ELCER ALBENIS ZAMORA JEREZ

MODELOS LALDA PARA PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS DE CRESCIMENTO E DE CONVERSÃO ALIMENTAR EM SUÍNOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Zootecnia, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

Z25m
2016

Zamora Jerez, Elcer Albenis, 1975-
Modelos LALDA para predição genômica de características de crescimento e de conversão alimentar em suínos / Elcer Albenis Zamora Jerez. - Viçosa, MG, 2016.
x, 44f. : il. ; 29 cm.

Inclui anexos.
Orientador: Fabyano Fonseca e Silva.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 31-35.

1. Suínos - Genômica. 2. Nutrição animal. 3. Suíno - Crescimento. 4. Suíno - Ganho de peso. 5. Suíno - Desempenho. 6. Suíno - Alimentação e rações. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-graduação em Zootecnia. II. Título.

CDD 22. ed. 636.4

ELCER ALBENIS ZAMORA JEREZ

MODELOS LALDA PARA PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS DE CRESCIMENTO E DE CONVERSÃO ALIMENTAR EM SUÍNOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Zootecnia, para obtenção do título de *Doctor Scientiae*.

APROVADO: 25 de julho de 2016

Marcos Deon Vilela de Resende

José Marcelo Soriano Viana

Rodrigo Oliveira de Lima

Camila Ferreira Azevedo

Fabyano Fonseca e Silva
(Orientador)

A meus pais Teresa e Alvaro
A meu esposo Fernando
A meu filho Francesco Emanuel
A meu anjo Gabrielle Geneve

Dedico

AGRADECIMENTOS

Agradeço primeiramente a Deus pela vida e pela oportunidade de vivenciar toda esta experiência no Brasil.

A Universidade Federal de Viçosa, a Pró-Reitoria de Pesquisa e Pós-Graduação e ao Departamento de Zootecnia pela oportunidade de realização do doutorado.

Ao professor Fabyano Fonseca e Silva pelos ensinamentos e o apoio no momento mais crucial, a professora Simone Elisa Faccione Guimaraes pela oportunidade de aprender no Laboratório de Biotecnologia Animal.

Aos meus colegas de doutorado pelo acompanhamento e paciência em estes anos, em especial a Carol, Margareth, Walmir, Evelyze, Joashy e todos os que com suas explicações me libraram de mais de uma dúvida.

Agradeço infinitamente a minha mãe Gladys Teresa Jerez por seu amor e apoio incondicional, a meu pai Alvaro Zamora Sanchez, que desde o céu me abençoa, a meu esposo Fernando Sanchez Castellanos por tanto amor e carinho e a meu amado filho Francesco Emanuel por encher de luz meu caminho e a meu anjo da guarda Gabrielle Geneve quem guia meus passos.

Aos funcionários do Departamento de Zootecnia: Fernanda, Gabriel, Adilson, Edson, Venâncio e Rosana, Adriano, Elísio, José Lino e Sebastião pela competência.

Aos demais professores, colegas e funcionários do Departamento de Zootecnia que de alguma forma, direta ou indireta, contribuíram para a conclusão deste curso.

SUMÁRIO

LISTA DE FIGURAS ---------- VI

LISTA DE TABELAS ---------- VII

RESUMO ---------- VIII

ABSTRACT ---------- X

1 INTRODUÇÃO ---------- 1

2 REFERENCIAL TEÓRICO ---------- 2

2.1 Melhoramento genético para conversão alimentar e seus componentes ---------- 2

2.2 Incorporação de informações genômicas no melhoramento animal ---------- 3

2.3 Introdução aos modelos LALDA para predição genômica ---------- 4

2.4 Modelos Bayesianos para predição genômica ---------- 7

2.4.1 Introdução à inferência Bayesiana ---------- 7

2.4.2 Métodos de Monte Carlo via Cadeias de Markov (MCMC) ---------- 9

2.4.3 Alfabeto Bayesiano em seleção genômica ---------- 10

3 MATERIAL E MÉTODOS ---------- 12

3.1 Descrição da população e dados fenotípicos ---------- 12

3.2 Extração de DNA e genotipagem para marcadores SNPs ---------- 13

3.3 Modelo estatístico para detecção de QTLs via análise de ligação (LA) ---------- 14

3.4 Modelos LALDA Bayesianos para predição genômica ---------- 16

3.5 Comparação dos modelos LALDA Bayesianos para predição genômica ---------- 18

4 RESULTADOS E DISCUSSÃO ---------- 19

4.1 QTLs para GPD e CR ---------- 19

4.2 QTLs para CA ---------- 20

4.3 Ajuste dos modelos LALDA ---------- 22

4.4 Análise de validação cruzada para os modelos LALDA ---------- 26

4.5 Estimativas de herdabilidade ---------- 28

5 CONCLUSÃO---------- 30

REFERÊNCIAS BIBLIOGRÁFICAS---------- 31

ANEXOS ---------- 36

Input de Programa R para leitura e formatação dos arquivos de dados fenotípicos, genotípicos e de pedigree conforme software QXPAK ---------- 36

Input de Arquivo de parâmetros do software QXPAK para implementar o modelo com efeito aleatório genotípico de QTL via matriz IBD ---------- 39

Input de Programa R para implementação dos modelos LALDA Bayesianos (BRR, BA, BB, BC e BL) – análise de validação cruzada ---------- 40

LISTA DE FIGURAS

Figura 1. Perfis da estatística dos testes de razão de verossimilhanças (TRV) relacionados com a comparação entre o modelo LA (análise de ligação) completo (com efeito aleatório de QTL e poligênico) e o modelo nulo (apenas com o efeito poligênico) para as características consumo de ração diário (CR) e ganho de peso diário (GPD) (a); e conversão alimentar (CA) (b). 22

LISTA DE TABELAS

Tabela 1. Número de observações, média, mediana, desvio padrão (dp), mínimo (min) e máximo (max) para as observações fenotípicas das características consumo de ração diário (CR), ganho de peso diário (GP) e consumo alimentar (CA) calculadas entre 77 e 105 dias de idade. 13

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 suínos (Sus scrofa).....14

Tabela 3. Valores de DIC (Deviance Information Criterion) para modelos LALDA (alfabeto Bayesiano + poligênico aditivo + genotípico de QTL) e LDA (alfabeto Bayesiano + poligênico aditivo) para as características ganho de peso diário (GPD), consumo diário de ração (CR) e conversão alimentar (CA). 24

Tabela 4. Capacidade preditiva (correlação entre preditos e observados) para modelos LALDA (alfabeto Bayesiano + poligênico aditivo + genotípico de QTL) e LDA (alfabeto Bayesiano + poligênico aditivo) para as características ganho de peso diário (GPD), consumo diário de ração (CR) e conversão alimentar (CA).....27

RESUMO

ZAMORA JEREZ, Elcer Albenis, D Sc., Universidade Federal de Viçosa, julho de 2016**. Modelos LALDA para predição genômica de características de crescimento e de conversão alimentar em suínos.** Orientador: Fabyano Fonseca e Silva.

Recentemente, as duas principais fontes de informações para estudos genéticos via marcadores moleculares (LA - linkage analysis, e LDA - linkage disequilibrium analysis) foram combinadas (originando o termo "LALDA") para fins de seleção genômica (SG). Os resultados foram satisfatórios, inclusive superando o modelo LDA (tradicional de SG) em termos de capacidade preditiva em diferentes aplicações a dados simulados e reais. A característica conversão alimentar (CA), e seus componentes (consumo de ração-CR e ganho de peso diário-GPD), são de grande importância econômica para a suinocultura moderna, uma vez que os custos com alimentação representarem a maior parte do custo total de produção. Desta forma, a utilização da SG para tais características se justifica e deve ser motivo de pesquisas na área de Melhoramento Genético Animal. Neste sentido, objetivou-se propor uma metodologia para implementação dos modelos LALDA para predição genômica utilizando softwares livres, bem como aplicar a referida proposta a dados reais de GPD, CR e CA em uma população F2 (Piau x comercial) de suínos. A proposta foi implementada em dois passos distintos. No primeiro, foram identificados efeitos significativos de QTL em posições específicas do genoma para as características GPD, CR e CA via ajuste de modelos que consideraram o efeito aleatório de QTL via matriz IBD (identity by descent) genotípica. No segundo, estas matrizes calculadas nas posições em questão foram utilizadas para inserir o efeito aleatório genotípico de QTL adicionalmente aos feitos aleatórios de marcadores SNPs e poligênico aditivo (baseado em matriz de parentesco tradicional) nos modelos Bayesianos de predição genômica (Bayesian Ridge Regression - BRR, Bayes A - BA, Bayes B - BB, Bayes C – BC e Bayesian LASSO - BL). Foram realizadas análises de qualidade de ajuste e de capacidade preditiva a fim de comprovar a eficiência dos modelos propostos. Em síntese, o modelo LALDA via BA mostrou a melhor qualidade de ajuste via DIC (Deviance Information Criterion) e maior capacidade preditiva quando comparado com os demais modelos LALDA (BRR, BB, BC e BL) para todas características estudadas. Embora de forma discreta, esta superioridade também se verificou ao comparar o modelo em questão com modelos alternativos que não contemplaram o efeito aleatório de QTL (modelos LDA

tradicionais de SG), ou seja, o modelo LALDA proposto mostrou-se eficiente e plausível de ser implementado por meio de softwares livres (QXPAK e R).

ABSTRACT

ZAMORA JEREZ, Elcer Albenis, D Sc., Universidade Federal de Viçosa, July, 2016**. LALDA models for genome prediction for growth and feed conversion traits in pigs.** Advisor: Fabyano Fonseca e Silva.

Recently, the two main sources of information for genetic studies via molecular markers (LA - linkage analysis, and LDA - linkage disequilibrium analysis) has been combined (emerging the term "LALDA") for genomic selection (GS) purposes. The results were satisfactory, even outperforming LDA model (traditional GS) in terms of predictive capacity in different applications to simulated and real data in animal breeding. The trait feed conversion ratio (FCR) and its components (feed intake – FI, weight daily gain – WDG) are very important for the modern pig industry, since the feed costs represent the largest part of the total production costs. Thus, the use of GS for these traits can be justified and represents an interesting research topic into the area of Animal Breeding. In this context, we aimed to propose a LALDA methodology for genomic prediction using free software, as well as to apply the proposed model to real data of WDG, FI and FCR from an F2 pig population (Piau x commercial). LALDA methodology was implemented in two different steps. At the first one, significant QTLs were identified for all traits by using mixed models that considered the QTL random effect via genotypic IBD (identity by descent) matrix. At the second, these matrices calculated at the positions of significant QTLs were used to insert the genotypic QTL random effect additionally to random SNPs markers (traditional GS model) and polygenic additive (based on traditional pedigree relationship matrix) effects in Bayesian models of genomic prediction (Bayesian Ridge Regression - BRR, Bayes a - BA, Bayes B - BB Bayes C - BC and Bayesian LASSO - BL). The goodness of fit and predictive capacity analyses was realized to test the efficiency of the proposed LALDA models. In summary, the LALDA model via BA showed the best fitting through DIC (Deviance Information Criterion) and higher predictive capacity when compared to other LALDA models (BRR, BB, BC and BL) for all traits. Although slightly, the superiority of the LALDA models was verified in relation to alternative models that did not included the genotypic QTL random effect (traditional LDA models for GS). In summary, the proposed LALDA model was efficient and available to be implemented through free software (QXPAK e R).

1 INTRODUÇÃO

O conceito fundamental da seleção genômica (SG) é o de LD (linkage disequilibrium) entre marcador e QTL, de forma que a associação entre eles é uma propriedade da população como um todo. Devido a esta propriedade populacional, é esperado que esta associação seja compartilhada por todos indivíduos e persista por várias gerações (Resende et al., 2012). Diferentemente da análise LDA (linkage disequilibrium analysis), a análise de ligação (linkage analysis - LA) considera apenas o desequilíbrio de ligação que existe dentro de famílias ou cruzamentos específicos, o qual é eliminado por recombinações após algumas poucas gerações. Assim, espera-se que a associação entre marcadores e QTLs permaneça ativa apenas dentro de famílias, e por poucas gerações (Resende et al., 2012).

