de regressão não linear com resposta univariada y_i são da forma

$$y_i = f(x_i, \theta) + \varepsilon_i \ , \ i = 1, \ldots , n \qquad (1)$$

em que y_i representa a observação da variável dependente, $f(x_i, \theta)$ é a função esperança ou função resposta conhecida, x_i representa a observação da variável independente, $\theta = \left[\theta_1, \theta_1, \cdots, \theta_p\right]'$ é um vetor de parâmetros p dimensional desconhecido; ε_i representa o efeito do erro aleatório não observável suposto NIID com média zero e variância desconhecida σ^2.

Na Tabela 1 estão as funções $f(x_i, \theta) = f\left(x_i, \beta_1, \beta_2, \ldots, \beta_p\right)$ dos principais modelos de regressão não linear utilizados para descrever curvas de crescimento animal (SILVEIRA, 2011). Em geral, esses modelos têm por objetivo descrever uma trajetória assintótica da variável dependente peso, em função da variável independente tempo. Na maioria das vezes a diferença entre esses modelos é dada pela definição do ponto de inflexão da curva, que lhe confere uma forma sigmoide, porém para alguns modelos este ponto pode não existir.

Segundo Silveira (2011), os modelos apresentados na Tabela 1 são os mais utilizados para descrever curvas de crescimento. De modo geral, nesses modelos o parâmetro β_1 representa o peso adulto e o parâmetro β_3 a taxa de maturidade, ou velocidade de crescimento. O modelo Richards, que apresenta o parâmetro β_4, possui ponto de inflexão variável, cuja localização é determinada pelo parâmetro em questão. Os demais modelos ou apresentam o ponto de inflexão fixo, ou não o possuem, como é o caso do modelo Brody. Em geral, não há uma interpretação prática para o parâmetro β_2, sendo este uma constante matemática.

Tabela 1 – Modelos de regressão não linear para descrever curvas de crescimento

Modelo	**Modelo de Crescimento**
Richards	$y_i = \frac{\beta_1}{\left(1+e^{(\beta_2-\beta_3 x_i)}\right)^{\frac{1}{\beta_4}}} + e_i$
Gompertz	$y_i = \beta_1 e^{\left(-e^{\left(\beta_2-\beta_3 x_i\right)}\right)} + e_i$
Logístico	$y_i = \frac{\beta_1}{\left(1+e^{(\beta_2-\beta_3 x_i)}\right)} + e_i$
Brody	$y_i = \beta_1\left(1-\beta_2 e^{-\beta_3 x_i}\right) + e_i$
von Bertalanffy	$y_i = \beta_1\left(1-\beta_2 e^{-\beta_3 x_i}\right)^3 + e_i$

Perotto *et al*. (1992) relatam que Brody, Logístico, Gompertz e von Bertalanffy são casos especiais do modelo Richards. Este último, por ter o parâmetro β_4 flexível, que determina o ponto de inflexão, pode assumir a forma dos outros quatro modelos em questão. Por ter um parâmetro a mais, o modelo Richards geralmente apresenta excelentes ajustes, porém é de difícil convergência.

Para ajustar os modelos de regressão não linear é necessário estimar os parâmetros por meio de algum método de estimação, sendo o mais usual o dos Quadrados Mínimos Ordinários.

2.2.1 Método dos Quadrados Mínimos Ordinários

Considerando o modelo (1) do item 2.2 na forma matricial, tem-se:

$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$,

em que

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad f(\mathbf{x},\boldsymbol{\theta})=\begin{bmatrix} f(\mathbf{x_1},\boldsymbol{\theta}) \\ f(\mathbf{x_2},\boldsymbol{\theta}) \\ \vdots \\ f(\mathbf{x_n},\boldsymbol{\theta}) \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A soma dos quadrados dos erros aleatórios (SQE) deverá ser minimizada por $\boldsymbol{\theta}$, portanto a função de mínimos quadrados para um modelo não linear é dada por:

$$SQE(\boldsymbol{\theta})=\sum_{i=1}^{n}\left[y_i - f(x_i,\boldsymbol{\theta})\right]^2,$$

que pode ser representada matricialmente por:

$$SQE(\boldsymbol{\theta})=\left[y - f(\boldsymbol{\theta})\right]'\left[y - f(\boldsymbol{\theta})\right].$$

Segundo Souza (1998), em modelos não lineares não se pode fazer afirmações gerais sobre as propriedades dos estimadores de quadrados mínimos, como não tendenciosidade e variância mínima, exceto para grandes amostras, os chamados resultados assintóticos. Para melhor compreensão do processo de obtenção desses estimadores, utilizou-se a seguinte notação de diferenciação matricial:

$$f(\boldsymbol{\theta})=\underset{n}{\begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ \vdots \\ f_n(\boldsymbol{\theta}) \end{bmatrix}_1} \qquad \text{e} \qquad F(\theta)=\frac{\partial f(\boldsymbol{\theta})}{\partial \theta'}=\underset{n}{\begin{bmatrix} \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_p} \\ \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}_p},$$

em que

$f(\boldsymbol{\theta})$ é uma função vetor coluna $n\times 1$ de um argumento p dimensional $\boldsymbol{\theta}$; e $F(\theta)$ é a matriz jacobiana de $f(\boldsymbol{\theta})$. Dessa forma, o estimador de mínimos quadrados, $\hat{\boldsymbol{\theta}}$, satisfaz a equação $\left.\frac{\partial SQE(\boldsymbol{\theta})}{\partial \theta}\right|_{\theta=\hat{\theta}} = \varnothing$, que representa a minimização de interesse.

sendo

$$\frac{\partial SQE(\theta)}{\partial \theta'}=\frac{\partial}{\partial \theta'}\left[y - f(\theta)\right]'\left[y - f(\theta)\right]=-2\left[y - f(\theta)\right]' F(\theta),$$

tem-se:

$$F'\left(\hat{\theta}\right)\left[y - f\left(\hat{\theta}\right)\right] = \varnothing.$$

Portanto, o sistema de equações normais (SEN) para modelos não lineares é dado por:

$$\begin{bmatrix} \frac{\partial f_1\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_1} & \frac{\partial f_2\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_1} & \cdots & \frac{\partial f_n\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_1} \\ \frac{\partial f_1\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_2} & \frac{\partial f_2\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_2} & \cdots & \frac{\partial f_n\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_1\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_p} & \frac{\partial f_2\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_p} & \cdots & \frac{\partial f_n\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\theta}_p} \end{bmatrix} \cdot \left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} f_1\left(\hat{\boldsymbol{\theta}}\right) \\ f_2\left(\hat{\boldsymbol{\theta}}\right) \\ \vdots \\ f_n\left(\hat{\boldsymbol{\theta}}\right) \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

2.2.2 Processos Iterativos e Método dos Quadrados Mínimos Ordinários

Para o SEN não linear descrito no item anterior não existe uma solução explícita, sendo assim a solução para o sistema deve ser obtida por meio de processos iterativos. Um dos métodos iterativos é a linearização da função não linear, chamado Método de Gauss-Newton, que se resume ao seguinte procedimento.

Seja o modelo não linear $y_i = f\left(x_i, \theta\right) + \varepsilon_i$, e $\hat{\theta}_o$ um valor tal que $F'\left(\hat{\theta}_o\right)\left[Y - f\left(\hat{\theta}_o\right)\right] \approx 0$. Aproximando $f\left(\hat{\theta}\right)$ pelo ponto $\hat{\theta}_o$ por uma TSA (*Taylor Series Expansion*) de primeira ordem, tem-se:

$$f\left(\hat{\theta}\right) \approx f\left(\hat{\theta}_o\right) + F\left(\hat{\theta}_o\right)\left(\hat{\theta} - \hat{\theta}_o\right) \tag{2}$$

$$F'\left(\hat{\theta}\right)\left[Y - f\left(\hat{\theta}\right)\right] \approx \varnothing \tag{3}$$

Aplicando (2) em (3): $F'\left(\hat{\theta}\right)\left[Y - f\left(\hat{\theta}_o\right) - F\left(\hat{\theta}_o\right)\left(\hat{\theta} - \hat{\theta}_o\right)\right] \approx \varnothing$, e multiplicando à esquerda, ambos os lados da igualdade, por $\left[F'\left(\hat{\theta}\right)\right]^{-1}$, obtém-se:

$$Y - f\left(\hat{\theta}_o\right) - F\left(\hat{\theta}_o\right)\hat{\theta} + F\left(\hat{\theta}_o\right)\hat{\theta}_o \approx \varnothing.$$

Logo, $F\left(\hat{\theta}_o\right)\hat{\theta} \approx F\left(\hat{\theta}_o\right)\hat{\theta}_o + \left[Y - f\left(\hat{\theta}_o\right)\right]$. Multiplicando novamente à esquerda, ambos os lados da igualdade, por $\left[F\left(\hat{\theta}_o\right)\right]^{-1}$, verifica-se que: $\hat{\theta} \approx \hat{\theta}_o + \left[F\left(\hat{\theta}_o\right)\right]^{-1}\left[Y - f\left(\hat{\theta}_o\right)\right]$.

Fazendo $\hat{\theta} = \hat{\theta}_{k+1}$ e $\hat{\theta}_o = \hat{\theta}_k$, tem-se para a k-ésima iteração, a expressão (4), que representa o processo iterativo conhecido como Gauss-Newton:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \left[F\left(\hat{\theta}_k\right)\right]^{-1}\left[Y - f\left(\hat{\theta}_k\right)\right] \quad (4)$$

Esse processo iterativo prossegue até que algum critério adotado para convergência seja atingido. No presente trabalho adotou-se o seguinte critério: quando o máximo de $d_j < \delta$, sendo $d_j = \left|\frac{\left(\hat{\theta}_{j,k+1} - \hat{\theta}_{jk}\right)}{\hat{\theta}_{jk}}\right|$, para j=1, 2,..., p, interrompe-se o processo. O valor de δ foi especificado de acordo com a facilidade de convergência do modelo estudado.

2.3 Análise genética de curvas de crescimento

Um grande interesse dos melhoristas de animais é a obtenção de valores genéticos genômicos para pesos em idades não contempladas pelo conjunto de dados amostrais de peso-idade. Isto permite a identificação precoce de animais geneticamente superiores, o que reduz custos e tempo para avaliação de reprodutores jovens.

Para obtenção desses valores genéticos genômicos, pode-se utilizar uma abordagem composta por duas etapas. Na primeira, ajustam-se modelos de crescimento aos dados de peso-idade de cada animal submetido à seleção, com isso obtêm-se as estimativas dos parâmetros dos modelos. Na segunda, aplica-se a seleção (genética ou genômica), considerando como fenótipos as estimativas dos parâmetros obtidas na etapa anterior.

Silva (2010) utilizou a metodologia de modelos mistos a fim de realizar predições de valores genéticos para pesos não contemplados nos conjuntos de dados de bovinos da raça Nelore. Em seus estudos, o procedimento em duas etapas (*Two Step*) também foi adotado e o modelo de regressão não linear Brody modificado foi empregado para realizar estas predições.

Pong-Wong e Hadjipavlou (2010) utilizaram uma metodologia composta por ajuste de modelo de crescimento e Seleção Genômica Ampla em conjuntos de dados simulados. Para o problema em questão, os autores adotaram dois passos subsequentes: primeiramente ajustaram o modelo de crescimento Gompertz a dados simulados de peso-idade e obtiveram os parâmetros estimados para cada indivíduo avaliado e, posteriormente, a GWS foi aplicada para obter valores genéticos genômicos para os fenótipos (parâmetros estimados) provenientes do primeiro passo.

Em estudos que envolvem Seleção Genômica Ampla é importante ter um bom conhecimento sobre os métodos estatísticos utilizados.

2.4 Métodos estatísticos para a Seleção Genômica Ampla

2.4.1 Método RR-BLUP/GWS

Este método utiliza preditores do tipo BLUP (*Best Linear Unbiased Prediction*), mas os efeitos de marcadores não são ajustados como variáveis classificatórias, mas sim como explicativas ou explanatórias. Portanto, essas são variáveis regressoras ajustadas como covariáveis de efeitos aleatórios, ou seja, os fenótipos são regredidos com base nessas covariáveis. O nome mais apropriado é Regressão Aleatória (*Random Regression*) do tipo BLUP (RR-BLUP) aplicada à Seleção Genômica Ampla (RR-BLUP/GWS), sendo esta um tipo especial da regressão de cumeeira (*Ridge Regression*) (RESENDE *et al*., 2010).