Wientjes et al. (2013) relataram que as acurácias da SG também dependem de informações sobre LD decorrentes de estruturas familiares recentes, indicando que a LA também pode contribuir para o sucesso da SG. Além disso, a LA pode apontar para regiões genômicas que são herdadas por um ancestral comum ao se utilizar o conceito de matrizes IBDs (identity by descent), fazendo com que tais regiões sejam melhores candidatas para a identificação de mutações causais relacionadas com o fenótipo de interesse (Bercovici et al., 2010).

Recentemente, estas duas fontes de informações (LA e LDA) foram combinadas (surgindo o termo "LALDA") para fins de seleção genômica (Boichard et al., 2012, Luan et al., 2012 e Wientjes et al. 2013) e apresentou resultados satisfatórios, inclusive superando o modelo LDA (tradicional de SG) em termos de capacidade preditiva em diferentes aplicações a dados simulados e reais de bovinos leiteiro.

Dada a importância econômica da característica conversão alimentar (CA) e de seus componentes (consumo de ração-CR e ganho de peso diário-GPD) para a suinocultura moderna, visto que os custos com alimentação representarem a maior parte do custo total de produção, a utilização da SG para tais características se justifica e deve ser motivo de pesquisas na área de Melhoramento Genético Animal.

Diante do exposto, objetivou-se propor uma metodologia para implementação dos modelos LALDA para predição genômica utilizando softwares livres, bem como aplicar a referida proposta a dados reais de GPD, CR e CA de uma população experimental de suínos.

2 REFERENCIAL TEÓRICO

2.1 Melhoramento genético para conversão alimentar e seus componentes

A conversão alimentar (CA), definida como a razão entre o consumo de ração (CR) e o ganho de peso (GP), é a medida de eficiência alimentar mais utilizada na produção de suínos para o abate (Losinger, 2000). Geralmente a CA é mensurada em termos de quilograma de matéria seca ingerida por quilograma de ganho de peso. Estudos de CA e de seus componentes (CR e GP) são impulsionados pelo fato dos custos com alimentação representarem a maior parte do custo total de produção, de forma que pequenos incrementos na CA podem ter uma grande influência na rentabilidade final de um sistema de produção.

Diferentes fatores impactam a CA em suínos, dentre os quais se destacam os genéticos. Estes podem afetar a CA de forma direta (sendo CA o próprio fenótipo avaliado) ou indireta (sendo CR e o GP os fenótipos avaliados). De forma geral, os suínos têm demonstrado alta capacidade de responder à seleção para CA, uma vez que as estimativas de herdabilidade para ganho de peso diário têm variado de 0,13 a 0,40; e para conversão alimentar de 0,19 a 0,42 (Torres Filho, 2001). Adicionalmente, a correlação genética entre ganho de peso diário e conversão alimentar é negativa (Roso et al., 1995, Torres Filho, 2001), o que favorece a seleção, pois o melhoramento genético de suínos visa menores conversões alimentares e maiores ganhos de peso.

Embora a herdabilidade para CA e seus componentes sejam relativamente altas, a mensuração desta característica requer medidas individuais de consumo de alimento, o que implica em manter os animais em baias individuais de forma que o consumo seja mensurado manualmente (alto custo com mão-de-obra) ou por meio de sistemas automáticos sofisticados (alto custo de aquisição e manutenção). Em síntese, por ser considerada uma característica de alto custo e difícil mensuração, o número de observações fenotípicas para CA é muito menor em comparação a outras características, tais como pesos em idades específicas e ganhos de pesos entre estas idades.

Uma opção para tratar este problema é buscar identificação ao nível de DNA dos locos envolvidos na determinação da CA e incorporar estas informações às metodologias tradicionais de melhoramento. O aumento no número de marcadores de DNA, impulsionados pelo desenvolvimento de métodos de genotipagem automatizados e ferramentas sofisticadas de bioinformática, tornaram esta incorporação possível e

executável. Atualmente a maioria dos programas de melhoramentos genéticos de suínos utilizam tais informações (amplamente denominadas de informações genômicas).

Existem diferentes propostas para explorar informações genômicas em programas de melhoramento animal. Em geral, estas se diferenciam em duas grande classes segundo o conceito de genética quantitaiva considerado, análise de ligação ou de desequilíbrio de ligação. Porém, recentemente foram propostos modelos combinando ambos os conceitos simultaneamente a fim de explorar de forma mais efetiva toda informação genômica disponível visando aumentar o ganho de seleção principalmente para características dispendiosas e de difícil mensuração, tal como a CA.

2.2 Incorporação de informações genômicas no melhoramento animal

Com o advento de painéis de marcadores SNP (single nucleotide polymorphism), os quais garantem alta cobertura do genoma, e com um decréscimo significativo nos custos de genotipagens mediante estes painéis, a seleção genômica (Genomic Selection - GS) tornou-se popular e acessível na área de melhoramento genético animal. A eficácia da SG foi comprovada para a maioria dos animais de produção, fazendo com que esta tivesse alta aceitação por parte da comunidade científica e da indústria agropecuária (Habier, 2007).

O princípio básico da SG é que por meio da utilização de um painel de marcadores de considerável densidade, pode-se assumir que cada locus de característica quantitativa (QTL) esteja em desequilíbrio de ligação (linkage disequilibrium analysis - LDA) com marcadores em suas proximidades, fazendo com que os valores genéticos, neste caso denominados de GEBVs (genomic estimated breeding values), sejam preditos com maior acurácia em relação ao método tradicional baseado na matriz de parentesco observada. Portanto, o conceito fundamental por trás da SG é o de LD entre marcador e QTL, de forma que a associação entre eles passa a ser uma propriedade da população como um todo. Devido a esta propriedade populacional, é esperado que esta associação seja compartilhada por todos indivíduos e persista por várias gerações (Resende et al., 2012).

Diferentemente da análise LD, a análise de ligação (linkage analysis - LA) considera apenas o desequilíbrio de ligação que existe dentro de famílias ou cruzamentos específicos, o qual é quebrado por recombinações após algumas poucas gerações. Assim, espera-se que a associação entre marcadores e QTLs permaneça ativa apenas dentro de famílias, e por poucas gerações (Resende et al., 2012). Habier et al. (2007) e Wientjes et

al. (2013) relataram que as acurácias da SG também dependem de informações sobre LD decorrentes de estruturas familiares recentes, indicando que a LA também pode contribuir para o sucesso da GS. Além disso, a LA pode apontar para regiões genômicas que são herdadas por um ancestral comum ao se utilizar o conceito de matrizes IBDs (identical by descent), fazendo com que tais regiões sejam melhores candidatas para a identificação de mutações causais relacionadas com o fenótipo de interesse (Bercovici et al., 2010).

Recentemente (i.e. nos últimos 5 anos), com o objetivo de combinar estas duas fontes de informação (LA e LDA) para o sucesso da detecção de QTLs, algumas pesquisas foram realizadas e evidenciaram a relevância da metodologia LD (Hernández-Sánchez et al., 2009; Bercovici et al., 2010; Pikkuhookana e Sillanpaa, 2014). Tais metodologias mostraram-se satisfatórias na área de melhoramento genético animal, como demonstrado nos trabalhos de Olsen et al. (2009), Brand et al. (2009), Baes et al. (2010) e Sun et al. (2014), dentre outros.

Diferentemente destes trabalhos que visaram a detecção pontual de QTLs, Dassonneville et al. (2012), Boichard et al. (2012) e Luan et al. (2012) utilizaram um modelo contendo valores genéticos genômicos (contemplando a informação de LD) e de QTLs (contemplando a informação de LA) para predição do mérito genético. Dependendo das populações utilizadas, principalmente para aquelas derivadas de cruzamentos específicos (por exemplo F1, F2 e compostos), o modelo LDLA mostrou-se satisfatório e superou o modelo LD (tradicional de GS) em termos de habilidade de predição. Tal sucesso pode ser evidenciado no trabalho de Boichard et al. (2012), no qual é relatado que a avaliação genética oficial de gado de leite na França utiliza o modelo LALDA para predições genômicas do mérito genético dos animais.

2.3 Introdução aos modelos LALDA para predição genômica

Utilizando-se notação matricial, o modelo uni-característico considerando apenas um QTL proposto por Fernando e Grossman (1989) é utilizado como referência para implementação da metodologia LALDA. Este pode ser descrito da seguinte forma:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{ZT\upsilon} + \mathbf{e} \qquad (1)$$

em que: **y** é o vetor de medidas fenotípicas de n indivíduos, **X** e **Z** são, respectivamente, as matrizes de incidência dos vetores de efeitos fixos (**β**) e aleatórios genéticos de indivíduos (**u**), **υ** é o vetor de efeitos aleatórios alélicos de QTL para cada individuo,

$\boldsymbol{\upsilon} = [\upsilon_1^p, \upsilon_1^m, ..., \upsilon_i^p, \upsilon_i^m, ..., \upsilon_n^p, \upsilon_n^m]'$, cuja matriz de incidência é **ZT**, sendo $\mathbf{T} = \mathbf{I}_n \otimes [1\ 1]$, e **e** representa o vetor de efeitos residuais. Na presente notação, os termos p e m em $\boldsymbol{\upsilon}$ indicam, respectivamente, a origem paterna e materna dos alelos. Assumindo-se que os efeitos aleatórios (**u**, $\boldsymbol{\upsilon}$ e **e**) são independentes, e que cada um se distribui segundo uma normal, tem-se: $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{A})$, sendo **A** a matriz de parentesco entre os indivíduos; $\boldsymbol{\upsilon} \sim N(\mathbf{0}, \sigma_\upsilon^2 \mathbf{G}_\upsilon)$, sendo $\mathbf{G}_\upsilon$ a matriz IBD alélica, e $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, sendo $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ a matriz de (co)variância residual.

A proposta do modelo LDLA apresentada por Luan et al. (2012) é baseada no modelo acima, porém a matriz de parentesco tradicional (**A**) é substituída pela matriz de parentesco genômico (**G**) proposta por Van Raden (2008), a qual é dada por: $\mathbf{G} = \mathbf{MM}' / 2 \sum_{i=1}^{I} q_i(1-q_i)$, sendo **M** a matriz de genótipos (N linhas e p colunas, em que N é o número de animais genotipados e I é o número de marcadores) e q_i a menor frequência alélica de cada marcador i. Assim, os efeitos genéticos aleatórios podem ser denominados de valores genéticos, de tal forma que $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{G})$.

O modelo apresentado em (1) precisa ser ajustado separadamente para cada marcador, uma vez que para cada um deles tem-se uma matriz $\mathbf{G}_\upsilon$. Para cada um destes marcadores, deve-se comparar este modelo com o modelo sem efeito de QTL, ou seja, $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$. Esta comparação geralmente é realizada mediante teste de razão de verossimilhanças (TRV), sendo o modelo com efeito de QTL considerado o modelo completo e o modelo sem este efeito o modelo nulo. Dessa forma, caso se verifique efeito significativo de QTL para vários marcadores (q=1,2,..,Q), o modelo em (1) pode ser generalizado para uma estrutura de múltiplos QTLs como apresentado a seguir:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Z}_1\mathbf{T}\boldsymbol{\upsilon}_1 + \mathbf{Z}_2\mathbf{T}\boldsymbol{\upsilon}_2 + ... + \mathbf{Z}_Q\mathbf{T}\boldsymbol{\upsilon}_Q + \mathbf{e}.$$

A grande vantagem do modelo LALDA em relação ao modelo tradicional de predição genômica (GBLUP) é a utilização das estimativas de $\boldsymbol{\upsilon}$ para predição de mérito genético de animais jovens e/ou que ainda não tiveram seus fenótipos coletados. Esta predição torna-se especialmente interessante quando se trabalha com características de carcaça, que geralmente são medidas após o abate dos animais, e características longitudinais de crescimento, as quais necessitam de avaliações durante toda a vida do animal, do nascimento até o abate. Dessa forma, segundo Luan et al. (2012) o valor genético genômico de um indivíduo será dado pelo seu próprio valor genômico u

(proveniente da análise LD) mais os efeitos alélicos de cada um dos QTLs de dos parentais (proveniente da análise LA), isto é: GEBV = u + $\upsilon_1^p + \upsilon_1^m + \upsilon_2^p + \upsilon_2^m + ... + \upsilon_Q^p + \upsilon_Q^m$, sendo υ_q^p e υ_q^m os efeitos alélicos paternos (υ^p) e maternos (υ^m) de cada QTL q.