A predição via RR-BLUP/GWS é descrita a seguir, com base em Resende (2008).

O seguinte modelo linear misto geral é ajustado para estimar os efeitos dos marcadores:

$$y = Xb + Zm + e, \quad (5)$$

em que y é um vetor de observações fenotípicas; b é um vetor de efeitos fixos; m é o vetor de efeitos dos marcadores assumidos como aleatórios; e se refere ao vetor de erros aleatórios; e X e Z são as matrizes de incidência para b e m. A matriz de incidência Z contém os valores 0, 1 e 2 para o número de alelos.

As equações de modelo misto para predição de m por meio do método RR-BLUP/GWS equivalem a:

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

em que σ_g^2 se refere à variância genética da característica, σ_e^2 à variância residual e *n* é função do número total de marcadores ponderados por suas frequências alélicas, sendo dado por $n = 2\sum_i p_i(1-p_i)$, em que p_i é a frequência do alelo *i*.

Nesse método considera-se que cada loco explica $(1/n)\sigma_g^2$, ou seja, partes iguais da variância genética são atribuídas a todos os locos.

O valor genético genômico do indivíduo *j* é dado por:

$$\left(G\hat{B}V = \hat{y}_j = \hat{\mu} + \sum_i Z_i\hat{m}_i\right)$$

em que Z_i equivale a 0, 1 ou 2 para os genótipos dos tipos aa, Aa e AA, respectivamente, para marcadores bialélicos e codominantes como os SNPs.

2.4.2 Métodos Bayesianos

Meuwissen *et al*. (2001) apresentam diversos métodos como possíveis abordagens para predição de valores genéticos com base em informações genômicas. Para isso, o seguinte modelo linear básico foi proposto:

$$y = 1\mu + \sum_i X_i g_i + e \qquad (6)$$

em que y é o vetor de fenótipos; 1 é o vetor de mesma dimensão de y com todas as entradas iguais a 1; μ é a média da característica estudada; g_i é o vetor de efeitos aditivos dos diferentes haplótipos (ou diferentes marcadores SNPs) do segmento *i* do genoma; X_i é a matriz de incidência que relaciona efeitos de haplótipos aos fenótipos contidos em y; e e é o vetor de resíduos do modelo. O somatório ($\sum_i$) indica que é em relação a todos marcadores simultaneamente.

A seguir tem-se a descrição de alguns desses métodos:

- Bayes A

Meuwissen *et al*. (2001) apresentam também uma metodologia para estimar, por meio da abordagem Bayesiana, os parâmetros do modelo (6) estendido, no qual

diferentes componentes de variância $\sigma^2_{g_i}$ são atribuídos para cada segmento considerado na análise. Considerou-se, em um segundo nível hierárquico, um modelo para $\sigma^2_{g_i}$ com o intuito de realizar inferências a respeito desses parâmetros.

Os autores assumiram, como densidades *a priori,* distribuição normal para os efeitos de haplótipo, uma constante para a média geral e distribuição qui-quadrado invertida com parâmetro de escala para a variância residual. Uma distribuição da mesma família dessa última foi considerada *a priori* para a variância do efeito de cada segmento, com o intuito de que a distribuição *a posteriori* de cada variância seja uma combinação daquela com as informações contidas nas observações. Essas distribuições utilizadas na construção da densidade *a priori* conjunta resultam em condicionais completas *a priori* com forma conhecida, o que possibilita a utilização de amostrador de Gibbs (GEMAN; GEMAN, 1984) para gerar amostras da densidade conjunta *a posteriori* (e por consequência as marginai*s a posteriori* de interesse).

- Bayes B

Adicionalmente aos modelos já apresentados, Meuwissen *et al.* (2001) desenvolveram uma abordagem Bayesiana alternativa. Os autores reconheceram como um problema no método Bayes A o fato de a distribuição das variâncias dos efeitos de haploides não apresentar uma massa de densidade no valor 0. Esta característica seria interessante para essa distribuição, uma vez que a maior parte dos segmentos não apresenta variância genética (por não carregarem segregação do QTL por falta de desequilíbrio de ligação com o mesmo). O método Bayes B utiliza densidade *a priori* com massa de densidade em $\sigma^2_{g_i} = 0$. Considera-se que $\sigma^2_{g_i} = 0$ com probabilidade π, enquanto $\sigma^2_{g_i} \sim inv - \chi^2(v, S)$ com probabilidade $1 - \pi$, S é um parâmetro de escala e v é o número de graus de liberdade.

Em princípio, é possível construir um amostrador de Gibbs para essa abordagem, mas ao fazê-lo a cadeia de Markov não visita todo o espaço amostral necessário, uma vez que $\sigma^2_{g_i} = 0$ tem probabilidade extremamente baixa se ${g_i}'g_i$ é maior que zero, e ${g_i}'g_i = 0$ tem probabilidade extremamente baixa se $\sigma^2_{g_i} > 0$. Considerando y como vetor de observações livre de efeitos da média e efeitos genéticos, com exceção de haplótipos da região *i*, a solução para a amostragem de g_i

e $\sigma^2_{g_i}$ pode ser obtida por meio de $p(\sigma^2_{g_i}, g_i|y) \sim p(\sigma^2_{g_i}|y)p(g_i|\sigma^2_{g_i}, y)$, de modo que a amostragem de $\sigma^2_{g_i}$ não seja função de g_i. Entretanto, os autores não obtiveram $p(\sigma^2_{g_i}|y)$ de forma fechada, ou seja, como sendo uma distribuição de probabilidade conhecida. Assim, foi necessária a utilização do algoritmo Metropolis-Hastings (GELMAN *et al*., 2004) para conseguir amostras de $p(\sigma^2_{g_i}|y)$.

- LASSO Bayesiano

Como já apresentado por Meuwissen *et al*. (2001), a regressão Bayesiana pode ser utilizada nas situações em que se têm mais marcadores (covariáveis) do que observações, atribuindo distribuição *a priori* aos coeficientes de regressão. Essas distribuições impõem regularização no ajuste do modelo, sob a forma de encurtamento dos coeficientes de regressão (*shrinkage*). Entretanto, os autores também demonstraram que a forma dessa regularização deve ser diferenciada, pois se espera que quando se tem SNPs de todo o genoma de um indivíduo muitos marcadores estarão em regiões que não influenciam o valor mensurado da característica, enquanto poucos estarão em desequilíbrio de ligação com alelos que influenciam a característica.

Uma abordagem interessante para a situação descrita é a utilização do método de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*, TIBSHIRANI, 1996), que combina seleção de variáveis e regularização via encurtamento dos coeficientes de regressão. A implementação Bayesiana da regressão LASSO (PARK; CASELLA, 2008) foi adaptada para seleção genômica por De Los Campos *et al*. (2009). Nesta adaptação, informações de parentesco e outras covariáveis que não sofrem o efeito da regularização são consideradas no modelo.

A metodologia LASSO consiste na obtenção de estimadores de coeficientes de regressão que resolvam o seguinte problema de otimização:

$$min\{\sum_i (y_i - x_i\beta)^2 + \lambda(t)\ \sum_j|\beta_j|\}$$

em que y_i são observações do indivíduo *i;*β é o vetor de coeficientes de regressão; e x_i é um vetor que contém covariáveis. $\sum_j|\beta_j|$ é a soma dos valores absolutos dos coeficientes de regressão contidos no vetor β, de modo que soluções nas quais os coeficientes de regressão se afastam de 0 sofrem penalização. Adicionalmente, $\lambda(t)$ é

um parâmetro de suavização que controla a força da regularização. Quando este último parâmetro é igual a zero, não há regularização. No LASSO Bayesiano, esse parâmetro controla a precisão da distribuição *a priori* atribuída aos coeficientes de regressão.

A implementação desse tipo de regularização envolve encurtamento mais forte no sentido de que alguns coeficientes de regressão tenham valores iguais a zero, o que pode ser demonstrado de diversas maneiras. Uma alternativa é pela própria implementação Bayesiana do LASSO (De LOS CAMPOS *et al*., 2009). Esta implementação impõe como distribuição *a priori* dos p coeficientes de regressão um produto de densidades exponenciais duplas: $p(\beta|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right)$. Por sua vez, a regressão Bayesiana-padrão utiliza distribuição normal multivariada:

$$p\left(\beta|\sigma_\beta^2\right) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left(-\frac{\beta_j}{2\sigma_\beta^2}\right).$$

As duas distribuições podem ser comparadas na Figura 1, na qual se observa que a densidade *a priori* utilizada no LASSO Bayesiano (curva sólida) apresenta maior massa de densidade no valor zero e caudas mais robustas, exercendo maior encurtamento sobre coeficientes de regressão próximos de 0 e menor encurtamento sobre coeficientes de regressão distantes de zero.

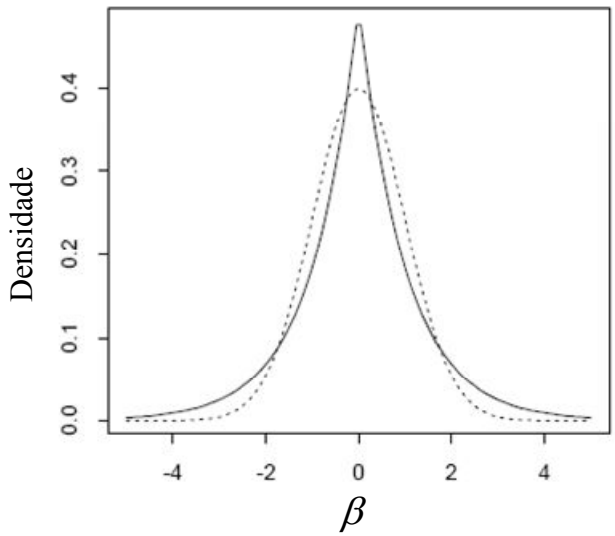


Figura 1 – Densidades das distribuições normal (curva pontilhada) e exponencial dupla (curva sólida), ambas com médias iguais a zero e variâncias iguais à unidade.

O modelo proposto por De Los Campos *et al*. (2009) é representado por:

$y_i = \mu + x'_{ri}\beta_r + x'_{li}\beta_l + e_i$, em que y_i é um fenótipo mensurado no indivíduo *i*; μ é média da característica estudada; x'_{ri} e x'_{li} são covariáveis atribuídas ao individuo *i* a serem tratadas por regressão Bayesiana-padrão e LASSO Bayesiano, respectivamente; β_r e β_l são coeficientes de regressão Bayesiana-padrão e LASSO Bayesiano, respectivamente; e e_i é o resíduo aleatório do modelo. Adicionalmente, será considerado que $e \sim N(0, \sigma_e^2)$.

Para construção da distribuição conjunta *a priori* dos parâmetros, os autores exploram o fato de a distribuição exponencial dupla poder ser representada como uma mistura de densidades normais com parâmetro de escala, com o processo de mistura de variâncias controlado por distribuição exponencial. Na distribuição *a priori* conjunta construída, a densidade atribuída aos coeficientes de regressão regularizados por LASSO será: $\prod_{j=1}^{p} N\left(\beta_{j1}|0, \sigma_e^2\tau_j^2\right)$, resultando em variância específicas para cada coeficiente de regressão. Por sua vez, a distribuição *a priori* para o parâmetro de escala τ_j^2 será representada por: $\prod_{j=1}^{p} \exp\left(\tau_j^2|\lambda\right)$, pela qual o parâmetro de suavização λ influencia o ajuste dos coeficientes de regressão. A informação *a priori* para esse parâmetro é dada por uma distribuição com hiperparâmetros conhecidos. Caso a distribuição escolhida seja conjugada, é possível obter amostras da distribuição *a posteriori* conjunta por meio de um amostrador de Gibbs.

Ao ser comparado com outros métodos de regressão Bayesiana, como o Bayes A (MEUWISSEN *et al*., 2001), as diferenças estão na distribuição *a priori* atribuídas às variâncias específicas para diferentes marcadores. No método Bayes A, a distribuição marginal *a priori* dos coeficientes de regressão são distribuições de t, que levam vantagem em relação à distribuição normal por apresentarem maior densidade em 0. Entretanto, a densidade em zero para a mesma marginal *a priori* no LASSO Bayesiano é ainda maior, o que é um aspecto desejável dessa metodologia.

3 MATERIAL E MÉTODOS

3.1 Dados simulados

3.1.1 Descrição dos dados

Os dados simulados utilizados neste estudo são oriundos do QTLMAS2009 (*Workshop of Quantitative Trait Loci Mapping and Marker Assisted Selection*), realizado na Holanda, em abril de 2009. O conjunto de dados consiste de 2.025 indivíduos de duas gerações. Todos os indivíduos têm informações completas de marcadores. Existem 453 marcadores SNPs, que estão aleatoriamente distribuídos sobre cinco cromossomos. Os primeiros 25 indivíduos são pais, 20 fêmeas e cinco machos. Os 2.000 indivíduos restantes são descendentes, 100 famílias de irmãos completos, uma de cada combinação entre machos e fêmeas. Cada família de irmãos completos tem 20 descendentes.

Cinquenta famílias (população de treinamento) possuem registros fenotípicos de produção (pesos), as outras 50 (população de validação) não têm informações fenotípicas. Fenótipos foram registrados em cinco momentos distintos (0, 132, 265, 397, 530 dias). As famílias fenotipadas foram escolhidas de tal forma que cada fêmea tem pelo menos 40 descendentes fenotipados, enquanto cada macho tem 100 descendentes fenotipados. A Figura 2 é uma representação gráfica da estrutura de combinação utilizada para simular a geração 2.

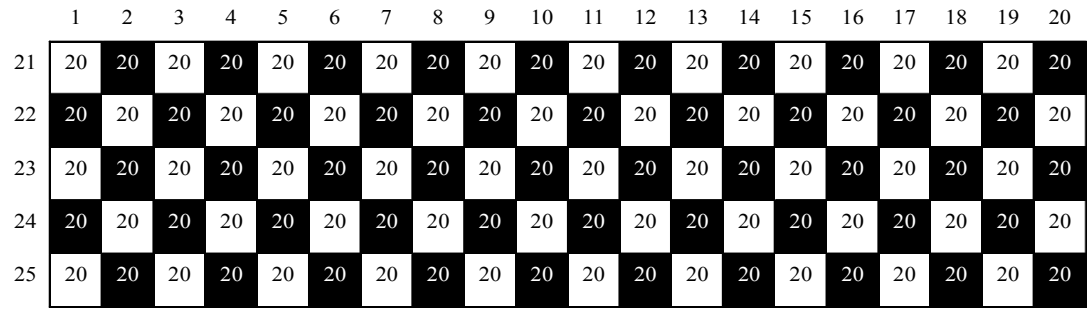


Figura 2 – Representação gráfica da segunda geração simulada. Cada *i,j* representa uma família de irmãos completos simulada pelo combinação de uma fêmea *i* e um macho *j*. Células pretas representam famílias de irmãos completos das quais dados fenotípicos foram simulados; células brancas representam famílias de irmãos completos das quais dados fenotípicos não foram simulados. Cada família de irmãos completos consiste de 20 descendentes.

Todo o conjunto de dados utilizado está disponível em: <http://www.qtlmas 2009.wur.nl/UK/Dataset/>.

3.1.2 Metodologia

As análises foram realizadas em dois passos subsequentes. Primeiramente, aos dados de peso-idade de cada indivíduo da população de treinamento ajustou-se o seguinte modelo de crescimento Logístico reparametrizado:

$$y_i = \frac{\emptyset_1}{1 + exp[(\emptyset_2 - t_i)/\emptyset_3]} + e_i \qquad (7)$$

em que y_i é o peso no tempo t_i; $\emptyset_1$ é o peso adulto (peso assintótico); $\emptyset_2$ é a abscissa do ponto de inflexão da curva; $\emptyset_2 + \emptyset_3$ é a abscissa referente ao ponto no qual y_i corresponde a aproximadamente 73% do peso adulto; e e_i é o efeito do erro aleatório. Em termos práticos, quanto menor os valores $\emptyset_2$ e $\emptyset_3$ e maior o valor de $\emptyset_1$, mais eficiente é o padrão de crescimento. A Figura 3 mostra um esboço do modelo (7), com a interpretação gráfica dos três parâmetros.

Para ajustar o modelo de regressão não linear (7) aos dados de crescimento, utilizou-se o método dos quadrados mínimos ordinários, cujas soluções foram obtidas por meio do processo iterativo de Gauss-Newton. Os ajustes foram realizados por meio do PROC MODEL do programa SAS® (SAS, 2003).

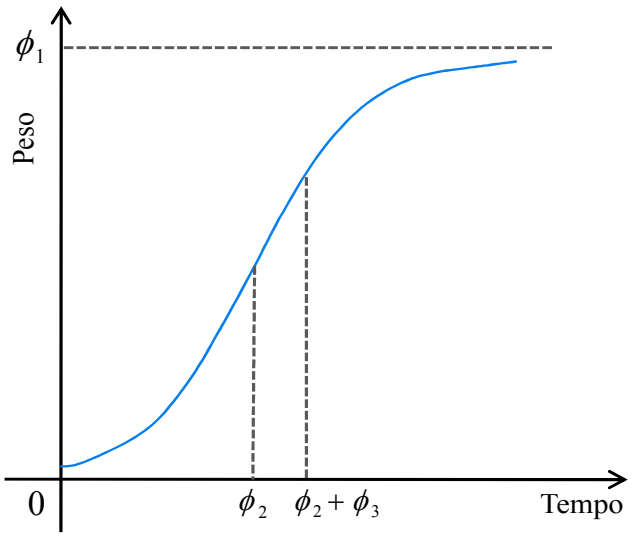


Figura 3 – Curva de crescimento Logística. A linha horizontal é o peso assintótico, representado pelo parâmetro $\emptyset_1$ no modelo de crescimento Logístico. A primeira linha vertical (tempo $= \emptyset_2$) é a abscissa do ponto de inflexão da curva. A segunda linha vertical (tempo $= \emptyset_2 + \emptyset_3$) é a abscissa do ponto onde o peso é aproximadamente 73% do peso adulto.

Em um segundo passo, as estimativas dos parâmetros, obtidas para cada indivíduo no passo anterior, foram consideradas como fenótipos em análises utilizando o método LASSO Bayesiano. Nessas análises a utilização dos parâmetros estimados como fenótipos permite compreender a arquitetura genética de toda a trajetória, uma vez que todas as pesagens são resumidas nessas estimativas.

No método LASSO Bayesiano considerou-se o seguinte modelo:

$$Y = 1\mu + X\beta + \varepsilon\,,$$

em que Y é o vetor de estimativas obtidas para $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$; μ é a média geral; β é o vetor de efeito dos marcadores; X é a matriz de incidência destes marcadores (marcadores codominantes SNPs codificados como 0, 1 ou 2, de acordo com o número de cópias de um dos alelos do loco marcador); e o termo ε corresponde ao erro aleatório, $\varepsilon \sim N(0, \sigma_e^2)$.

Para exemplificar a execução do método, descreve-se o vetor Y de estimativas obtidas para $\emptyset_1$, a matriz X_T de incidência para os efeitos dos marcadores SNPs para indivíduos da população de treinamento e β vetor de efeitos dos marcadores a ser estimado:

$$Y = \begin{bmatrix} 19{,}53 \\ 37{,}53 \\ 31{,}30 \\ \vdots \\ 36{,}48 \end{bmatrix}_{1000\times1}, \quad X_T = \begin{bmatrix} 1 & 0 & 2 & \dots & 1 \\ 1 & 0 & 1 & \dots & 2 \\ 0 & 0 & 1 & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 2 & 1 & 1 & 0 & 0 \end{bmatrix}_{1000\times453} \quad e \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}_{453\times1}$$

A metodologia consiste na obtenção das estimativas dos coeficientes de regressão ($\hat{\beta}_j, com\ j = 1, 2, \cdots, p$) que resolvam o seguinte problema de otimização:

$$min\{(Y - X_T\beta)'(Y - X_T\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

em que $\sum_{j=1}^{p} |\beta_j|$ é a soma dos valores absolutos dos coeficientes de regressão contidos no vetor β, de modo que soluções nas quais os coeficientes de regressão se afastam de 0 sofrem penalização. Adicionalmente, λ é um parâmetro de suavização que controla a força da regularização. Quando esse último parâmetro é igual a zero, não há regularização. No LASSO Bayesiano, esse parâmetro controla a precisão da distribuição *a priori* atribuída aos coeficientes de regressão. A execução desse tipo de regularização envolve encurtamento mais forte no sentido de que alguns coeficientes de regressão tenham valores iguais a zero.

As estimativas dos efeitos dos marcadores ($\hat{\beta}_j, com\ j = 1, 2, \cdots, p$) e da média $\hat{\mu}$ foram obtidas com a utilização do pacote BLR (*Bayesian Linear Regression*), disponível do programa R (R *Development Core Team*, 2010). A implementação Bayesiana da regressão LASSO (PARK; CASELLA, 2008) presente nesse pacote foi adaptada para seleção genômica por De Los Campos *et al*. (2009).

De posse do vetor estimado $\hat{\beta}$ e da média $\hat{\mu}$, prosseguiu-se com a obtenção do valor genético genômico predito $\left(G\hat{B}V_T\right)$ para cada indivíduo da população de treinamento, como exemplificado a seguir para marcador codominante (SNP):

$$G\hat{B}V_T = \hat{\mu} + \left(X_T \times \hat{\beta}\right)$$

$$G\hat{B}V_T = \begin{bmatrix} 30{,}4976 \\ 30{,}4976 \\ 30{,}4976 \\ \vdots \\ 30{,}4976 \end{bmatrix}_{1000\times1} + \begin{bmatrix} 1 & 0 & 2 & \dots & 1 \\ 1 & 0 & 1 & \dots & 2 \\ 0 & 0 & 1 & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 2 & 1 & 1 & 0 & 0 \end{bmatrix}_{1000\times453} \times \begin{bmatrix} -0{,}1871 \\ -0{,}0990 \\ 0{,}0887 \\ \vdots \\ 0{,}1773 \end{bmatrix}_{453\times1} = \begin{bmatrix} 28{,}8778 \\ 33{,}2067 \\ 36{,}3633 \\ \vdots \\ 40{,}6342 \end{bmatrix}_{1000\times1}$$

A obtenção do valor genético genômico predito para cada indivíduo da população de validação $\left(G\hat{B}V_V\right)$ foi realizada da mesma forma, tomando o cuidado em substituir matriz X_T pela matriz X_V, que se refere à incidência para os efeitos dos marcadores codominantes SNPs para indivíduos da população de validação.

$$G\hat{B}V_V = \hat{\mu} + \left(X_V \times \hat{\beta}\right)$$

$$G\hat{B}V_V = \begin{bmatrix} 30{,}4976 \\ 30{,}4976 \\ 30{,}4976 \\ \vdots \\ 30{,}4976 \end{bmatrix}_{1000\times1} + \begin{bmatrix} 0 & 2 & 1 & \dots & 0 \\ 0 & 2 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 2 & 2 & 1 & 1 \end{bmatrix}_{1000\times453} \times \begin{bmatrix} -0{,}1871 \\ -0{,}0990 \\ 0{,}0887 \\ \vdots \\ 0{,}1773 \end{bmatrix}_{453\times1} = \begin{bmatrix} 33{,}7189 \\ 38{,}4528 \\ 36{,}9261 \\ \vdots \\ 35{,}9370 \end{bmatrix}_{1000\times1}$$

Os coeficientes de correlação e regressão linear envolvendo valores genéticos genômicos verdadeiros (GBV fixados na simulação) e preditos ($G\hat{B}V$) foram utilizados a fim de medir a capacidade do método em predizer de forma acurada e não viesada, respectivamente. A correlação fornece a acurácia, e no caso em que envolve valores fenotípicos e $G\hat{B}V$ recebe o nome de capacidade preditiva. O coeficiente de regressão indica a presença de viés, de modo que coeficientes de regressão abaixo de 1 indicam que os valores genéticos genômicos são superestimados e apresentam variabilidade além da esperada, e acima de 1 indicam que os valores genéticos genômicos estimados apresentam variabilidade aquém da esperada. 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 (RESENDE *et al*., 2010).

Com o intuito de identificar animais de melhor desempenho, foi realizada uma análise de agrupamento considerando como variáveis todos os valores genômicos estimados $\left(G\hat{B}Vs\right)$ nas duas populações. Para tanto, foi utilizado o PROC CLUSTER do programa SAS® (SAS, 2003) com o método centroide e distância entre dois grupos dada pela distância euclidiana quadrática entre os vetores de médias.

Por ordem de aplicação prática, optou-se por dividir os animais em dez grupos distintos, sendo o PROC TREE utilizado para visualização do dendrograma e verificação de animais pertencentes aos diferentes grupos obtidos pela discriminação estatística em relação aos vetores de $GBVs$ estimados.