Segundo Boichard et al. (2012), é possível obter um aumento significativo na acurácia de predição computada pela correlação entre GEBV e o verdadeiro mérito genético dos animais principalmente em população oriundas de cruzamentos específicos, tais como populações F1 e F2 de gado de leite utilizadas na França, linhagens comerciais de frangos e suínos e compostos raciais (cruzamentos industriais) geralmente utilizados em gado de corte.

Nagamine (2005) relata que a inversão da matriz $\mathbf{G}_\upsilon$ requerida no ajuste (expressão 2) do modelo (1) é bastante complexa, porque esta geralmente apresenta grande dimensão (2n x 2n) e muitas vezes, dependendo da posição no cromossomo, esta pode aproximar-se da condição de singularidade. Além disso, como (1) é ajustado separadamente para cada marcador, cada um desses ajustes necessita-se da inversão de uma matriz $\mathbf{G}_\upsilon$ diferente. Dessa forma, o autor demonstra ser possível a utilização de uma matriz IBD genotípica **Q** cuja dimensão é n x n, em vez da IBD alélica $\mathbf{G}_\upsilon$ de dimensão 2n x 2n. Neste enfoque é assumido que $\mathbf{Q} = (1/2)\mathbf{T}\mathbf{G}_\upsilon\mathbf{T}'$, em que $\mathbf{T} = \mathbf{I} \otimes [1\ 1]$. A dimensão desta matriz é quatro vezes menor que a de $\mathbf{G}_\upsilon$, e, além disso, $\mathbf{Q}^{-1}$ pode ser obtida por meio de algoritmos já empregados para a inversão da matriz de parentesco **A.**

A generalização do modelo (1) para a utilização de **Q** é dada por:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Zw} + \mathbf{e} \qquad (2)$$

em que: **y** é o vetor de medidas fenotípicas de n indivíduos, **X** é matriz de incidência dos vetor de efeitos fixos (**β**), **Z** a matriz de incidência dos valores genéticos genômicos (**u**) e genotípico de QTL (**w**), sendo $\mathbf{w} = [w_1, w_2, ..., w_n]'$, em que $w_i = \upsilon_i^p + \upsilon_i^m$, e **e** representa o vetor de efeitos residuais. Diante da suposição de independência e normalidade para os efeitos aleatórios (**u**, **w** e **e**) tem-se: $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{G})$, sendo **G** a matriz de parentesco genômico entre os indivíduos; $\mathbf{w} \sim N(\mathbf{0}, \sigma_w^2\mathbf{Q})$, sendo **Q** a matriz IBD genotípica, e $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, sendo $\mathbf{R} = \sigma_e^2\mathbf{I}_n$ a matriz de (co)variância residual.

Analogamente ao modelo apresentado em (1), o modelo em (2) também necessita ser ajustado separadamente para marcador, sendo que para cada um deles tem-se uma

matriz **Q** diferente. Assim, tamAbém se deve proceder a comparação via TRV com o modelo sem o efeito aleatório genotípico de QTL.

O sistema de equações associado ao modelo (2) é dado por:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}(\sigma_e^2/\sigma_u^2) & \mathbf{Z}'\mathbf{X} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z}+\mathbf{Q}^{-1}(\sigma_e^2/\sigma_w^2) \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{w}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} \quad (3)$$

O sistema em (3) é utilizado sob o enfoque frequentista, de forma que os componentes de variância são estimados geralmente via método REML (Máxima Verossimilhança Restrita). Uma vez obtidas as predições dos vetores **u** e **w**, os mesmos são somados a fim de se obter o vetor de predição do valor genético total, o qual é utilizado para selecionar os indivíduos geneticamente superiores de uma população em relação a uma dada característica.

O modelo descrito em (2) desconsiderando o efeito genotípico de QTL (**w**) é denominado GBLUP (Van Raden, 2008), e é equivalente ao modelo denominado RR-BLUP (Ridge Regression BLUP) proposto por Meuwissen et al. (2001) no primeiro artigo sobre seleção genômica. Embora tais modelos sejam amplamente utilizados na literatura, os mesmos assumem que, a priori, todos os marcadores apresentam a mesma variância.

Modelos mais flexíveis que permitem assumir variância específica para cada marcador e também tratar seleção de variáveis foram apresentadas por Meuwissen et al. (2001), porém tal flexibilidade se deve à abordagem Bayesiana implementada para tais modelos.

2.4 Modelos Bayesianos para predição genômica

2.4.1 Introdução à inferência Bayesiana

Um dos principais objetivos da estatística é realizar inferência sobre os parâmetros de um modelo. Na abordagem Frequentista, os parâmetros desconhecidos são considerados fixos e toda a análise é baseada nas informações contidas na amostra dos dados, ou seja, distribuições inferenciais são assumidas para os estimadores dos parâmetros, e não propriamente para os parâmetros, tal como na Inferências Bayesiana (Paulino et al., 2003).

Ao realizar inferência sobre os parâmetros de um modelo, a informação que se tem do parâmetro de interesse θ é de grande importância na estatística, porém o

verdadeiro valor do parâmetro θ é desconhecido. Segundo Paulino et al. (2003), o que é desconhecido, neste caso o parâmetro θ, é incerto, e toda a incerteza deve ser quantificada em termos de probabilidade. A inferência Bayesiana consiste de uma informação a priori dos dados amostrais e do cálculo da densidade a posteriori dos parâmetros. A informação a priori é dada pela densidade de probabilidade $\mathrm{P}(\theta)$, a qual expressa o conhecimento do pesquisador sobre os parâmetros a serem estimados.

A inferência Bayesiana trata o vetor de parâmetros desconhecidos como quantidades aleatórias e qualquer informação inicial sobre elas pode ser representada por modelos probabilísticos para θ. Assim, tal abordagem permite incorporar algum conhecimento sobre esses parâmetros antes que os dados tenham sido coletados, atribuindo assim distribuições de probabilidade. Essas distribuições podem ser obtidas através de análises anteriores, experiência do pesquisador na área em questão ou em revisões de literatura sobre o assunto que se deseja tratar.

Deste modo, para se realizar uma inferência Bayesiana é necessário assumir uma função densidade de probabilidade a priori $\mathrm{P}(\theta)$, que combinada com a função de verossimilhança $\mathrm{L}(y_1,...,y_n|\theta)$, por meio do teorema de Bayes, gera a função densidade de probabilidade a posteriori $\mathrm{P}(\theta|\mathrm{Y_n})$.

$$\mathrm{P}(\theta|\mathrm{Y_n}) = \frac{\mathrm{L}(\mathrm{Y_n}|\theta)\mathrm{P}(\theta)}{\int \mathrm{L}(\mathrm{Y_n}|\theta)\mathrm{P}(\theta)\mathrm{d}\theta}. \tag{4}$$

Em geral, o denominador em (4) é omitido por não depender de θ, sendo possível reescrever o Teorema de Bayes da seguinte maneira:

$$\mathrm{P}(\theta|\mathrm{Y}) \propto \mathrm{L}(\mathrm{Y_n}|\theta)\mathrm{P}(\theta), \tag{5}$$

ou seja, Posteriori $\propto$ Verossimilhança x Priori, onde $\propto$ representa proporcionalidade.

Toda a inferência sobre o parâmetro θ é realizada por meio da distribuição densidade a posteriori $\mathrm{P}(\theta|\mathrm{Y_n})$. Para se inferir em relação a qualquer elemento de θ, deve-se integrar a distribuição a posteriori conjunta dos parâmetros, $\mathrm{P}(\theta|\mathrm{Y_n})$, em relação a todos os outros parâmetros. Assim, se o interesse do pesquisador se concentra sobre determinado conjunto de θ, por exemplo, θ_1, necessita-se a obtenção da distribuição $\mathrm{P}(\theta_1|\mathrm{Y})$, denominada de distribuição marginal a posteriori, dada por:

$$\mathrm{P}(\theta_1|\mathrm{Y}) = \int_{\theta \neq \theta_1} \mathrm{P}(\theta|Y)d\theta_{\theta \neq \theta_1}. \tag{6}$$

A integração da distribuição conjunta a posteriori para a obtenção das marginais geralmente não é analítica, sendo necessário o uso de algoritmos iterativos especializados como o Gibbs Sampler e o Metropolis-Hastings. Estes algoritmos são denominados de algoritmos MCMC (Markov Chain - Monte Carlo). Para a utilização desses algoritmos, é necessário que se obtenha as distribuições condicionais completas para cada parâmetro, que são obtidas a partir das distribuições a posteriori.

2.4.2 Métodos de Monte Carlo via Cadeias de Markov (MCMC)

Os Métodos de Monte Carlo via Cadeias de Marcov (MCMC) são de grande importância para a estatística Bayesiana. Por meio dos mesmos é possível obter uma amostra das distribuições marginais a posteriori dos parâmetros de interesse por meio de um processo iterativo utilizando as distribuições de cada parâmetro condicionada aos demais parâmetros do modelo, que são denominadas distribuições condicionais completas a posteriori $P\left(\theta_i \middle| \theta_i,...,\theta_{i-1},...,\theta_{i+1},...,\theta_p,Y\right)$. Os valores gerados são considerados amostras aleatórias de uma determinada distribuição de probabilidade, e ao se tratar de um processo iterativo condicionado apenas a última iteração, tem-se o conceito de Cadeia de Markov (Gelfand, 2000). Os principais algoritmos MCMC são o Metropolis-Hastings e o amostrador de Gibbs.

O algoritmo Metropolis-Hastings é utilizado principalmente quando as distribuições condicionais completas a posteriori dos parâmetros não possuem forma fechada, ou seja, não é possível amostrar valores diretamente destas distribuições (i.e. não são caracterizadas de seus distribuições de probabilidade conhecidas). Para a implementação deste algoritmo é necessário considerar uma distribuição conhecida denominada de candidata, de forma que o valor gerado pela mesma é aceito ou não (via critério probabilístico) como sendo um valor da distribuição condicional completa desconhecida. O amostrador de Gibbs é um caso particular do algoritmo Metropolis-Hastings, e gera valores diretamente das distribuições condicionais completas a posteriori pois estas devem ser caracterizadas como distribuições de probabilidade conhecidas.

Como os algoritmos MCMC são processos iterativos, a avaliação da convergência se faz necessária. Dentre os métodos mais utilizados para avaliação da convergência das cadeias MCMC destacam-se o Heidelberger & Welch (1983), Geweke (1992) e Raftery & Lewis (1992).

2.4.3 Alfabeto Bayesiano em seleção genômica

Diante da abundância dos marcadores SNPs ao longo do genoma, os mesmos são utilizados no contexto de seleção genômica para assistir e melhorar diretamente a predição do mérito genético individual para características de interesse.

De forma geral, a utilização desta grande quantidade de informação genômica ainda é um desafio, pois na maioria das vezes não é possível estimar livremente o efeito de cada SNP sobre o fenótipo devido a problemas de multicolinearidade (diferentes marcadores com o mesmo perfil genotípico) e de dimensionalidade (número de marcadores muito maior que o número de animais genotipados, ou seja, o número de parâmetros a serem estimados é muito maior que o número de observações). De acordo com Gianola et al. (2013), tal situação demanda a utilização de métodos estatísticos que considerem a seleção de covariáveis (problema de multicolinearidade) e a regularização do processo de estimação (problema de dimensionalidade).

O modelo original proposto por Meuwissen et al. (2001) para seleção genômica é dado por:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{I} \mathbf{x}_i g_i + \mathbf{e}, \qquad (4)$$

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 efeito marcador SNP (i=1,2,...,p), $\mathbf{x}_i$ é matriz de incidência de cada marcador i, e **e** é o vetor de resíduos do modelo.

Tal como mencionando no item anterior, ao assumir que $g_i \sim N(0,\sigma^2) \; \forall \; i=1,2,...,I$ ou seja, considerar a priori que todos os marcadores dispõem de uma mesma variância, o modelo é denominado de Ridge Regression. Quando ajustado por sob o ponto de vista frequentista, o mesmo é denominado de RR-BLUP, e quando ajustado sob o enfoque Bayesiano de BRR (Bayesian Ridge Regression).