Dentro de cada grupo foram calculadas as médias dos $G\hat{B}Vs$, sendo essas substituídas no modelo de crescimento Logístico (7) para a construção de um gráfico contendo dez curvas, uma para cada grupo. Com a análise gráfica pode-se identificar o grupo mais eficiente, ou seja, aquele em que a curva ajustada demonstra melhor desempenho de produção no intervalo de tempo considerado (0 aos 530 dias).

Com o objetivo de detectar QTLs de efeitos mais expressivos em nível genômico, foi obtido o quantil 95% da distribuição empírica dos módulos dos efeitos estimados dos SNPs para análises que envolviam, respectivamente, estimativas de $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$ (Figura 4). Assim, foi possível identificar os SNPs com maiores efeitos em módulo e suas respectivas posições por meio do mapa de posição dos marcadores disponível no arquivo intitulado "*Map File*" fornecido pelo QTLMAS2009.

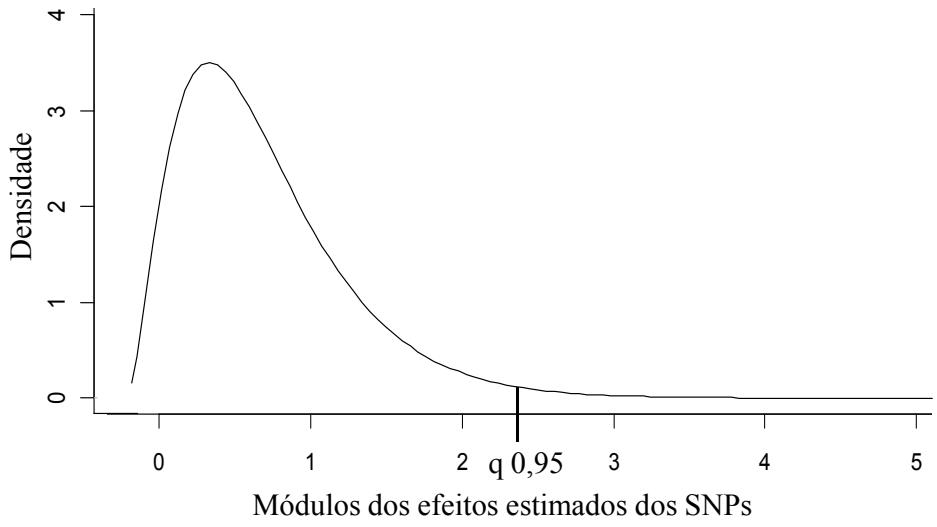


Figura 4 – Esquema ilustrativo do método de identificação de marcadores com maiores efeitos em módulo.

As posições encontradas para marcadores com maiores efeitos nas análises envolvendo, respectivamente, estimativas de $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$ foram comparadas com as posições simuladas dos marcadores disponíveis no arquivo nomeado *Simulation Values*, também fornecido pelo QTLMAS2009. Este procedimento foi realizado a fim de verificar a correspondência entre posições encontradas e simuladas de QTLs mais expressivos.

3.2 Dados reais

3.2.1 Descrição dos dados

A formação da população e a coleta dos dados fenotípicos foram realizadas na Granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa (UFV), em Viçosa, Minas Gerais, Brasil, no período de novembro de 1998 a julho de 2001.

Foi construída uma população F_2 (345 suínos) proveniente do cruzamento de dois varrões da raça naturalizada brasileira Piau, com 18 fêmeas de linhagem desenvolvida na UFV, pelo acasalamento de animais de raça comercial (Landrace × Large White × Pietrain). Indivíduos F_2 possuem informações fenotípicas (pesos) em sete momentos distintos (0, 21, 42, 63, 77, 105, 150 dias), exceto para alguns animais que possuem informações perdidas em determinadas idades. Maiores detalhes sobre a constituição da população F_2 foram descritos por Peixoto *et al*. (2006). A extração do DNA foi realizada no Laboratório de Biotecnologia Animal do Departamento de Zootecnia da Universidade Federal de Viçosa, e detalhes dos procedimentos usados podem ser encontrados em Peixoto *et al*. (2006).

Os SNPs utilizados para o mapeamento fino foram selecionados de acordo com seu espaçamento entre cromossomos que continham QTLs previamente detectados nessa população e foram distribuídos da seguinte forma nos cromossomos de *Sus scrofa*: SSC1 (85), SSC4 (71), SSC7 (84), SSC8 (42), SSC17 (36) e SSCX (66). A genotipagem para os 384 SNPs foi realizada via tecnologia Golden Gate/VeraCode®, que é uma plataforma robusta e flexível, para o leitor BeadXpress de Illumina, no Laboratório de Genética Animal (LGA), Embrapa Recursos Genéticos e Biotecnologia (CENARGEN), Brasília, DF. Destes, 66 SNPs foram descartados devido à ausência de amplificação, e dos 318 SNPs restantes 81 foram descartados por apresentar menor frequência alélica (MAF < 0,05). Após esses procedimentos, a distribuição de SNPs foi: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25) e SSCX (12), totalizando 237 marcadores (Tabela 2). As posições físicas desses marcadores foram obtidas em um banco de dados de suínos construído pelo Ensembl (disponível em: <http://www.ensembl.org>).

Tabela 2 – Número total de marcadores SNPs, comprimento dos cromossomos e distância média entre os marcadores nos cromossomos 1, 4, 7, 8, 17 e X de *Sus scrofa*

SSC	Número Total de Marcadores SNPs	Comprimento do Cromossomo (cM)	Distância Média (cM)
1	56	290	5,18
4	54	128	2,37
7	59	133	2,25
8	31	118	3,81
17	25	67	2,68
X	12	132	11,00

Para análise de ligação, construiu-se o mapa de ligação, considerando a mesma ordem dos marcadores adotada no mapa físico. As distâncias genéticas entre SNPs foram extrapoladas em relação às distâncias físicas (1Mb = 1cM).

3.2.2 Metodologia

As análises foram realizadas em dois passos subsequentes, como naqueles descritos no item 3.1.2 para dados simulados. No passo 1 foram efetuadas duas modificações: indivíduos em que o modelo Logístico não atingiu a convergência foram eliminados e parâmetros estimados do modelo foram corrigidos, visando à eliminação dos efeitos dos genitores e à desregressão dos valores genéticos. Segundo Resende *et al*. (2010), esses devem ser desregressados por três motivos: (i) não pode haver duas regressões: uma baseada em *pedigree* e outra baseada em marcadores; (ii) a matriz A (contendo coeficientes de parentesco de Wright) baseada em *pedigree* é menos precisa que a ZZ´ baseada em marcas; e (iii) influência de genes de grande efeito presentes em um dos genitores.

Ainda de acordo com Resende *et al*. (2010), os fenótipos originais também 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". O efeito da segregação mendeliana proporciona a análise da associação de alelos de marcas e de QTLs, ou seja, desequilíbrio de ligação (*linkage disequilibrium* - LD) livre de genealogia.

A correção para os fenótipos (parâmetros estimados do modelo Logístico) foi realizada por meio de duas formas distintas. Na primeira, corrigiram-se efeitos de

sexo, lote e presença do gene halotano, gene amplamente conhecido e divulgado na literatura especializada em melhoramento de suínos (BAND *et al*., 2005). Na segunda, apenas efeitos de sexo e lote foram corrigidos, sendo o gene halotano considerado um marcador adicional.

Especificamente para os dados reais, a fim de garantir maior precisão dos resultados, utilizou-se, além do método LASSO Bayesiano, o método RR-BLUP/GWS (RESENDE, 2008). Neste método considerou-se o modelo misto descrito em (5), no item 2.4.1:

$$y = Xb + Zm + e$$

em que y é um vetor de observações fenotípicas; b é um vetor de efeitos fixos; m é o vetor de efeitos dos marcadores assumidos como aleatórios; e se refere ao vetor de erros aleatórios; e X e Z são as matrizes de incidência para b e m. No presente estudo Xb foi considerado como sendo uma constante μ.

Da mesma forma que no método LASSO, também foram realizadas três análises, sendo em cada uma considerado em y o vetor de parâmetros estimados corrigidos obtidos para $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$, respectivamente.

A equação de modelos mistos utilizada para predição dos valores genômicos foi:

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

em que σ_g^2 se refere a variância genética da característica; σ_e^2 a variância residual; e n é o número de marcadores corrigidos por suas frequências alélicas. A metodologia em questão foi empregada por meio do programa Selegen Genômica (RESENDE, 2007), que também disponibiliza o valor de $\hat{m}$ e o valor genético genômico estimado de cada indivíduo $\left(G\hat{B}V = \hat{y}_j = \hat{\mu} + \sum_i Z_i \hat{m}_i\right)$.

Os métodos LASSO Bayesiano e RR-BLUP/GWS foram aplicados para a população total de 265 indivíduos, sendo a análise de validação cruzada realizada apenas no método RR-BLUP/GWS, por meio do programa Selegen Genômica (RESENDE, 2007). A validação cruzada foi realizada de acordo com um procedimento Jacknife, de maneira que em cada repetição um indivíduo foi removido do conjunto de dados para compor a população de validação e os outros 264 indivíduos foram utilizados na estimação dos valores genéticos genômicos, ou seja,

constituíam a população de treinamento. Uma vez estimados todos os efeitos, estes eram aplicados na população de validação para predizer o valor genético genômico e somados à média geral estimada para compor o fenótipo estimado na população de validação. Uma vez que em cada repetição um indivíduo foi utilizado na validação, ao final da análise obteve-se o fenótipo estimado para todos os 265 indivíduos, sem que isso comprometesse a independência necessária entre a estimação e a validação.

As análises por meio desses dois métodos também foram realizadas, considerando o efeito do gene halotano. Conhecido como gene do estresse, o halotano surgiu de uma mutação no cromossomo 6 de suínos e está associado com carne PSE (pálida, flácida e exsudativa). Sua presença no suíno contribui para o aumento do porcentual de carne na carcaça (BAND *et al*., 2005), porém provoca aumento de mortes súbitas, especialmente na movimentação e no transporte dos animais, quando não manejados adequadamente. O gene halotano tem sido explorado para condicionar aumento de carne na carcaça, cruzando-se machos terminais heterozigotos (Hal^{Nn}) com fêmeas homozigotas livres do alelo recessivo (Hal^{NN}). Esse procedimento objetiva chegar a uma progênie 50% Hal^{Nn} e 50% Hal^{NN}, com o aumento de 1 a 2% no conteúdo de carne nas carcaças e, supostamente, sem prejuízo para sua qualidade.

As análises em que o gene halotano é considerado como marcador são úteis para comparar as metodologias RR-BLUP/GWS e LASSO Bayesiano, uma vez que no primeiro método assume-se *a priori* que todos os locos explicam iguais quantidades da variação genética e no segundo assume-se *a priori* variâncias diferentes para efeitos de marcador. O método LASSO Bayesiano foi implementado, considerando um total de 10.000 amostras salvas para o algoritmo Gibbs Sampler, com um descarte inicial de 5.000 e um intervalo de 100 entre amostras salvas. Em relação à detecção de QTLs, considerando a análise dos dados reais, foram usados os mesmos procedimentos apresentados no item 3.1.2 (Figura 4).

4 RESULTADOS E DISCUSSÃO

4.1 Dados simulados

As acurácias (estimativas dos coeficientes de correlação entre GBV e $G\hat{B}V$) obtidas nas duas populações para os três parâmetros analisados foram elevadas, o que indica a capacidade do método LASSO em predizer de forma acurada os valores genéticos genômicos (Tabela 3). Nota-se, também, que as acurácias referentes à população de validação foram um pouco menores do que as da população de treinamento. Esse resultado já era esperado, uma vez que os valores genéticos genômicos dos indivíduos de validação são preditos com base nos efeitos dos marcadores estimados por meio da população de treinamento.

Ainda na Tabela 3, nota-se que as estimativas dos coeficientes de regressão linear que envolvem valores verdadeiros e preditos foram próximas de 1, mostrando que as avaliações foram não viesadas e efetivas em predizer as reais magnitudes das diferenças entre os indivíduos nas populações de treinamento e validação. Esse método de avaliação da eficiência é recomendado por Resende *et al*. (2010), que relataram que predições não viesadas relacionam-se com coeficientes de regressão próximos a 1.