Meuwissen et al. (2001) apresentam também um modelo Bayesiano (denominado Bayes A) no qual assume-se a priori que $g_i \sim N(0,\sigma_i^2)$, ou seja, que cada marcador apresenta uma variância específica. Considerou-se nesta proposta Bayesiana distribuição normal para os efeitos de marcadores, distribuição uniforme para a média geral e distribuição qui-quadrado invertida para a variância residual e para a variância do efeito de cada marcador.

No modelo Bayes A, as distribuições utilizadas na construção da densidade a posteriori conjunta resultam em condicionais completas a posteriori com forma conhecida, o que possibilita a utilização de amostrador de Gibbs para gerar amostras da densidade conjunta a posteriori (e por consequência, das marginais a posteriori de interesse). Ao final do processo MCMC, obtém-se as estimativas dos efeitos de cada marcador e do vetor de valores genéticos genômicos (GEBV) dado por: $\text{GEBV} = \sum_{i=1}^{I} \mathbf{x}_i \hat{g}_i$.

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 que as distribuições das variâncias dos efeitos de marcadores não apresentavam uma massa de densidade no valor 0. Esta característica seria interessante para esta distribuição, uma vez que para algumas características alguns marcadores não apresentam variância genética. O método Bayes B utiliza densidade a priori com massa de densidade em σ^2_{gi} =0. Considera-se que σ^2_{gi} =0 com probabilidade π, enquanto σ^2_{gi} ~ inv χ^2 (υ, S) com probabilidade $1 - \pi$.

Considerando **y** como vetor de observações livre de efeitos da média e efeitos genéticos com exceção do marcador i, a solução para a amostragem de g_i e σ^2_{gi} pode ser feita por meio de $p(\sigma^2_{gi}, g_i \mid \mathbf{y}) \sim p(\sigma^2_{gi} \mid \mathbf{y})\, p(g_i \mid \sigma^2_{gi}, \mathbf{y})$, de modo que a amostragem de σ^2_{gi} não seja função de g_i. Entretanto, os autores não obtiveram $p(\sigma^2_{gi} \mid \mathbf{y})$ 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 obter amostras de $p(\sigma^2_{gi} \mid \mathbf{y})$. A estimativa do vetor GEBV é obtida da mesma forma que mostrado para o modelo Bayes A.

Conforme já apresentado por Meuwissen et al. (2001) a regressão Bayesiana pode ser utilizada nas situações em que se tem mais marcadores (covariáveis) do que observações, uma vez que determinadas distribuições a priori impõem regularização no ajuste do modelo, sob forma de encurtamento dos coeficientes de regressão (shrinkage). Uma forma interessante de executar este encurtamento é por meio da regressão LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996), a qual combina seleção de variáveis via regularização dos coeficientes de regressão. A versão Bayesiana da regressão LASSO (BL) para seleção genômica foi idealizada por de Los Campos et al. (2009). De forma geral, esta consiste na obtenção de estimadores de coeficientes de regressão do modelo (4) que resolvam o seguinte problema de otimização: $\min\{[\mathbf{y} - (\mathbf{1}\mu + \sum_i \mathbf{x}_i g_i)]'[\mathbf{y} - (\mathbf{1}\mu + \sum_i \mathbf{x}_i g_i)] + \lambda \sum_i^p |g_i| \}$, em que $\sum_i^p |g_i|$ é a soma

dos valores absolutos dos coeficientes de regressão e λ é o parâmetro que controla a força da regularização, de forma que quando λ =0 não há regularização.

Na implementação do BL (de Los Campos et al., 2009) impõe-se como distribuição marginal a priori dos p coeficientes de regressão um produto de densidades exponenciais duplas: $p(\mathbf{g}|\lambda) = \prod_{i=1}^{p} \frac{\lambda}{2} \exp(-\lambda|g_i|)$. Por sua vez, os métodos Bayes A e B utilizam distribuição normal: $p(\mathbf{g}|\sigma_{gi}^2) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{gi}^2}} \exp\left(-\frac{g_i}{2\sigma_{gi}^2}\right)$. A estimativa do vetor GEBV via BL é obtida da mesma forma que mostrado para os modelos Bayes A e Bayes B. Devido a denominação de cada método via letra do alfabeto, todos estes métodos foram formalmente definidos na literatura como alfabeto Bayesiano para seleção genômica (Gianola et al., 2009; Habiel et al. 2011; Gianola 2013).

3 MATERIAL E MÉTODOS

3.1 Descrição da população e dados fenotípicos

Os dados utilizados são provenientes da população F2 (Piau x comercial) da granja de Melhoramento de Suínos do Departamento de Zootecnia da Universidade Federal de Viçosa, em Viçosa, MG, Brasil.

Para a formação desta população foram construídas inicialmente duas famílias provenientes do cruzamento de dois varrões da raça nativa brasileira Piau com 18 fêmeas originadas de linhagem desenvolvida na UFV pelo acasalamento de animais de raça comercial, Landrace x Large White, selecionadas para peso e precocidade. Dentre os machos F1 foram selecionados ao acaso 11 varrões, provenientes de diferentes leitegadas que foram cruzados (monta natural) com 54 fêmeas. Estes animais foram acasalados para a produção da geração F2, dividida em cinco diferentes lotes de nascimento.

Ao nascimento, os animais foram identificados individualmente utilizando o sistema de marcação via brincagem nas orelhas, foram também pesados e realizado o corte de dentes e aplicação de antibiótico. Entre o terceiro e quinto dia foi aplicada uma dosagem (3ml) de ferro injetável, para se prevenir anemia. A castração dos machos foi feita por volta do 10º dia de idade; quando também era colocada ração pré-inicial à disposição dos leitões. O desmame foi feito aos 21 dias de idade, quando os animais eram novamente pesados e recebiam mais uma dosagem (3ml) de ferro injetável, e eram então transferidos para a creche, onde ficaram até cerca de 60 dias de idade. Após esta data, foram transferidos para o setor de cria/recria onde ficaram até os 77 dias de idade, quando

iniciaram o teste de conversão alimentar. Este teste foi conduzido em galpão onde os animais ficaram em baias individuais por um período de 28 dias (77 a 105 dias de idade).

As observações das características fenotípicas foram feitas na geração F2 durante cerca de 150 dias (idade aproximada ao abate), quando os animais atingiam cerca de 65 kg de peso vivo.

No presente estudo, consideram-se apenas três características (avaliadas no período já mencionando de 77 aos 105 dias de idade), dadas pela conversão alimentar diária (CA), consumo de ração diário (CR) e ganho de peso diário (GP). A CA foi calculada como sendo CR dividida pelo GP. As estatísticas descritivas para estas características são apresentadas na Tabela 1.

Tabela 1. Número de observações, média, mediana, desvio padrão (dp), mínimo (min) e máximo (max) para as observações fenotípicas das características consumo de ração diário (CR), ganho de peso diário (GP) e consumo alimentar (CA) calculadas entre 77 e 105 dias de idade.

Característica	Unidade	N	Média	Mediana	dp	min	max
CR	Kg/dia	341	1,44	1,46	0,30	0,48	2,34
GP	Kg/dia	341	0,53	0,55	0,13	0,08	0,89
CA	Kg/Kg	341	2,78	2,71	0,59	1,53	5,25

3.2 Extração de DNA e genotipagem para marcadores SNPs

O DNA dos animais parentais, F1 e F2 foi extraído do sangue dos mesmos. As células, independentemente de suas procedências, foram mantidas a 60°C por uma hora em CTAB, sendo posteriormente centrifugadas e deproteinizadas em um banho de clorofórmio seguido de centrifugação, sendo o DNA então precipitado em etanol absoluto, e guardado a -20°C para uso posterior. As soluções de DNA para uso (na concentração aproximada de 25 ηg/μL) foi mantida em geladeira a 4°C. Para as amplificações, este DNA foi submetido à Reação em Cadeia da Polimerase (PCR) no Laboratório de Biotecnologia Animal do DZO/UFV.

Os marcadores SNPs 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 baixa frequência alélica (MAF < 0,05). Após estes procedimentos, a distribuição de SNPs foi a seguinte: 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 através do banco de dados de suínos disponível no sistema Ensembl (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 suínos (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

3.3 Modelo estatístico para detecção de QTLs via análise de ligação (LA)

Para empregar o modelo LALDA proposto no presente trabalho, primeiramente foi ajustado o modelo de detecção de QTL proposto por Fernando e Grossman (1989). Tal modelo foi ajustado separadamente para cada cromossomo e cada característica (CA, CR e GP). O modelo em questão é dado por:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Zw} + \mathbf{e} \qquad (1)$$

em que: **y** é o vetor de medidas fenotípicas de n indivíduos; **X** é matriz de incidência dos vetor de efeitos fixos (**β**) representados por sexo, lote e genótipo do gene halotano; **Z** a matriz de incidência dos efeitos poligênicos (**u**) e genotípico de QTL (**w**), sendo $\mathbf{w} = [w_1, w_2, ..., w_n]'$, e **e** representa o vetor de efeitos residuais. Diante da suposição de independência e normalidade para os efeitos aleatórios (**u**, **w** e **e**) tem-se: sendo **A** é a

matriz de parentesco tradicional entre os indivíduos; $\mathbf{w} \sim N(\mathbf{0},\sigma_w^2\mathbf{Q})$, sendo **Q** a matriz IBD genotípica, e $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$, sendo $\mathbf{R} = \sigma_e^2\mathbf{I}_n$ a matriz de (co)variância residual.

A obtenção da matriz IBD genotípica (**Q**) é dada como se segue. As matrizes $\mathbf{G}_\upsilon$ (IBD alélicas) foram obtidas por meio do software QXPAK.5 (Pérez-Enciso & Misztal, 2011), o qual considera a seguinte metodologia. Em uma população formada pelo cruzamento entre duas raças C e D, um indivíduo i apresenta em um determinado lócus, respectivamente, os alelos paternos e maternos g_i^p e g_i^m, cujos efeitos aditivos são υ_i^p e υ_i^m. A (co)variância entre estes efeitos, considerando o lócus em questão, provenientes de dois indivíduos diferentes, i e q, é definida por:

$$\mathrm{Cov}(\upsilon_i^h,\upsilon_q^{h'}) = \frac{1}{2}\sum_{h=1}^{2}\sum_{h'=1}^{2} P(\upsilon_i^h \equiv \upsilon_q^{h'} | \upsilon_i^h \in C)\sigma_C^2 + \frac{1}{2}\sum_{h=1}^{2}\sum_{h'=1}^{2} P(\upsilon_i^h \equiv \upsilon_q^{h'} | \upsilon_i^h \in D)\sigma_D^2, \qquad (2)$$

em que: $P(\upsilon_i^h \equiv \upsilon_q^{h'} | \upsilon_i^h \in C)$ é a probabilidade dos alelos g_i^h e $g_q^{h'}$ (cujos efeitos aditivos são υ_i^h e $\upsilon_q^{h'}$) serem idênticos por descendência e provenientes da raça C; h é o índice que representa a origem paterna (h=1) e materna (h=2) e σ_C^2 é a variância dos efeitos aditivos para a raça C. A mesma notação é assumida para a raça D.