De acordo com item 3.1.2, foi realizada uma análise de agrupamento considerando como variáveis os valores genéticos genômicos estimados $(G\hat{B}Vs)$ para cada parâmetro do modelo Logístico (7). Por razões práticas, considerou-se o número de grupos igual a dez, de modo que ao substituir os parâmetros do modelo

Tabela 3 – Estimativas dos coeficientes de correlação (r) e regressão linear (b) envolvendo valores genéticos genômicos verdadeiros (GBV) e preditos $(G\hat{B}V)$ nas populações e parâmetros analisados

Populações	**Coeficientes**	**Parâmetros**		
		$\emptyset_1$	$\emptyset_2$	$\emptyset_3$
Treinamento	$r_{GBV_T,G\hat{B}V_T}$	0,93	0,80	0,91
	$b_{GBV_T,G\hat{B}V_T}$	1,05	0,96	0,98
Validação	$r_{GBV_V,G\hat{B}V_V}$	0,92	0,79	0,85
	$b_{GBV_V,G\hat{B}V_V}$	1,02	1,14	0,89

Logístico pelos $(G\hat{B}Vs)$ médios de cada parâmetro (ϕ_1, ϕ_2 e ϕ_3) de cada grupo foram obtidas dez curvas de crescimento distintas, uma para cada grupo, as quais podem ser consideradas geneticamente diferentes. Essas curvas estão apresentadas na Figura 5.

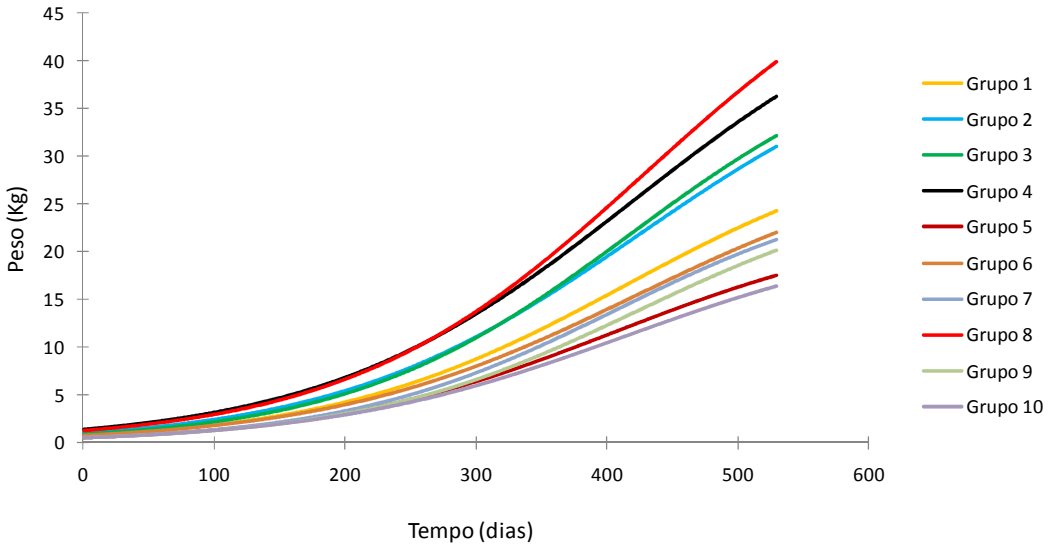


Figura 5 – Curvas de crescimento genéticas referentes ao ganho de peso (em kg) no período de 0 a 530 dias, para dez grupos provenientes da análise de agrupamento.

Dessa forma, nota-se o maior desempenho dos grupos 8, 4, 3 e 2 em relação aos demais. Dentre estes, o grupo 8 foi o que apresentou maior eficiência de crescimento dentro da amplitude de tempo considerada, portanto os indivíduos que compõem esse grupo (1928, 1930, 1932, 1941 e 1942) são, em princípio, aqueles destinados à seleção, tendo em vista toda a trajetória da curva de crescimento. Vale ressaltar que a estratégia usada permitiu selecionar indivíduos geneticamente superiores, levando em consideração o processo de crescimento como um todo, e não

aqueles superiores apenas em relação a pesos em tempos específicos da curva. Essa estratégia também foi usada com sucesso por Silva (2010) em estudos que envolviam curvas de crescimento de gado Nelore e por Pong-Wong e Hadjipavlou (2010), em estudos de simulação com seleção genômica para curvas de crescimento.

Para identificação dos marcadores com maiores efeitos em módulo, utilizou-se o procedimento representado pela Figura 4 do item 3.1.2. Assim, foi obtido o quantil 95% da distribuição empírica dos módulos dos efeitos estimados dos SNPs para análises que incluíam, respectivamente, estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$ (Figura 6).

Sob esse enfoque, em cada análise foram encontrados 23 marcadores acima do quantil 95%, ou seja, 23 marcadores com efeitos mais expressivos em nível genômico. A Tabela 4 fornece as posições reais assumidas na simulação para os seis marcadores de maiores efeitos e as posições dos seis marcadores de efeitos mais expressivos que mais se aproximaram das reais. A grande correspondência entre as posições em questão indica a eficiência do método LASSO em encontrar marcadores de grande efeito. Embora não apresentadas, as posições dos outros 17 SNPs encontradas também foram próximas das simuladas, confirmando a presença de marcadores com efeitos mais expressivos nessas regiões genômicas.

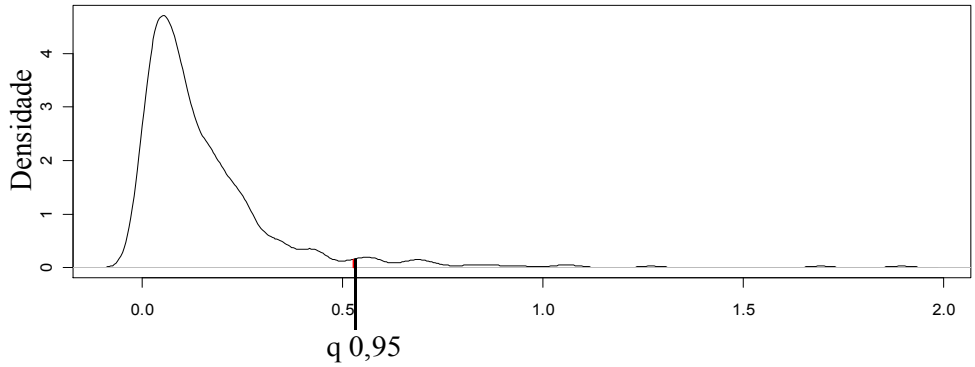
De modo geral, os marcadores de maiores efeitos identificados pelas posições encontradas (Tabela 4) podem ser usados diretamente na localização de QTLs. Esse procedimento pode ser realizado, pois a Seleção Genômica Ampla (GWS), idealizada por Meuwissen *et al*. (2001), preconiza que a quantidade de marcadores é densa o suficiente para que eles estejam em desequilíbrio de ligação direto com o QTL.

4.2 Dados reais

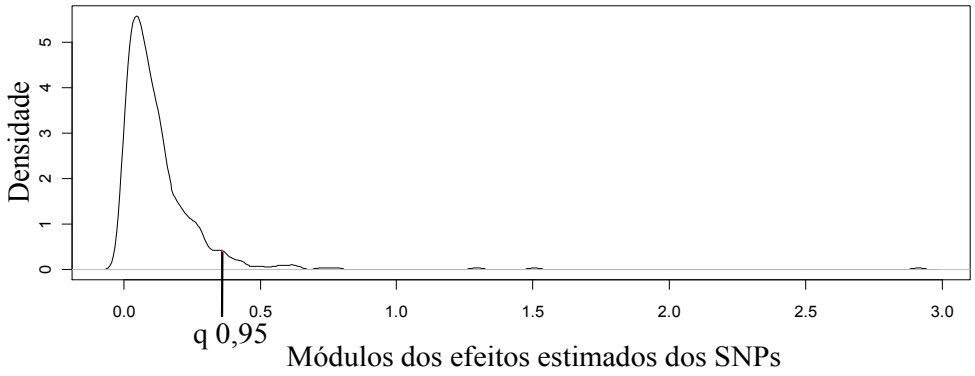
Dos 345 suínos submetidos à seleção, 80 foram eliminados pelo fato de o modelo de crescimento Logístico não ter atingido a convergência ou ter assumido valores irreais para as estimativas dos parâmetros. Para os 265 animais restantes, obteve-se um R^2 médio de aproximadamente 0,99.

Como descrito no item 3.2.2, a correção para os fenótipos (parâmetros estimados do modelo Logístico) foi realizada por meio de duas formas distintas. Na primeira, corrigiram-se efeitos de sexo, lote e presença do gene halotano. Na

(a)



(b)



(c)

Figura 6 – Distribuições empíricas dos módulos dos efeitos estimados dos SNPs (a, b, c) para análises envolvendo, respectivamente, estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Em cada distribuição q 0,95 indica o valor do quantil 95%.

segunda, apenas efeitos de sexo e lote foram corrigidos, sendo o gene halotano considerado como um marcador adicional. Os resultados obtidos pela utilização desses dois procedimentos estão, respectivamente nos itens 4.2.1 e 4.2.2.

Tabela 4 – Posições simuladas e encontradas de marcadores de maiores efeitos para os parâmetros analisados

Parâmetros					
$\emptyset_1$		$\emptyset_2$		$\emptyset_3$	
Posição Simulada	**Posição Encontrada**	**Posição Simulada**	**Posição Encontrada**	**Posição Simulada**	**Posição Encontrada**
0,4245	0,4447	0,5425	0,5309	0,8765	0,9137
1,0455	1,0356	1,3302	1,3445	1,4889	1,4849
1,8864	1,8831	2,0686	2,0515	2,2622	2,2202
2,8984	2,8827	2,5609	2,5826	3,0962	3,0663
3,6979	3,7004	3,3652	3,3582	3,8639	3,8537
4,7719	4,7396	4,5971	4,6264	4,3148	4,2936

4.2.1 Fenótipos corrigidos para efeitos de sexo, lote e presença do gene halotano

De acordo com os resultados apresentados na Tabela 5, constata-se que as capacidades preditivas (estimativas dos coeficientes de correlação $r_{y,G\hat{B}V}$) obtidas pelos dois métodos, LASSO Bayesiano e RR-BLUP/GWS, foram elevadas para os três parâmetros analisados. Isso significa que os dois métodos empregados foram capazes de predizer valores genéticos genômicos de forma acurada. Porém, percebe-se que as estimativas dos coeficientes de regressão linear entre valores observados e preditos ($b_{y,G\hat{B}V}$) foram superiores a 1 nas análises que incluem $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Por sua vez, isso indica que o estimador dos efeitos de marcadores apresenta viés, já que a constatação de estimadores não viesados é caracterizada pelo valor do coeficiente de regressão linear igual a unidade.

Pela análise da Tabela 5, nota-se uma semelhança entre os resultados obtidos pelos métodos LASSO Bayesiano e RR-BLUP/GWS. Em geral, esse fato ocorreu porque nessas análises não foram considerados genes de grande efeito (halotano não foi considerado como marcador), verificando-se a condição que todos os locos explicam iguais quantidades da variação genética, assumida pelo método RR-BLUP/GWS.

A herdabilidade calculada, dada pela fórmula $h^2 = 1/\{1 + [1/(2n/\lambda^2)]\}$, em que n é o número de locos marcadores corrigidos para suas frequências alélicas e λ é o parâmetro de regularização do método LASSO Bayesiano, foi de 0,55, 0,68 e

Tabela 5 – Estimativas dos coeficientes de correlação (r) e regressão linear (b) envolvendo valores fenotípicos (y), valores genéticos genômicos preditos ($G\hat{B}V$) e valores genéticos genômicos verdadeiros (GBV) nos métodos e parâmetros analisados. N é o número de indivíduos e M é o número de marcadores considerados

Parâmetro	Método	N	M	$r_{y,G\hat{B}V}$	$b_{y,G\hat{B}V}$	$r_{GBV,G\hat{B}V}$
	RR-BLUP/GWS-VC*	264	237	0,55	1,00	0,74
$\emptyset_1$	LASSO Bayesiano	265	237	0,80	1,34	-
	RR-BLUP/GWS	265	237	0,80	1,37	-
	RR-BLUP/GWS-VC*	264	237	0,44	1,00	0,63
$\emptyset_2$	LASSO Bayesiano	265	237	0,75	1,51	-
	RR-BLUP/GWS	265	237	0,76	1,59	-
	RR-BLUP/GWS-VC*	264	237	0,54	1,00	0,70
$\emptyset_3$	LASSO Bayesiano	265	237	0,82	1,35	-
	RR-BLUP/GWS	265	237	0,82	1,39	-

* Validação cruzada.

0,61, respectivamente, para os parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Este resultado indica que esses parâmetros podem ser considerados como importantes fenótipos para estudos genéticos que visam selecionar indivíduos de acordo com suas curvas de crescimento. Além disso, as acurácias ($r_{GBV,G\hat{B}V}$) obtidas pelo método RR-BLUP/GWS para os parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$ na validação cruzada foram, respectivamente, 0,74, 0,63 e 0,70, que podem ser consideradas de moderadas a altas (Tabela 5).

A Figura 7 contempla curvas de crescimento genéticas (itens 3.1.2 e 4.1) referentes ao ganho de peso (em kg) no período de 0 a 150 dias, para dez grupos provenientes da análise de agrupamento. Diferentemente do resultado para dados simulados, não se tem uma clara distinção entre as curvas construídas a partir das médias dos $G\hat{B}Vs$ de cada grupo (itens 3.1.2 e 4.1). Ainda assim, percebe-se uma pequena superioridade do grupo 9 em relação aos demais, portanto indivíduos que constituem esse grupo (550, 583, 584, 1011 e 1012) podem, em princípio, ser destinados à seleção, tendo em vista toda a trajetória da curva de crescimento.

A Figura 8 mostra as distribuições empíricas dos módulos dos efeitos estimados dos SNPs para análises que incluem o método LASSO Bayesiano as e estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Em cada análise foram encontrados 12 marcadores acima do quantil 95%, ou seja, 12 marcadores com efeitos mais expressivos em nível genômico. Detalhes sobre esses marcadores de maiores efeitos

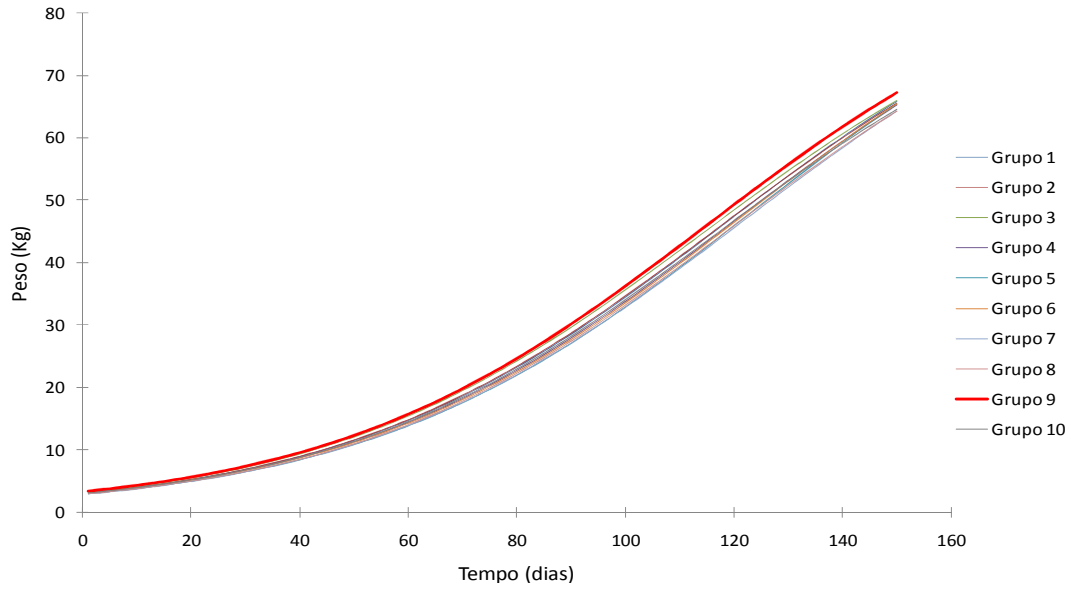


Figura 7 – Curvas de crescimento genéticas referentes ao ganho de peso (em kg) no período de 0 a 150 dias, para dez grupos provenientes da análise de agrupamento.

podem ser visualizados nas Tabelas 6, 7 e 8, para análises que envolvem estimativas de $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$, respectivamente.

Para o parâmetro $\boldsymbol{\emptyset_1}$, peso adulto, (Tabela 6) foram encontrados marcadores de maiores efeitos, principalmente, nos cromossomos SSC1 e SSC4. O marcador ALGA0010089 encontrado no cromossomo 1 na posição 233,3944 cM está de acordo com o trabalho de Quintanilla *et al*. (2002) os quais detectaram QTL significativo para ganho de peso no final do cromossomo 1 em populações de suínos provenientes do cruzamento entre as raças Large White e Meishan. Os marcadores ALGA0021973 e ALGA0021974 encontrados no início do cromossomo 4, respectivamente nas posições 0,2781 e 0,3173 cM concordam com os resultados obtidos por Kim *et al.* (2006) ao utilizarem marcadores microssatélites para detectar QTLs em suínos Yorkshire e com os resultados descritos por Liu *et al.* (2008) em estudos de detecção de QTL em populações de suínos provenientes do cruzamento das raças Duroc x Pietrain. Nestes trabalhos os autores citados encontraram, respectivamente, QTLs mais expressivos nas posições 0 e 4 cM para a característica ganho de peso.

De acordo com resultados das Tabelas 6, 7 e 8, é importante mencionar que o marcador ALGA0010089 figurou-se entre os mais expressivos para os três parâmetros considerados, indicando que a posição de tal marcador contém

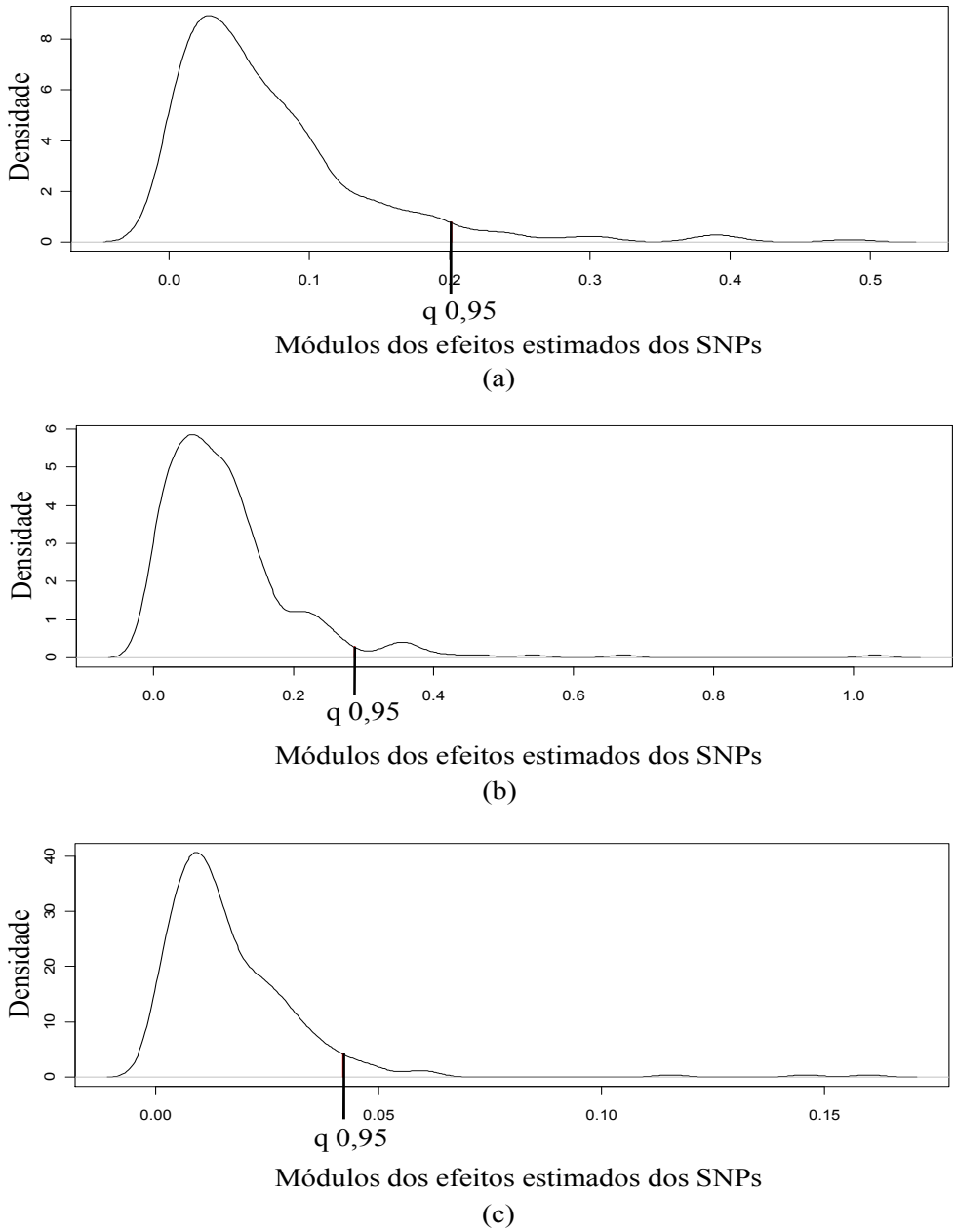


Figura 8 – Distribuições empíricas dos módulos dos efeitos estimados dos SNPs (a, b, c) para análises envolvendo, respectivamente, estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Em cada distribuição q 0,95 indica o valor do quantil 95%.

informações genéticas relacionadas com o crescimento, portanto deve ser melhor explorada em relação a possíveis genes que atuam sobre este processo.

Tabela 6 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_1$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0027861	-0,2076	SSC4	105,0104
ALGA0049546	-0,2218	SSC8	60,0367
ALGA0038559	-0,2337	SSC7	10,3512
ALGA0047440	0,2366	SSC8	15,0429
ALGA0043769	-0,2465	SSC7	100,6632
ALGA0044984	-0,2782	SSC7	120,6290
MARC0051258	-0,3012	SSCX	112,2210
ALGA0025813	0,3086	SSC4	70,2850
ALGA0049219	0,3834	SSC8	55,0076
ALGA0021974	-0,3872	SSC4	0,3173
ALGA0021973	0,4008	SSC4	0,2781
ALGA0010089	0,4850	SSC1	233,3944

Tabela 7 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_2$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0043769	-0,3184	SSC7	100,6632
ALGA0040318	0,3398	SSC7	35,2897
ALGA0025813	0,3483	SSC4	70,2850
ALGA0006721	0,3538	SSC1	142,0169
ALGA0044984	-0,3578	SSC7	120,6290
ALGA0038559	-0,3737	SSC7	10,3512
ALGA0038213	-0,3748	SSC7	5,3398
ALGA0005071	0,4176	SSC1	80,4417
ALGA0049219	0,4673	SSC8	55,0076
ALGA0047440	0,5420	SSC8	15,0429
ALGA0021973	0,6705	SSC4	0,2781
ALGA0010089	1,0308	SSC1	233,3944

Tabela 8 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_3$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0029781	0,0432	SSC4	127,9160
ALGA0044298	-0,0432	SSC7	110,6366
ALGA0093241	0,0461	SSC17	10,0672
ALGA0006721	0,0476	SSC1	142,0169
ALGA0021974	-0,0495	SSC4	0,3173
ALGA0050287	-0,0495	SSC8	66,5612
ALGA0093254	0,0567	SSC17	10,2795
ALGA0037853	-0,0602	SSC7	0,4704
ALGA0029483	0,0616	SSC4	123,2808
MARC0051258	-0,1155	SSCX	112,2210
ALGA0010089	0,1459	SSC1	233,3944
ALGA0047440	0,1601	SSC8	15,0429

4.2.2 Fenótipos corrigidos para efeitos de sexo e lote (Halotano como marcador adicional)

As análises foram realizadas com fenótipos corrigidos apenas para efeitos de sexo e lote. O gene halotano foi considerado um marcador adicional, a fim de identificar sua influência no crescimento dos animais. Esse gene possui grande efeito para ganho de peso e qualidade da carne, sendo amplamente conhecido e divulgado na literatura especializada em melhoramento de suínos (BAND *et al*., 2005).

De acordo com os resultados descritos na Tabela 9, observa-se que as capacidades preditivas (estimativas dos coeficientes de correlação $r_{y,G\hat{B}V}$) obtidas nos métodos LASSO Bayesiano e RR-BLUP/GWS foram elevadas para os três parâmetros analisados. Vale ressaltar que os valores de $r_{y,G\hat{B}V}$ para os parâmetros $\emptyset_1$ e $\emptyset_2$ foram superiores aos valores encontrados nas análises em que o gene halotano não foi considerado como marcador (Tabela 5).