Para estimar as probabilidades $P(\upsilon_i^h \equiv \upsilon_q^{h'} | \upsilon_i^h \in C)$ e $P(\upsilon_i^h \equiv \upsilon_q^{h'} | \upsilon_i^h \in D)$, o software em questão utiliza um procedimento de amostragem Gibbs, cuja descrição detalhada é apresentada em Pérez-Enciso et al. (2000). Uma vez obtidas tais estimativas, a matriz $\mathbf{G}_\upsilon$ é então confeccionada para cada posição considerada, sendo esta dada por:

$$\mathbf{G}_\upsilon = \begin{pmatrix}
\boxed{\mathrm{Cov}(\upsilon_1^1,\upsilon_1^1)} & \overline{\mathrm{Cov}(\upsilon_1^1,\upsilon_1^2)} & \mathrm{Cov}(\upsilon_1^1,\upsilon_2^1) & \mathrm{Cov}(\upsilon_1^1,\upsilon_2^2) & \cdots & \mathrm{Cov}(\upsilon_1^1,\upsilon_n^1) & \mathrm{Cov}(\upsilon_1^1,\upsilon_n^2) \\
\underline{\mathrm{Cov}(\upsilon_1^2,\upsilon_1^1)} & \underline{\mathrm{Cov}(\upsilon_1^2,\upsilon_1^2)} & \mathrm{Cov}(\upsilon_1^2,\upsilon_2^1) & \mathrm{Cov}(\upsilon_1^2,\upsilon_2^2) & \cdots & \mathrm{Cov}(\upsilon_1^2,\upsilon_n^1) & \mathrm{Cov}(\upsilon_1^2,\upsilon_n^2) \\
\mathrm{Cov}(\upsilon_2^1,\upsilon_1^1) & \mathrm{Cov}(\upsilon_2^1,\upsilon_1^2) & \boxed{\mathrm{Cov}(\upsilon_2^1,\upsilon_2^1)} & \overline{\mathrm{Cov}(\upsilon_2^1,\upsilon_2^2)} & \cdots & \mathrm{Cov}(\upsilon_2^1,\upsilon_n^1) & \mathrm{Cov}(\upsilon_2^1,\upsilon_n^2) \\
\mathrm{Cov}(\upsilon_2^2,\upsilon_1^1) & \mathrm{Cov}(\upsilon_2^2,\upsilon_1^2) & \underline{\mathrm{Cov}(\upsilon_2^2,\upsilon_2^1)} & \underline{\mathrm{Cov}(\upsilon_2^2,\upsilon_2^2)} & \cdots & \mathrm{Cov}(\upsilon_2^2,\upsilon_n^1) & \mathrm{Cov}(\upsilon_2^2,\upsilon_n^2) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\mathrm{Cov}(\upsilon_n^1,\upsilon_1^1) & \mathrm{Cov}(\upsilon_n^1,\upsilon_1^2) & \mathrm{Cov}(\upsilon_n^1,\upsilon_2^1) & \mathrm{Cov}(\upsilon_n^1,\upsilon_2^2) & \cdots & \boxed{\mathrm{Cov}(\upsilon_n^1,\upsilon_n^1)} & \overline{\mathrm{Cov}(\upsilon_n^1,\upsilon_n^2)} \\
\mathrm{Cov}(\upsilon_n^2,\upsilon_1^1) & \mathrm{Cov}(\upsilon_n^2,\upsilon_1^2) & \mathrm{Cov}(\upsilon_n^2,\upsilon_2^1) & \mathrm{Cov}(\upsilon_n^2,\upsilon_2^2) & \cdots & \underline{\mathrm{Cov}(\upsilon_n^2,\upsilon_n^1)} & \underline{\mathrm{Cov}(\upsilon_n^2,\upsilon_n^2)}
\end{pmatrix}$$

O software QXPAK.5 armazena cada matriz $\mathbf{G}_\upsilon$ (uma para cada posição k do cromossomo) em arquivos denominados zran.k, por exemplo para a posição 1 esta matriz é armazenada no arquivo zran.10000, para a posição 2 no arquivo zran.20000, e assim

continuamente até a última posição a ser testada (por default, adotou-se cada posição como sendo 1 cM).

Para se inferir em relação a presença de QTL, o software emprega o teste de razão de verossimilhanças (RV), cuja estatística é dada por: $\log RV = -2\ln(L_0/L_1)$, em que L_0 e L_1 são, respectivamente, os valores do logaritmo da função de verossimilhança para os modelos sem e com efeito genotípico de QTL. A distribuição Qui-quadrado (χ^2) com d graus de liberdade, sendo d a diferença entre o número de parâmetros dos dois modelos, é utilizada para obtenção dos valores limites (threshold) aproximados que definem a significância de uma dada posição testada via correção FDR (false discovery rate) para múltiplos de teste.

Uma vez detectada a presença significativa de QTL para as características em uma dada posição, as matrizes IBD alélicas salvas pelo software nestas posições foram então inseridas no software R e transformadas em matrizes IBD genotípicas ($\mathbf{Q} = (1/2)\mathbf{TG}_{\upsilon}\mathbf{T}'$) (Nagamine, 2005) associados ao efeito aleatório de QTL a ser considerado nos modelos LALDA Bayesianos para predição genômica propostos no presente estudo.

É importante ressaltar que caso um QTL tenha sido declarado como significativo exatamente em uma posição físico de algum marcador considerado, tal marcador tem seu efeito excluído do componente LDA do modelo LALDA, sendo contemplado apenas no componente LA deste mesmo modelo por meio do efeito aleatório de QTL que o mesmo representa. Este procedimento é realizado com o intuito de evitar que um mesmo marcador seja considerado simultaneamente no modelo LALDA no componente LA e no componente LDA, o que poderia ocasionar multicolinearidade e consequentemente partição da informação relacionada ao marcador em questão.

3.4 Modelos LALDA Bayesianos para predição genômica

Para implementação dos modelos LALDA, os modelos tradicionais de predição genômica (alfabeto Bayesiano descritos no item Referencial Teórico) foram adaptados para comportar efeitos genotípicos (**w**) de QTL (sendo **Q** a matriz de covariância associados a eles). Além destes efeitos de QTL, tais modelos também comportaram os efeitos de marcadores SNPs (g_i) e o poligênico aditivo (**u**) assumindo a matriz de parentesco tradicional (**A**). A inclusão deste último efeito se deve ao pequeno número de marcadores no presente estudo (somente 237) de forma que a utilização apenas dos efeitos

de marcadores SNPs não resultaram em valores de acurácia satisfatórios considerando análises prévias.

Diante do exposto, o modelo LALDA geral considerado no presente estudo, considerando apenas um QTL (cujo efeito é dado por w) foi dado por:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{I} \mathbf{x}_i g_i + \mathbf{Zu} + \mathbf{Zw} + \mathbf{e}, \qquad (3)$$

em que: **y** é o vetor de fenótipos pré-corrigidos para efeitos fixos (sexo, lote e genótipo do gene halotano), **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 efeito marcador SNP (i=1,2,...,p), $\mathbf{x}_i$ é matriz de incidência de cada marcador i, e os demais efeitos tal como descrito no modelo em (1).

Os diferentes submodelos associados a (3) foram definidos de acordo com as distribuições a priori para os efeitos de marcadores, ou seja, foram considerados os seguintes modelos do alfabeto Bayesiano: Bayesian Ridege Regression (BRR), Bayes A (BA), Bayes B (BB), Bayes C (BC) e Bayesian LASSO (BL).

As distribuições a priori para efeitos dos marcadores assumidas para os modelos BRR, BA, BB, BC e BL são dadas respectivamente por: $g_i \sim N(0,\sigma^2)\ \forall\ i=1,2,...,I$, $g_i \sim N(0,\sigma_i^2)$, $g_i \sim (1-\gamma_i)N(0,\sigma_{i0}^2=0) + \gamma_i N(0,\sigma_{i1}^2)$, $g_i \sim (1-\gamma_i)N(0,\sigma_0^2=0) + \gamma_i N(0,\sigma_1^2)$ e $g_i \sim N(0,\tau_i^2\sigma_e^2)$. Especificamente para os modelos BB e BC, a probabilidade de gerar a variável indicadora binária γ_i (relacionada com a seleção de variáveis) foi gerada de uma distribuição Beta(α_1, α_2). Especificamente para o modelo BL, a distribuição a priori para o parâmetro τ_i^2 foi assumida como sendo uma Exponencial, $\tau_i^2 \sim \text{Exp}(\lambda^2)$, na qual o parâmetro λ denominado de "penalização" ou "regularização" foi assumido por pertencer a distribuição Gama, tal que $\lambda^2 \sim G(\phi_1,\phi_2)$.

Em termos de práticos, vale ressaltar que a diferença entre os modelos BRR, BA, BB, BC e BL está relacionada com a arquitetura genética que os mesmos representam. O modelo BRR assume a priori que todos os marcadores apresentam a mesma variância; diferentemente o modelo BA assume uma variância para cada marcador. Já o modelo BB, tal como o BA, também assume uma variância para cada marcador, porém exerce adicionalmente uma seleção de variáveis (i.e., assume que alguns marcadores não têm efeitos sobre a característica em questão) realizadas por meio da mistura de distribuições Normais baseadas na proporção de valores 1 e 0 gerados para γ_i. O modelo BC, tal como

BB, também assume esta seleção de variáveis, porém diferentemente do BA e similarmente ao BRR, assume apenas uma única variância para todos os marcadores. O modelo BL, tal como o BB, assume uma variância para cada marcador e também a seleção de variáveis, porém esta seleção não está fundamenta em misturas de distribuições, e sim na utilização de um parâmetro de regularização (λ) que direcionada para zero marcadores com efeitos irrelevantes.

As distribuições a priori para todos os componentes de variância referentes ao modelo (3) foram assumidas como sendo Qui-quadrada invertida escalada, de forma que exceto para os modelos BB e BC, o algoritmo amostrador de Gibbs foi utilizado para gerar amostras das distribuições marginais a posteriori para todos os parâmetros. Para estes modelos mencionados, a geração da variável latente indicadora (γ_i) foi realizada por meio do algoritmo Metropolis-Hastings.

Os modelos BRR, BA, BB, BC e BL contento os efeitos aleatórios poligênico aditivo e de QTL foram ajustados por meio do pacote BGLR do R considerando 100.000 iterações, 30.000 de burn-in e 5 the thin. A análise de convergência, apenas verificada para a cadeia da variância do erro, foi realizada no software boa (Bayesian Output Analysis) do R. Todos os códigos utilizados são apresentados de forma detalhada no Apêndice.

3.5 Comparação dos modelos LALDA Bayesianos para predição genômica

Todos os modelos (BRR, BA, BB, BC e BL) representados em (3) foram comparados considerando o arquivo completo de observações, ou seja, foram avaliados quanto a qualidade de ajuste. Estes modelos também foram comparados com suas versões reduzidas, no qual o efeito genotípico de QTL (**w**) foi removido. Tais comparações foram realizadas por meio do DIC (Deviance Information Creiterion) conforme proposto por Spiegelhalter et al. (2002), no qual $\mathrm{DIC} = \mathrm{D}(\theta) + 2\,\mathrm{p_D}$, em que $\mathrm{D}(\theta)$ é a deviance e $\mathrm{p_D}$ é o número efetivo de parâmetros no modelo.

Adicionalmente, as análises de qualidade de ajuste, análises de capacidade preditiva via validação cruzada também foram realizadas. Neste sentido, os dados originais foram divididos em cinco grupos independentes (five-fold) de dados (quatro grupos contento 68 animais e um grupo contendo 69 animais), de forma que cinco análises diferentes foram realizadas, sendo que em cada uma delas um destes grupos não foi

considerado a fim de representar a população de validação. Ao final da análise de validação cruzada o resultado foi expresso como a correlação entre o vetor de vetor (**y**) de fenótipos pré-corrigidos para efeitos fixos (sexo, lote e genótipo do gene halotano) e o valor genético total ($\hat{\mathbf{y}}$), de forma que $\hat{\mathbf{y}} = \sum_{i=1}^{I} \mathbf{x}_i \hat{g}_i + \mathbf{Z}\hat{\mathbf{u}} + \mathbf{Z}\hat{\mathbf{w}}$ para o modelo completo (LALDA) e para o modelo sem o efeito aleatório de QTL, ou seja, o modelo LDA.

Uma vez identificado o melhor modelo, os componentes de variância considerando o ajuste do mesmo aos dados completos foram então utilizados para estimar a herdabilidade (h^2) de cada característica. Para o modelo LALDA h^2 foi defina como sendo:

$$\hat{h}^2 = \frac{\hat{\sigma}_g^2 + \hat{\sigma}_u^2 + \hat{\sigma}_w^2}{\hat{\sigma}_g^2 + \hat{\sigma}_u^2 + \hat{\sigma}_w^2 + \hat{\sigma}_e^2}, \qquad (4)$$

em que: $\hat{\sigma}_g^2 = \sum_{i=1}^{I} 2\hat{p}_i(1-\hat{p}_i)\hat{\sigma}_i^2$, na qual $\hat{\sigma}_i^2$ é a variância explicada por cada marcador e p_i as estimativas das frequências alélicas destes mesmos marcadores; $\hat{\sigma}_u^2$ e $\hat{\sigma}_w^2$ são os componentes de variância associados aos efeitos poligênico aditivo e de QTL.

4 RESULTADOS E DISCUSSÃO

Nas Figuras 1a (para as características CR e GPD) e 1b (para a característica CA) são apresentados os perfis da estatística dos testes de razão de verossimilhanças (TRV) relacionados com a comparação entre o modelo LA completo (com efeito aleatório de QTL, **w**, e poligênico, **u**) e o modelo nulo (apenas com o efeito poligênico).

4.1 QTLs para GPD e CR

Nota-se que foi reportado um efeito significativo de QTL para cada uma das características avaliadas. Para GPD e CR estes QTLs encontram-se no cromossomo 1, respectivamente nas posições 91 e 101 cM. Para CA, o QTL encontra-se na posição 136 cM do cromossomo 8.

Em relação aos QTLs indicados no cromossomo 1, Paixão et al. (2013) utilizou dados da mesma população F2 (Piau x comercial) considerada no presente estudo, porém considerando o método de regressão fixa proposto por Haley e Knott (1992), e identificaram efeitos significativos de QTL para as características GPD e CR respectivamente nas posições 93 cM e 102 cM do cromossomo 1. Os autores ainda

calcularam o intervalo de confiança para tais efeitos, os quais foram de 85 a 100 cM para GPD e de 95 a 120 cM para CR. Liu et al. (2007) estudaram uma população de animais cruzados Duroc x Pietran e também identificaram efeito significativo de QTL para GPD no cromossomo 1 na região entre 90,6 e 93,2 cM. Jiao et al. (2014) também identificaram uma região cromossômica significativa no SSC 1 (140-166 cM) relacionada com GPD em uma população Duroc.

Além dos QTLs reportados no cromossomo 1, estudos relatam importantes QTLs identificados para CR e GPD em outros cromossomos. Gilbert et al. (2010) ao utilizaram dados de CR obtidos de uma população de retrocruzamento entre Large White e Pietrain identificaram QTLs com efeitos significativos no SSC2 a 2 cM, no SSC6 a 83 cM e no SSC9 a 104 cM. Liu et al. (2007) analisaram dados de GPD em uma polução cruzada (Duroc x Pietran) e reportaram relevantes QTLs no SSC9 a 9.2 cM e no SSC10 a 59.2 e 79.3 cM.

4.2 QTLs para CA

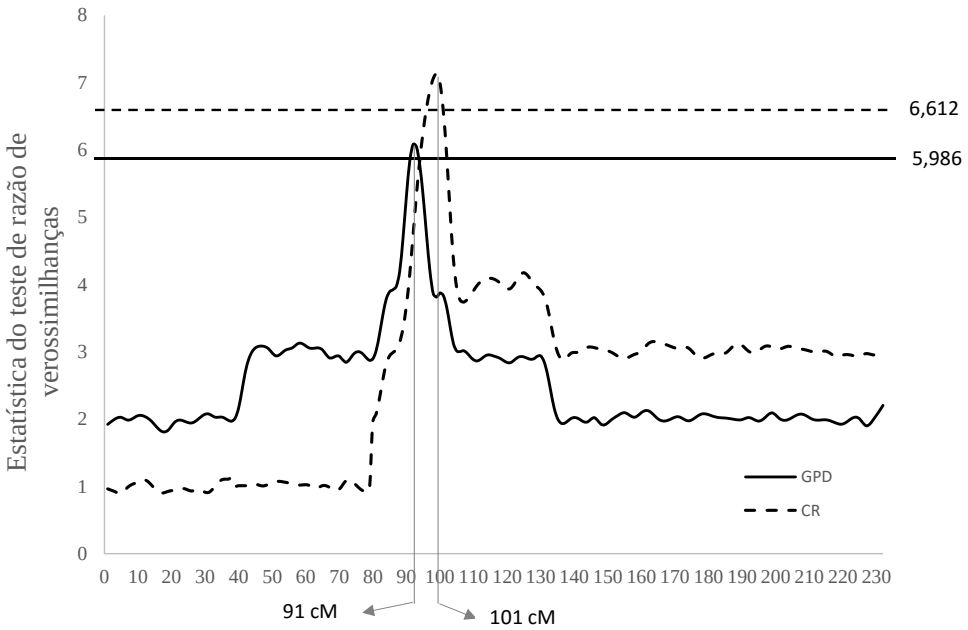
Também utilizando dados da mesma população F2 (Piau x comercial) considerada no presente estudo, Souza (2008) encontrou efeito significativo de QTL via método de Haley e Knott (1992) para a característica CA na região compreendida entre 138 e 148,9 cM do cromossomo 8.

Comparativamente com outras características, dada a dificuldade de mensuração e o alto custo envolvido na fenotipagem para CA, são poucos os estudos objetivando detecção de QTLs para esta característica. Similarmente aos resultados apresentados no presente trabalho, Beeckmann et al. (2003) também relatou significância do efeito de QTL para CA no cromossomo 8 entre as posições 96,3 e 106 cM utilizando animais provenientes de cruzamentos das raças puras Meishan, Pietrain e porco selvagem europeu. Zhang et al. (2009) utilizou animais cruzados Duroc Branco x Erhualian chinês utilizando marcadores microssatélites e identificados QTLs significativos para CA no SSC8 a 96 cM, o qual explicou 3,27% da variância fenotípica. Wang et al. (2015), já utilizando o conceito de associação genômica ampla (GWAS), reportou região significativa para CA em animais Duroc entre 97 e 119 cM do cromossomo 8.

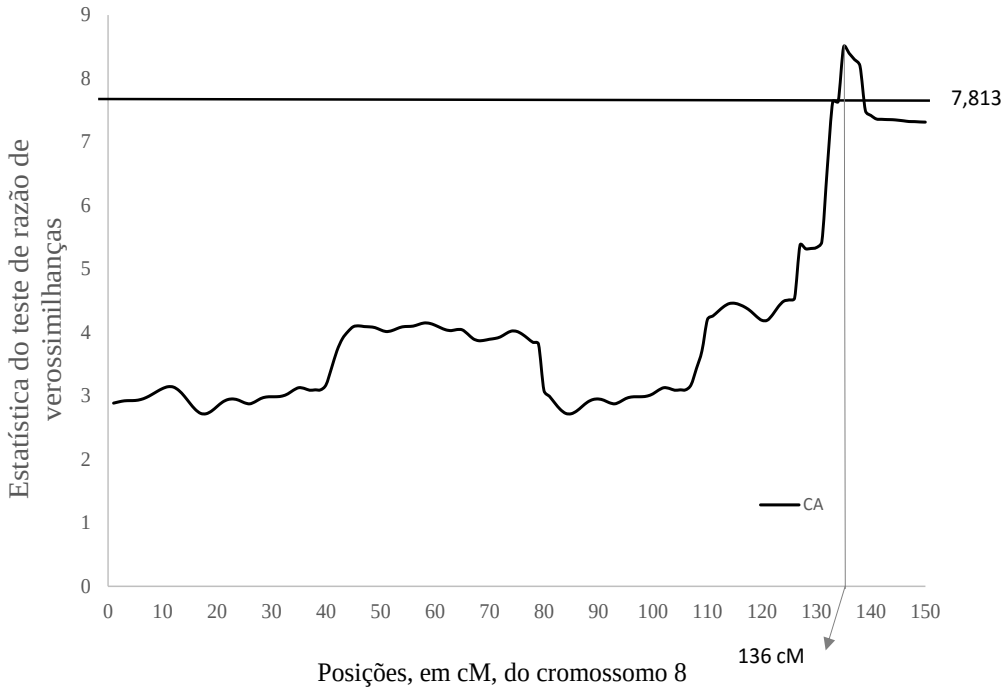
Além do cromossomo 8, estudos relatam a presença de QTLs para conversão alimentar em outros cromossomos. Duthie et al. (2008) utilizaram animais de três gerações de irmãos completos provenientes do cruzamento entre Pietran e linhagem comercial genotipados para 51 marcadores microsatélites e detectaram QTLs para

conversão alimentar no SSC2 a 3 cM, SS4 a 20 cM, e no SSC10 a 89 cM. Zhang et al. (2009) ao estudar animais cruzados (Duroc Branco x Erhualian chinês) também reportaram efeitos significativos de QTLs para CA no SSC3 a 74 cM e no SSC7 a 59 cM, os quais explicaram, respectivamente, 3,82% e 7,63% da variância fenotípica. Gilbert et al. (2010) estudaram uma população de retrocruzamento de Large White e Pietrain genotipada via marcadores microssatélites e detectaram QTLs para CA no SSC6 a 125 cM e no SSC7 a 74 cM.

Goutam et al. (2014) utilizaram GWAs para identificação de QTLs para CA e identificaram marcadores SNPs significativos no SSC4 (63.8 e 64 Kb) e no SSC14 (12.04, 12,14. 12.15, 12.18, 12.20, 12.21, 12.23, 12.49 Kb). Os autores comentam que os genes HIF1AN (fator 1 de Hipoxia inducida e inibidor da subunidade alfa) e LBX1 (Ladybird Homeobox 1), localizados respctivamentes no SSC4 e SSC17, são dois genes candidatos para CA em suínos.



(a)



(b)

Figura 1. Perfis da estatística dos testes de razão de verossimilhanças (TRV) relacionados com a comparação entre o modelo LA (análise de ligação) completo (com efeito aleatório de QTL e poligênico) e o modelo nulo (apenas com o efeito poligênico) para as características consumo de ração diário (CR) e ganho de peso diário (GPD) (a); e conversão alimentar (CA) (b).

4.3 Ajuste dos modelos LALDA

Uma vez que as posições dos QTLs reportados nos itens anteriores não correspondem exatamente as posições dos marcadores SNPs no mapa físico, os efeitos genotípicos de QTL nestas posições foram incluídos nos modelos Bayesianos de predição genômica.

Respectivamente para as posições dos efeitos significativos de QTL para GPD (91 cM) e CR (101 cM) no SSC1, e CA (136 cM) no SSC8, os marcadores SNPs mais próximos estavam situados nas posições 80,5 (ALGA0005078) e 107,2 (ALGA0005838) cM no SSC1; e nas posições 132,3 (ALGA0049550) e 138,2 (ALGA0050287).

As matrizes IBD genotípicas calculadas pelo software QXPAK nas posições indicadas na Figura 1 foram utilizadas como matrizes de covariâncias associadas aos efeitos genotípicos de QTL (componente LA) adicionados aos modelos de regressão Bayesianos (LAD) a fim de compor os modelos LALDA propostos no presente estudo. É importante ressaltar que dado o pequeno número de marcadores SNPs (237), o efeito poligênico aditivo fundamentado na matriz de parentesco tradicional também foi contemplado nos modelos em questão.

Nesta primeira análise comparativa entre os modelos LALDA, foram considerados todas as 341 observações das três cara caraterísticas estudadas (GPD, CA e CR), ou seja, foi considerado o arquivo completo de dados a fim de testar a qualidade de ajuste de cada modelo via DIC (Deviance Information Criterion), e não suas capacidades preditivas (explorada posteriormente sobre o enfoque de validação cruzada).

A convergência das cadeias MCMC para a variância residual de todos os modelos considerados foi verificada via teste de Geweke, no qual a hipótese de nulidade representa a situação de convergência (estacionaridade) da cadeia. Os p-valores referentes a este teste variaram de 0,0564 (LALDA Bayes A) a 0,9876 (LALDA BRR), ou seja, todos os modelos apresentaram convergência para a variância residual. Uma vez que o pacote BGLR do software R apenas reporta a cadeia MCMC para a variância residual, para os demais parâmetros do modelo não foi possível verificar a convergência das cadeias mencionadas.

É importante ressaltar que a característica CA, por ser definida como a razão entre as características CR e GPD, teoricamente tende a apresentar ausência de normalidade. Uma vez que a distribuição Normal foi assumida na construção da função de verossimilhança, os valores das características foram submetidos ao teste de aderência de Shapiro-Wilk via função shapiro.test do software R. Todas as características se distribuíram segundo uma distribuição Normal (p-valores variaram entre 0,0656 para LALDA Bayes B e 0,0786 para LALDA Bayes BC), o que valida, de certa forma, a opção por esta distribuição na construção da função de verossimilhança. Faz-se necessário comentar que este fato de testar as caracter´siticas não implica em um confundimento entre abordagens frequentista e Bayesiana, mas sim em uma forma de conferir se a distribuição para assumida para os dados observados está em consonância com a distribuição assumida teoricamente para os mesmos na função de verossimilhança.