Para o parâmetro $\emptyset_1$ (peso adulto), nota-se que a capacidade preditiva no método LASSO foi superior à encontrada no método RR-BLUP/GWS (Tabela 9). Essa superioridade deve-se a dois fatores. Primeiramente, porque o gene halotano possui grande efeito para ganho de peso e qualidade da carne (BAND *et al*., 2005). Segundo, porque o método LASSO considera variâncias diferentes para efeitos de

Tabela 9 – Estimativas dos coeficientes de correlação (r) e regressão linear (b) envolvendo valores fenotípicos (y), valores genéticos genômicos preditos ($G\hat{B}V$) e valores genéticos genômicos verdadeiros (GBV) nos métodos e parâmetros analisados. N é o número de indivíduos e M é o número de marcadores considerados

Parâmetro	Método	N	M	$r_{y,G\hat{B}V}$	$b_{y,G\hat{B}V}$	$r_{GBV,G\hat{B}V}$
$\emptyset_1$	RR-BLUP/GWS-VC*	264	238	0,80	1,50	0,91
	LASSO Bayesiano	265	238	0,99	1,04	-
	RR-BLUP/GWS	265	238	0,95	1,45	-
$\emptyset_2$	RR-BLUP/GWS-VC*	264	238	0,52	1,01	0,71
	LASSO Bayesiano	265	238	0,79	1,40	-
	RR-BLUP/GWS	265	238	0,79	1,42	-
$\emptyset_3$	RR-BLUP/GWS-VC*	264	238	0,54	1,10	0,76
	LASSO Bayesiano	265	238	0,80	1,43	-
	RR-BLUP/GWS	265	238	0,80	1,49	-

* Validação cruzada.

marcador ao realizar as análises, o que não é contemplado pelo RR-BLUP/GWS, onde se pressupõe a igualdade de variâncias.

Ainda com relação à Tabela 9, as estimativas dos coeficientes de regressão linear entre valores observados e preditos ($b_{y,G\hat{B}V}$) foram superiores a 1 nas análises que envolvem os três parâmetros. No parâmetro $\emptyset_1$, visualiza-se que a estimativa do coeficiente aproximou-se da unidade no método LASSO Bayesiano e foi elevada para os métodos RR-BLUP/GWS-VC e RR-BLUP/GW. Isso indica que no método LASSO o estimador dos efeitos de marcadores é não viesado e consequentemente melhor.

A herdabilidade calculada, assim como no item 4.2.1, foi de 0,77, 0,53 e 0,52, respectivamente para os parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Este resultado indica que esses parâmetros podem ser considerados como importantes fenótipos para estudos genéticos que visam selecionar indivíduos de acordo com suas curvas de crescimento. Particularmente para o parâmetro $\emptyset_1$ (peso adulto), a herdabilidade aumentou de 0,55 para 0,77 quando o gene halotano foi incluído. Além disso, as acurácias obtidas para os parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$ foram, respectivamente, 0,91, 0,71 e 0,76, superando os valores encontrados nas análises em que o gene halotano não foi considerado como marcador (Tabela 5). Essas acurácias são consideradas altas para programas de seleção.

A Figura 9 contempla curvas de crescimento genéticas (itens 3.1.2 e 4.1) referentes ao ganho de peso (em kg) no período de 0 a 150 dias, para dez grupos provenientes da análise de agrupamento. Nota-se também que não existe uma clara distinção entre as curvas construídas a partir das médias dos $G\hat{B}Vs$ de cada grupo (itens 3.1.2 e 4.1). Contudo, percebe-se uma pequena superioridade dos grupos 4 e 7 em relação aos demais. Os indivíduos representados pelos números 409, 412, 461, 494, 558, 591, 781, 887, 889, 890, 891 e 1135 fazem parte do grupo 4 e os indivíduos representados pelos números 500, 517, 519, 589, 714, 754, 939, 941, 958 e 1023 são integrantes do grupo 7. Todos eles podem ser destinados à seleção, tendo em vista a trajetória da curva genética de crescimento.

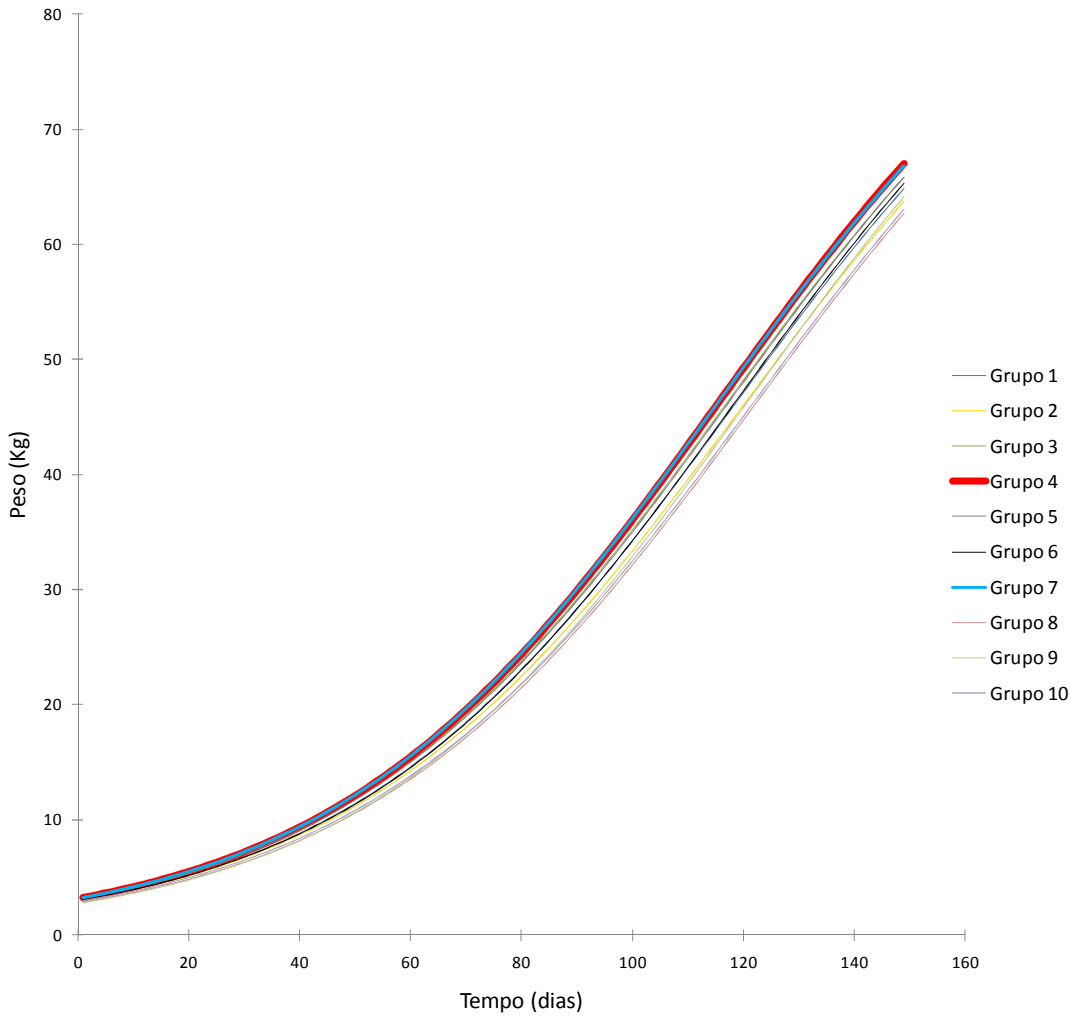


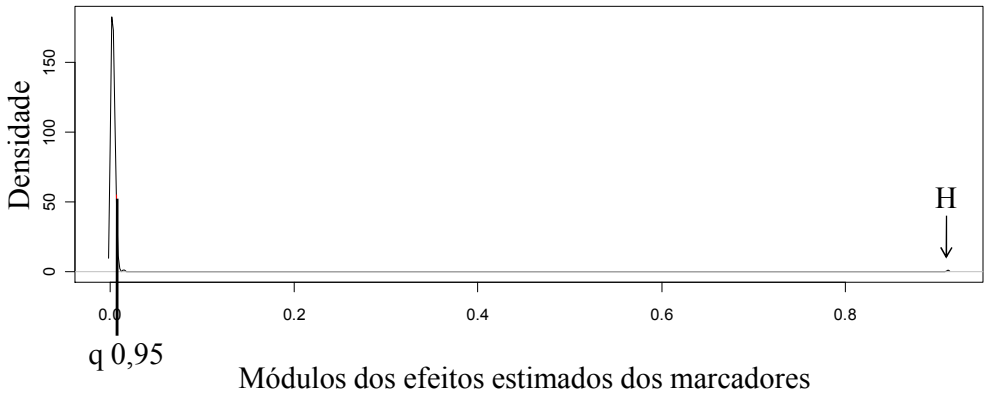
Figura 9 – Curvas de crescimento referentes ao ganho de peso (em Kg) no período de 0 a 150 dias, para dez grupos provenientes da análise de agrupamento.

A Figura 10 ilustra as distribuições empíricas dos módulos dos efeitos estimados dos marcadores para análises que envolvem o método LASSO Bayesiano e as estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. É interessante ressaltar que o marcador halotano pode ser visualizado em (a). Por ter grande efeito no ganho de peso, distingue-se claramente dos outros.

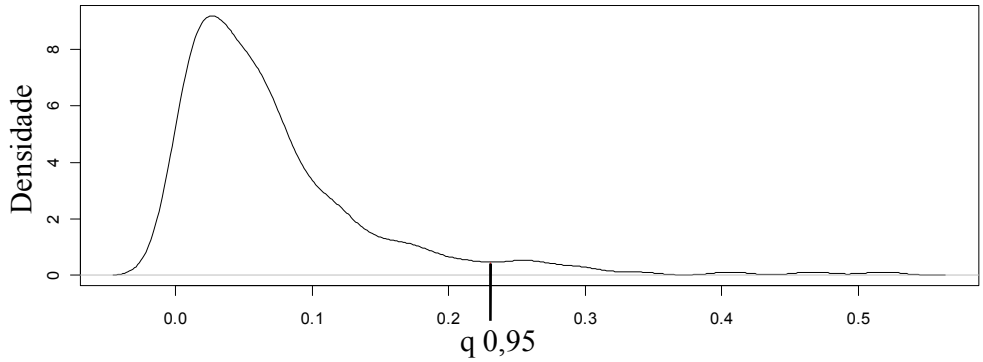
Em cada análise foram encontrados 12 marcadores acima do quantil 95%, ou seja, 12 marcadores com efeitos mais expressivos em nível genômico. Detalhes sobre esses marcadores de maiores efeitos podem ser visualizados nas Tabelas 10, 11 e 12, respectivamente, para análises que envolvem estimativas de $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$.

Em relação ao parâmetro $\boldsymbol{\emptyset_1}$, peso adulto (Tabela 10), foram encontrados marcadores de maiores efeitos principalmente nos cromossomos SSC6, SSC7 e SSC17. É importante realçar a grande magnitude do efeito do gene halotano (-0,9114), destacando-se dos demais. O marcador ALGA0042216 encontrado no cromossomo 7, na posição 60,4304 cM, está de acordo com o trabalho de Yue *et al*. (2003), que detectaram QTL de grande efeito na região central do cromossomo SSC7 para a característica peso ao abate, em suínos provenientes do cruzamento das raças Meishan, Pietrain e European Wild Boar. Rohrer (2000), em estudos que envolviam detecção de QTL para a característica peso ao acabamento em suínos oriundos do cruzamento das raças Meishan x Large White, também encontrou QTL de grande efeito na região central do cromossomo SSC7. O marcador ALGA0095662 encontrado no cromossomo 17 na posição 45,2687 cM concorda com os resultados obtidos por Pierzchala *et al*. (2003), que constataram QTLs para peso ao abate em populações provenientes do acasalamento de suínos das raças Meishan, Pietrain e European Wild Boar.