A Tabela 3 mostra os valores de DIC para todos os modelos comparados (alfabeto bayesiano - BRR, BA, BB, BC e BL – com efeito poligênico aditivo, considerados como modelos LDA; e estes mesmos modelos com adição do efeito genotípico de QTL, considerado como modelo LALDA).

Tabela 3. Valores de DIC (Deviance Information Criterion) para modelos LALDA (alfabeto Bayesiano + poligênico aditivo + genotípico de QTL) e LDA (alfabeto Bayesiano + poligênico aditivo) para as características ganho de peso diário (GPD), consumo diário de ração (CR) e conversão alimentar (CA).

	GPD		**CR**		**CA**	
Alfabeto bayesiano[1]	**LALDA**	**LDA**	**LALDA**	**LDA**	**LALDA**	**LDA**
BRR	2624,2	2626,7	4109,1	4117,6	3288,0	3297,2
BA	2622,4	2624,8	4103,2	4113,2	3284.6	3293,8
BB	2623,9	2625,8	4106,2	4116,8	3286,8	3296,1
BC	2625,3	2628,1	4111,3	4119,1	3291.7	3299,0
BL	2623,4	2625,0	4105,3	4115,2	3286,1	3295,7

[1] Bayesian Ridge Regression (BRR), Bayes A (BA), Bayes B (BB), Bayes C (BC) e Bayesian LASSO (BL).

Os modelos LALDA apresentaram melhor qualidade de ajuste em relação aos modelos LDA ao considerar todos os modelos do alfabeto bayesiano (BRR, BA, BB, BC e BL). A diferença entre os mesmos foram mais evidentes para as características CR e CA, e menos expressiva para a característica GPD. Em termos de qualidade de ajuste, os modelos LALDA propostos no presente trabalho mostraram-se efetivos devido principalmente ao fato de se utilizar um pequeno número de marcadores. A abordagem proposta possibilitou que regiões relevantes (efeito de QTL significativo) que a princípio não seriam consideradas diretamente nas análises LDA (regressão Bayesiana) por não estarem fisicamente na posição dos SNPs considerados, pudessem ser adicionadas e contribuir de alguma forma para a melhoria da qualidade de ajuste.

Faz-se necessário comentar que em situações nas quais as posições de QTLs com efeitos significativos na análise LA coincidem com as posições físicas dos SNPs a serem considerados na análise LDA, estudos complementares devem ser realizados a fim de evitar colinearidade entre os efeitos. Por exemplo, devem ser considerados modelos LALDA contemplando o marcador no componente LDA (variável regressora) e omitindo o mesmo do componente LA (removendo o efeito genotípico de QTL), e vice-versa. Assim, estes diferentes modelos também devem ser comparados quanto a qualidade de ajuste e discutidos em termos do conceito geral de análise de ligação e de desequilíbrio de ligação.

Neste sentido, uma vez que o modelo contemplando o marcador no componente LDA seja superior em termos de qualidade de ajuste, pode-se inferir que a associação

entre marcador e QTL é uma propriedade da população como um todo, de forma que tal associação seja compartilhada por todos indivíduos e persista por várias gerações.

Diferentemente, uma vez que o modelo assumindo ao marcador no componente LA seja o de melhor ajuste, é possível inferir que o desequilíbrio de ligação e existe dentro de famílias ou cruzamentos específicos, o qual pode ser dissolvido por recombinações após algumas poucas gerações. Assim, espera-se que a associação entre marcadores e QTLs permaneça ativa apenas dentro de famílias, e por poucas gerações (Resende et al., 2012).

Esta diferenciação de modelos poderia ter aplicabilidade, por exemplo, em programas de melhoramento que utilizam predições genômicas considerando gerações futuras como população de seleção, de forma que modelos LALDA apontando para regiões relevantes no componente LA demandariam uma reestimação dos efeitos em um menor intervalo de tempo (ou seja, a população de treinamento deveria ser atualizada com maior frequência).

O modelo BA, seguido pelos modelos BB e BL, apresentaram os menores valores de DIC em ambas abordagens (LALDA e LDA) para todas as características avaliadas, indicando que o fato de considerar a priori uma variância para cada marcador implicou em melhor ajuste do modelo, independente de se incluir ou não o efeito genotípico de QTL.

A superioridade do BA quando comparada aos modelos BB e BL permite inferir que a seleção de variáveis, implementada via mistura de distribuições Normais no BB e parâmetro de regularização no BL, não mostrou-se efetiva. Além disso, ao se considerar modelos que assumem seleção de variáveis quando a mesma não é demandada implicou, de certa forma, em prejuízos na qualidade do ajuste. Talvez isto possa ser explicado por algum tipo de sobreajuste ("overfiting") em relação aos parâmetros (parâmetro "pi" no BB e "lambda" no BL) que direcionam a seleção de variáveis.

Em síntese, o pequeno número de marcadores SNPs utilizados (e elevado espaçamento entre os mesmos dentro de cada cromossomo) seria o fator responsável pela falta de efetividade dos modelos que exercem controle de variáveis e assume uma variância para cada marcador (BB e BL) em relação ao modelo que contempla apenas esta última situação (BA). A ineficiência da seleção de variáveis também foi verificada ao nível dos dois modelos com menor qualidade de ajuste, uma vez que o BC que assume uma mesma variância para todos os marcadores e exerce seleção de variáveis mostrou

qualidade inferior ao BRR que dispõe da mesma pressuposição quanto as variâncias dos marcadores, porém não considera a seleção de variáveis.

Conforme comentado anteriormente, o DIC avalia os modelos quanto a qualidade de ajuste a um certo conjunto de dados. Porém, o conceito de capacidade preditiva, o qual é de extrema relevância para a área de seleção genômica, não é contemplado no DIC. De forma geral, segundo Bishop (2007), um modelo que se ajusta bem a um conjunto de dados pode não ser um bom modelo para predizer novos valores com base nas estimativas dos parâmetros obtidas. Geralmente este problema é denominado de sobreajuste ("overfitting") e deve ser investigado via análise de validação cruzada.

4.4 Análise de validação cruzada para os modelos LALDA

Os mesmos modelos comparados no item 4.3 também foram submetidos a uma análise de validação cruzada particionada em cinco arquivos de dados ("5-fold cross-validation"), com quatro deles contendo 68 animais e um contendo 69 animais. Foram então realizadas cinco análises, de modo que em cada uma um arquivo foi removido do conjunto de dados para compor a população de validação, e os outros quatro foram utilizados na estimação dos valores genéticos totais.

Para o modelo LALDA, na população de treinamento foram estimados os efeitos de cada marcador bem como o vetor de valor genético poligênico (considerando todos indivíduos na matriz de parentesco) e o vetor de efeito genotípico de QTL (considerando todos os indivíduos na matriz IBD genotípica). Desta forma, os indivíduos da população de validação participaram da análise na população de treinamento ao considerar seus valores fenotípicos como não observados ("missing values"). Por utilizar todos os indivíduos na matriz de parentesco e na matriz IBD, via inferência Bayesiana (tal como BLUP tradicional) foi possível estimar um vetor de valores genéticos poligênicos ($\hat{\mathbf{u}}$) e aditivos de QTL ($\hat{\mathbf{w}}$) para os mesmos na própria população de treinamento. Assim, o valor genético genômico ($\hat{\mathbf{g}}$) do componente LDA para os indivíduos da população de validação foi obtido pela multiplicação do vetor de efeitos de marcadores estimados na população de treinamento (g_i) pela matriz de genótipos dos indivíduos da população de validação, o que corresponde a $\hat{\mathbf{g}} = \sum_{i=1}^{I} \mathbf{x}_i^{val} \hat{g}_i$, sendo $\mathbf{x}_i^{val}$ o vetor de genótipos do marcador i para os indivíduos da população de validação. Em resumo, o vetor de valores genéticos total foi dado por $\hat{\mathbf{y}} = \hat{\mathbf{g}} + \hat{\mathbf{u}} + \hat{\mathbf{w}}$, o qual foi correlacionado com o vetor de fenótipos pré-corrigidos para efeitos sistemáticos para calcular a capacidade preditiva mostrada na Tabela 4.

Tabela 4. Capacidade preditiva (correlação entre preditos e observados) para modelos LALDA (alfabeto Bayesiano + poligênico aditivo + genotípico de QTL) e LDA (alfabeto Bayesiano + poligênico aditivo) para as características ganho de peso diário (GPD), consumo diário de ração (CR) e conversão alimentar (CA).

	GPD			CR			CA			Media de Eficiência
Alfabeto bayesiano[1]	LALDA	%	LDA	LALDA	%	LDA	LALDA	%	LDA	%
BRR	0,32	14	0,28	0,26	13	0,23	0,28	17	0,24	15
BA	0,36	16	0,31	0,29	16	0,25	0,32	19	0,27	17
BB	0,33	14	0,29	0,27	13	0,24	0,30	20	0,25	16
BC	0,32	14	0,28	0,26	18	0,22	0,27	17	0,23	16
BL	0,33	14	0,29	0,26	13	0,23	0,29	21	0,24	16

[1] Bayesian Ridge Regression (BRR), Bayes A (BA), Bayes B (BB), Bayes C (BC) e Bayesian LASSO (BL).

De forma geral, embora as diferenças em termos de capacidade preditivas entre modelos LALDA e LDA tenham sido pouco expressivas, nota-se na Tabela 4 que os modelos LALDA tenderam a apresentar maiores valores de correlação entre valores genéticos preditos e valores fenotípicos pré-corrigidos para efeitos sistemáticos. Isto foi observado para todas as características. Mesmo utilizando um conjunto reduzido de observações, apenas 68 animais na população de validação, tais resultados podem ser considerados animadores, mesmo porque corroboram com aqueles obtidos para qualidade de ajuste via DIC contemplados no item anterior.

No presente estudo, embora as diferenças entre as acurácias proporcionadas por cada método tenham sido relativamente de baixas magnitudes para todas as características estudadas (GPD, CR e CA), nota-se que os métodos que contemplam, a priori, uma variância por marcador (BA, BB e BL, nesta ordem), apresentaram uma melhor qualidade de predição em relação aos métodos que consideram, a priori, todos os marcadores igualmente relevantes para característica estudada (BRR e BC). Novamente, vale ressaltar que o pequeno número de marcadores considerados no presente estudo privilegiou o método BA que não exerce seleção de variáveis.

Embora existam poucas referências comparando os modelos LALDA com os modelo LDA, Boichard et al. (2012) estudando toda a população de gado Holandês da França, reportou que as capacidades preditivas para as características produção total de leite, produção total de proteína, produção total de gordura, porcentagem de proteína no leite, porcentagem de gordura no leite e fertilidade foram 0,60 e 0,56, 0,57 e 0,55, 0,66 e

0,59, 0,73 e 0,73, 0,81 e 0,72, e 0,39 e 0,35, respectivamente para os modelos LALDA e LDA. Em geral, embora os autores tenham utilizado um número muito maior de animais e marcadores, o modelo LALDA mostrou-se satisfatório e vem sendo atualmente utilizado para a avaliação genética nacional de gado de leite na França.

Também na área de gado de leite, Luan et al. (2012) estudando uma população de animais Pardo Suíço da Itália, compararam a capacidade preditiva de modelos combinando informações de análise de ligação e de análise de desequilíbrio de ligação. Os autores encontraram os valores de 0,60 e 0,59, 0,60 e 0,58, e 0,63 e 0,61, respectivamente para os modelos LALDA e LDA considerando as características produção total de leite, produção total de proteína, produção total de gordura. Embora as diferenças tenham sido pouco expressivas, os autores argumentam que as mesmas tendem a aumentar ao se trabalhar com populações cruzadas com diferentes estruturas de famílias dentro de cada raça pura.

4.5 Estimativas de herdabilidade

Uma vez que o modelo LALDA via Bayes A (BA) mostrou-se como o de melhor ajuste por apresentar menor valor de DIC, e também maiores valores de capacidade preditiva na análise de validação cruzada, o mesmo foi escolhido para calcular as estimativas de variâncias genéticas aditivas e herdabilidade para cada uma das características estudadas.