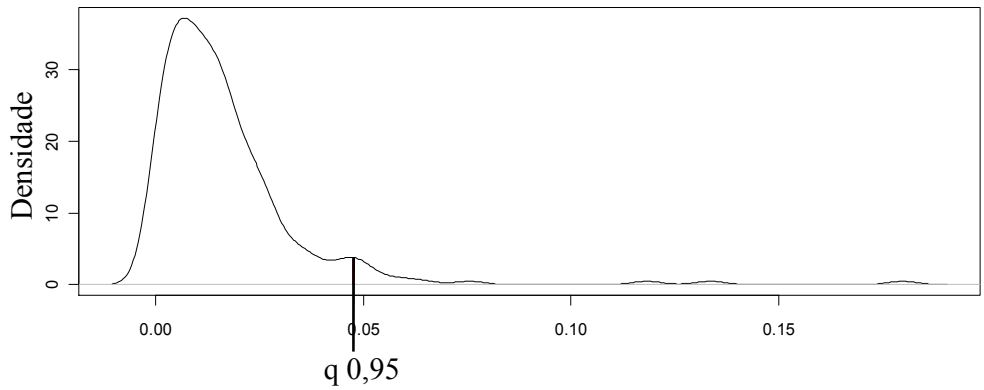
Ao comparar as Tabelas 6, 7 e 8 do item 4.2.1, respectivamente, com as Tabelas 10, 11 e 12 deste item, nota-se que não existem marcadores de maiores efeitos em comum para análises que incluem o parâmetro $\emptyset_1$, ou seja, a incorporação do gene halotano como marcador adicional teve grande influência na detecção de QTLs. Já para o parâmetro $\emptyset_2$, os marcadores ALGA0010089, ALGA0021973, ALGA0025813, ALGA0044984, ALGA0047440 e ALGA0049219 apresentaram efeitos mais expressivos para as duas formas abordadas (4.2.1, 4.2.2). Os resultados obtidos com análises envolvendo $\emptyset_3$ tiveram pouca influência com a incorporação do gene halotano, uma vez que foram encontrados nove marcadores de efeitos mais expressivos em comum nas duas formas de proceder. Em parte, esses resultados



(a)



(b)



(c)

Figura 10 – Distribuições empíricas dos módulos dos efeitos estimados dos SNPs (a, b, c) para análises envolvendo, respectivamente, estimativas dos parâmetros $\emptyset_1$, $\emptyset_2$ e $\emptyset_3$. Em cada distribuição q 0,95 indica o valor do quantil 95%. Em (a), H indica o marcador halotano.

Tabela 10 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_1$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0000022	-0,0065	SSC1	0,2928
ALGA0044519	0,0066	SSC7	115,2341
ALGA0026446	0,0067	SSC4	85,0137
ALGA0111404	-0,0076	SSCX	100,7676
ALGA0042327	0,0078	SSC7	65,5626
ALGA0039607	0,0080	SSC7	26,4277
ALGA0098944	0,0080	SSCX	0,0645
ALGA0024031	-0,0084	SSC4	20,2485
ALGA0099785	0,0092	SSCX	35,1721
ALGA0095662	-0,0092	SSC17	45,2687
ALGA0042216	0,0148	SSC7	60,4304
HALOTANO	-0,9114	SSC6	112,0000

Tabela 11 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_2$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0027861	-0,2437	SSC4	105,0104
ALGA0047440	0,2474	SSC8	15,0429
ALGA0021974	-0,2580	SSC4	0,3173
ALGA0006708	-0,2587	SSC1	141,3861
ALGA0025813	0,2676	SSC4	70,2850
ALGA0029483	0,2808	SSC4	123,2808
ALGA0049546	-0,2976	SSC8	60,0367
ALGA0044984	-0,3004	SSC7	120,6290
MARC0051258	-0,3382	SSCX	112,2210
ALGA0021973	0,4059	SSC4	0,2781
ALGA0049219	0,4668	SSC8	55,0076
ALGA0010089	0,5181	SSC1	233,3944

podem ser explicados pelo fato de o parâmetro $\emptyset_1$ estar diretamente relacionado com o peso, característica influenciada pelo gene halotano, o que não é o caso dos parâmetros $\emptyset_2$ e $\emptyset_3$, que estão diretamente relacionados à idade, como apresentado na Figura 3 do item 3.1.2.

Tabela 12 – Detalhes sobre os marcadores de efeitos mais expressivos encontrados através das análises envolvendo o método LASSO Bayesiano e estimativas do parâmetro $\emptyset_3$

Marcador	Efeito	Cromossomo	Posição (cM)
ALGA0050287	-0,0484	SSC8	66,5612
ALGA0029781	0,0492	SSC4	127,9160
ALGA0026787	0,0492	SSC4	90,3547
ALGA0047444	-0,0494	SSC8	15,1855
ALGA0006721	0,0510	SSC1	142,0169
ALGA0037853	-0,0558	SSC7	0,4704
ALGA0093254	0,0591	SSC17	10,2795
ALGA0006708	-0,0644	SSC1	141,3861
ALGA0029483	0,0756	SSC4	123,2808
MARC0051258	-0,1183	SSCX	112,2210
ALGA0010089	0,1336	SSC1	233,3944
ALGA0047440	0,1798	SSC8	15,0429

5 CONCLUSÕES

Os métodos estatísticos na Seleção Genômica Ampla mostraram-se eficientes no estudo de curvas de crescimento, considerando dados simulados e dados reais de peso-idade de suínos.

A GWS apresentou alta acurácia na seleção para a trajetória das curvas de crescimento e possibilitou a detecção de QTLs (posições nas quais se localizam os marcadores de maiores efeitos) para os parâmetros da curva de crescimento dos indivíduos considerados.

Na ausência de genes de grande efeito, os métodos RR-BLUP/GWS e LASSO Bayesiano produziram resultados semelhantes, entretanto o método LASSO Bayesiano apresentou maior eficiência quando o gene halotano, caracterizado como de grande efeito, foi incluído como marcador nas análises.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BAND, G. D. O.; GUIMARÃES, S. E. F.; LOPES, P. S.; PEIXOTO, J. D. O.; FARIA, D. A.; PIRES, A. V.; FIGUEIREDO, F. C.; NASCIMENTO, C. S.; GOMIDE, L. A. M. Relationship between the Porcine Stress Syndrome gene and carcass and performance traits in F2 pigs resulting from divergent crosses. *Genetics and Molecular Biology*, v. 28, p. 92-96, 2005.

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

FITZHUGH JR., H. A. Analysis of growth curves and strategies for altering their shapes. *Journal of Animal Science*, Champaign, v. 42, n. 4, p. 1036-1051, 1976.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. *Bayesian data analysis*. London: Chapman e Hall, 2004.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Transactions. *Pattern Anal.*, v. 6, p. 721-741, 1984.

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.

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

LIU, G.; KIM, J. J.; JONAS, E.; WIMMERS, K.; PONSUKSILI, S.; MURANI, E.; PHATSARA, C. Combined line-cross and half-sib QTL analysis in Duroc-Pietrain population. *Mammalian genome*: official journal of the International Mammali, v. 19, n. 6, p. 429-438, 2008.

KIM, C. W.; HONG, Y. H.; YUN, S.; LEE, S.; KIM, Y. H.; KIM, M.; CHUNG, K. H.; JUNG W. Y.; KW, E. J. Use of Microsatellite Markers to Detect Quantitative Trait Loci in Yorkshire Pigs. *Journal of Reproduction and Development*, v. 52, n. 2, p. 229-237, 2006.

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.

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

PEIXOTO, J. O.; GUIMARAES, S. E. F.; LOPES, P. S.; SOARES, M. A. M.; PIRES, A. V.; SILVA, M. V.; TORRES, R. A.; SILVA, M. A. E. Associations of leptin gene polymorphisms with production traits in pigs. *Journal of Animal Breeding and Genetics*, v. 123, p. 378-383, 2006.

PEROTTO, D.; CUE, R. I.; LEE, A. J. Comparison of nonlinear functions for describing the growth curve of three genotypes of dairy cattle. *Canadian Journal of Animal Science*, v. 72, n. 4, p. 773-782, 1992.

PIERZCHALA, M.; CIESLAK, D.; REINER G.; BARTENSCHLAGER H.; MOSER G.; GELDERMANN H. Linkage and QTL mapping for Sus scrofa chromosome 17. *Journal of Animal Breeding and Genetics*, v. 120, n. 1, p. 132-137, 2003.

PONG-WONG, R.; HADJIPAVLOU, G. A two-step approach combining the Gompertz growth with genomic selection for longitudinal data. *BMC Proceedings*, v. 4, S4, 2010 (Suppl. 1).

QUINTANILLA, R.; MILAN, D.; BIDANEL, J. P. A further look at quantitative trait loci affecting growth and fatness in across between Meishan and Large White pig populations. *Genetics Selection Evolution*, v. 34, n. 2, p. 193-210, 2002.

R DEVELOPMENT CORE TEAM. *R:* A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. Disponível em: <http://www.R-project.org>.

RESENDE, M. D. V. Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais. Colombo: EMBRAPA Florestas, 2008. 330p.

RESENDE, M. D. V. *Selegen genômica RR-BLUP*. Sistema de seleção genômica ampla (GWS) computadorizada via modelos lineares mistos, 2007 (CD-ROM).

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, v. 1, 2010. 79 p.

ROHRER, G. A. Identification of quantitative trait loci affecting birth characters andaccumulation of backfat and weight in a Meishan-White Composite resourcepopulation. *Journal of Animal Science*, v. 78, n. 10, p. 2547-2553, 2000.

SAS Institute Inc. *Statistical analysis system user's guide*. Version 9, 1. ed. Cary: SAS Institute, USA, 2003.

SILVA, N. A. M. *Seleção de modelos de regressão não lineares e aplicação do algoritmo SAEM na avaliação genética de curvas de crescimento bovinos de nelore*. 2010. 58 f. Tese (Doutorado) – Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2010.

SILVA, N. A. M.; AQUINO, L. H.; SILVA, F. F.; OLIVEIRA, A. I. G. Curvas de crescimento e influência de fatores não genéticos sobre as taxas de crescimento de bovinos da raça Nelore. *Ciência e Agrotecnologia*, Lavras, v. 28, n. 3, p. 647-654, 2004.

SILVEIRA, F. G.; SILVA, F. F.; CARNEIRO, P. L. S.; MALHADO, C. H. M.; MUNIZ, J. A. Análise de agrupamento na seleção de modelos de regressão não-lineares para curvas de crescimento de ovinos cruzados. *Ciência Rural*, v. 41, p. 692-698, 2011.

SOUZA, G. S. *Introdução aos modelos de regressão linear e não linear*. Brasília, DF: EMBRAPA-SPI, 1998. 505 p.

TEDESCHI, L. O.; BOIN, C.; NARDON, R. F.; LEME, P. R. Estudo da curva de crescimento de animais da raça Guzera e seus cruzamentos alimentados a pasto, com e sem suplementação. 1. Análise e seleção das funções não-lineares. *Revista Brasileira de Zootecnia*, v. 29, p. 630-637, 2000.

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

YUE, G.; STRATIL, A.; CEPICA, S.; SCHRÖFFEL Jr. J.; SCHRÖFFELOVA D.; FONTANESI, L.; CAGNAZZO, M.; MOSER, G.; BARTENSCHLAGER, H.; REI, G. Linkage and QTL mapping for Sus scrofa chromosome 7. *Journal of Animal Breeding and Genetics*, v. 120, n. 1, 2003.

7 APÊNDICE

Rotina utilizada no programa R (R Development Core Team, 2010) para implementação do método LASSO Bayesiano.

```
dados=read.table("Arquivo de dados.txt",h=T) #Arquivo contendo dados dos fenótipos (estimativas dos parâmetros) e dados dos marcadores.

y=dados[,2] #Vetor de estimativas do parâmetro ∅1.

X=as.matrix(dados[,-(1:4)]) #Matriz de incidência que relaciona os efeitos dos marcadores aos fenótipos contidos em y (matriz de covariáveis).

require(BLR)#Pacote exigido (BLR-Bayesian Linear Regression).

prior = list( varE = list(S=2, df=3), varBR = list(S=.0009, df=3),lambda=list(type='random',value=1800,shape=.1,rate=2e-5))

nIter=10000 #Amostras salvas para o algoritmo Gibbs Sampler.
burnIn=5000 #Descarte inicial.
thin=100 #Intervalo entre amostras salvas.

fit=BLR(y=y,XL=X,nIter=nIter,burnIn=burnIn,thin=thin, prior=prior)

fit$bL #Estimativas dos efeitos dos marcadores.

fit$mu #Estimativa da média geral.

fit$yHat #Valores genéticos genômicos preditos
         (fit$yHat = ĜBV = fit$mu + X%*%fit$bL).
```