Os valores de h^2 foram, respectivamente, 0,38, 0,21 e 0,23 para GPD, CR e CA. Uma vez que a ferramenta estatística utilizada nas análises, neste caso o pacote BGLR do software R, salva a cadeia MCMC apenas para a variância do erro, não foi possível obter a distribuição a posteriori para h^2, o que implica na ausência de inferências intervalares para este parâmetro. Dessa forma, as estimativas reportadas foram obtidas a partir das médias a posteriori para as variâncias genética aditiva ($\hat{\sigma}_g^2 = \sum_{i=1}^{I} 2\hat{p}_i(1-\hat{p}_i)\hat{\sigma}_i^2$), poligênica aditiva ($\hat{\sigma}_u^2$), genotípica de QTL ($\hat{\sigma}_w^2$) e residual ($\hat{\sigma}_e^2$).

Tendo em vista os resultados da Tabela 4 apresentados no item anterior, é possível associar os maiores valores de capacidade preditiva para GPD quando comparado as outras duas características a sua maior herdabilidade. Embora esta diferença reportada na Tabela 4 seja de pequena magnitude, há razões para acreditar que a mesma é proveniente desta maior estimativa de herdabilidade para GPD.

Em relação a estimativas de herdabilidade baseando-se apenas no efeito poligênico aditivo via matriz de parentesco tradicional, as estimativas de h^2 nesta mesma

população F2 foram reportadas por Mendonça et al. (2012) como sendo 0,40, 0,19 e 0,25, respectivamente para as características GPD, CR e CA. Portanto, nota-se que as estimativas reportadas no presente trabalho utilizando marcadores moleculares (sob os enfoques LDA e LA) adicionalmente a matriz de parentesco não resultou em alterações expressivas das mesmas. De acordo com outros estudos, as estimativas de herdabilidade para ganho de peso diário variam de 0,13 a 0,40; e para conversão alimentar de 0,19 a 0,42 (Torres Filho, 2001). Roso et al. (1995) reportou estimativas de herdabilidade iguais a 0,40, 0,25, respectivamente para as características GPD e CA.

5 CONCLUSÃO

Em síntese, o modelo LALDA via Bayes A (BA) mostrou a melhor qualidade de ajuste via DIC (Deviance Information Criterion) e maior capacidade preditiva quando comparado com os demais modelos LALDA (BRR, BB, BC e BL) para todas características estudadas. Embora de forma discreta, esta superioridade também se verificou ao comparar o modelo em questão com modelos alternativos que não contemplaram o efeito aleatório de QTL (modelos LDA tradicionais de SG), ou seja, o modelo LALDA proposto mostrou-se eficiente e plausível de ser implementado por meio de softwares livres.

REFERÊNCIAS BIBLIOGRÁFICAS

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007) Discrete multivariate analysis: theory and practice. Springer Science & Business Media.

Baes, C., Mayer, M., Tetens, J., Liu, Z., Reinhardt, F., Thaller, G., & Reinsch, N. (2010) Refined mapping of a QTL for somatic cell score on BTA27 in the German Holstein using combined linkage and linkage disequilibrium analysis. Canadian journal of animal science, 90(2), 169-178.

Beeckmann, P., Moser, G., Bartenschlager, H., Reiner, G., Geldermann, H. (2003) Linkage and QTL mapping for Sus scrofa chromosome 8. J Anim Breed Genet. 120:66–73.

Bercovici, S., Meek, C., Wexler, Y., & Geiger D. (2010) Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. Bioinformatics 26:175-182.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N. (2012) Genomic Selection in French Dairy Cattle. Anim Prod Sci 52: 115–120.

Brand, B., Baes, C., Mayer, M., Reinsch, N., Kühn, C. (2009) Identification of a two-marker-haplotype on Bos taurus autosome 18 associated with somatic cell score in German Holstein cattle. BMC Genet. 10:50.

Dassonneville, R., Fritz, S., Ducrocq, V., & Boichard, D. (2012) Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of dairy science, 95(7), 4136-4140.

De los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., & Cotes, J. M. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1), 375-385.

Duthie, C., Simma, G., Doeschl-Wilson, A., Kalm, E., Knap, P., & Roehe, R. (2008). Quantitative trait loci for chemical body composition traits in pigs and their positional associations with body tissues, growth and feed intake. Animal Genetics, 39 (2), 130-140.

Fernando, R. L., & Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol., 21: 467–477

Gelfand, I. M., & Silverman, R. A. (2000). Calculus of variations. Courier Corporation.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis (2nd ed.). Boca Raton: Chapman and Hall/CRC.

Geweke, J. (1992), Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. Bayesian Statistics 4 (Eds.) Oxford University Press, 169-193.

Gianola, D., De los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics, 183(1), 347-363.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. Genetics, 194(3), 573-596.

Gilbert, H., Riquet, J., Gruand, J., Billon, Y., Feve, K., Sellier, P., Noblet, J., Bidanel, J. P. (2010) Detecting QTL for feed intake traits and other performance traits in growing pigs in a Pietrain-Large White backcross. Animal, 4: 1308-1318.

Goutam, S., Kadlecová, V., Hornshøj, O., Nielsen, B., & Christensen, O. (2014). A genome-wide association scan in pig identifies novel regions associated with feed efficiency trait. Journal of Animal Science, 91 (3), 1041-1050.

Habier, D., Fernando, R. L., Dekkers, J. C. (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177: 2389-2397.

Habiel, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics, 12: 1471–1484.

Haley, C. S., Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69 (4), 315-324.

Hernandez-Sanchez, J., Grunchec, J. A., Knott, S. (2009) A web application to perform linkage disequilibrium and linkage analyses on a computational grid. Bioinformatics, 25: 1377-1383.

Heidelberger, P., & Welch, P. D. (1983) Simulation run length control in the presence of an initial transient. Oper. Res,. 31:1109–1144.

Jiao, S., Maltecca, C., Gray, K., & Cassady, P. (2014) Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: II. Genome wide association. Journal of Animal Science,(92), 2846–2860 .

Liu, G., Jennen, D. G.,Tholen, E., Juengst, H., Kleinwächter, T., Hölker, M. (2007) A genome scan reveals QTL for growth, fatness, leanness and meat quality in a Duroc-Pietrain resource population. Animal Genetics, 38 (3), 241-252.

Losinger, W. C., & Sampath, R. K. (2000) Economies of scale in the production of swine manure. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 52(3), 285-294.

Luan, T., Woolliams, J. A., Ødegård, J., Dolezal, M., Roman-Ponce, S. I., Bagnato, A., & Meuwissen, T. H. (2012) The importance of identity-by-state information for the accuracy of genomic selection. Genetics Selection Evolution, 44(1), 1-15.

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Mendonça, P., Lopes, P., Braccini Neto, J., Carneiro, P., Torres, R., Guimarães, S., & Veroneze, R. (2012) Estimação de parâmetros genéticos de uma população F2 de suínos. Revista Brasileira de Saúde e Produção Animal, 13(2), 330-343

Nagamine, Y. Transformation of QTL genotypic effects to allelic effects. (2005) Genet. Sel. Evol. 37, 579–584.

Olsen, H. G., Meuwissen, T. H., Nilsen, H., Svendsen, M; Lien, S. (2009) Fine mapping of quantitative trait loci on bovine chromosome 6 affecting calving difficulty. J. Dairy Sci, 91:4312–4322.

Paulino, C., Turkman, M. A., Murteira, B. (2003) Estatística Bayesiana. Fundação Calouste Gulbenkian, Lisboa.

Paixão, D.M., Braccini Neto J., Paiva, S. R., Carneiro, P. L. S., Pinto, A. P. G., Sousa, K. R. S., Nascimento, C. S., Verardo, L. L., Hidalgo, A. M., Lopes, O. S. (2013). Detecção de locos de características quantitativas nos cromossomos 1, 2, 3, 12, 14, 15 e X de suínos: Características de desempenho. Arq Bras Med Vet. 65:213–220.

Pérez-Enciso, M., Varona, L., & Rothschild, M. F. (2000). Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov Chain method. Genetics Selection Evolution, 32(5), 1.

Pérez-Enciso, M., & Misztal, I. (2011) Qxpak. 5: old mixed model solutions for new genomics problems. BMC bioinformatics, 12(1), 1.

Pikkuhookana, P., Sillanpaa, M.J. (2014) Combined linkage disequilibrium and linkage mapping: Bayesian multilocus approach. Heredity (Edinb) 112: 351-360.

R Development Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler. Bayesian statistics, 4(2), 763-773.

Resende, M. D. V.; Silva, F. F. ; Lopes, P. S. ; Azevedo, C. F. (2012). Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM). Estatística Espacial. 1. ed. v. 1. 291 p

Roso, V.M., Fries, L.A., Martins, E.S. (1995) Parâmetros genéticos em características de desempenho e qualidade carcaça em suínos da raça Duroc. Revista Brasileira de Zootecnia, v.24, n.2, p.310-16.

Souza, K.R.S. (2008).Características Quantitativas (QTL) nos cromossomos 5, 7 e 8 de suínos. (Mestrado em Genética e Melhoramento) - Universidade Federal de Viçosa.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583-639.

Sun, Y., Liu, R., Zhao, G., Zheng, M., Sun, Y., Yu, X., Li, P., Wen, J. (2014). Genome-Wide Linkage Analysis and Association Study Identifies Loci for Polydactyly in Chickens. G3 Gene|Genomic|Genetic 21;4(6):1167-72.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Torres, R.A. (2001). Avaliação genética de características de desempenho e reprodutivas em suínos. (Dissertação de Mestrado em Genética e Melhoramento Animal) Universidade Federal de Viçosa.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), 4414-4423.

Wientjes Y. C. J., Veerkamp R. F., Calus M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193: 621–631.

Zhang, Z., Ren, J., Ren, D., Ma, J., Guo, Y., & Huang , L. (2009). Mapping quantitative trait loci for feed consumption and feeding behaviors in a White Duroc x Chinese Erhualian resource population. Journal of Animal Science, (11), 3458-3463.

ANEXOS

Input de Programa R para leitura e formatação dos arquivos de dados fenotípicos, genotípicos e de pedigree conforme software QXPAK

```
setwd("/media/fabyano/SAMSUNG/backup_junho_2016/comp_novo_maio_2016/da
dos_elcer")
data=read.table("GWS.txt",h=T)
map=read.table("mapa.txt",h=T)
snp0=data[,65:301]  #only colunms of SNPs
snp01=data.frame(colnames(snp0),t(snp0))
colnames(snp01)=c("snp",data[,1])
snp02=merge(snp01,map, by=intersect("snp","snp"))
aux=snp02[,c(1,347,348)]
table(aux[,2])
# 1 17  4  7  8  x
# 56 25 54 59 31 12
snp=snp02[,-c(1,347,348)]

#chr1
snp_c1=snp[1:56,]
snp_c1t=t(snp_c1)
snp_c1t_two=matrix(0,nrow(snp_c1t),ncol(snp_c1t))
snp_c1t_two[snp_c1t==0]<-c("1 1")
snp_c1t_two[snp_c1t==1]<-c("1 2")
snp_c1t_two[snp_c1t==2]<-c("2 2")
write.table(snp_c1t_two,"snp_c1t_two1.txt",col.names=FALSE,row.names=F
ALSE, quote=FALSE)
snp_c1t_two2=read.table("snp_c1t_two1.txt")
chr1=rbind(t(matrix(c("chr1",rep(c("       "),(ncol(snp_c1t_two2))))))),
as.matrix(cbind( colnames(snp),snp_c1t_two2)))
chr1_1=chr1[chr1[,1]!=643&chr1[,1]!=851&chr1[,1]!=922&chr1[,1]!=1214,]
#IDs not specified in pedigree
write.table(chr1_1,"chr1.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)

#chr17
snp_c17=snp[57:81,]
snp_c17t=t(snp_c17)
snp_c17t_two=matrix(0,nrow(snp_c17t),ncol(snp_c17t))
snp_c17t_two[snp_c17t==0]<-c("1 1")
snp_c17t_two[snp_c17t==1]<-c("1 2")
snp_c17t_two[snp_c17t==2]<-c("2 2")
```

```
write.table(snp_c17t_two,"snp_c17t_two1.txt",col.names=FALSE,row.names
=FALSE, quote=FALSE)

snp_c17t_two2=read.table("snp_c17t_two1.txt")

chr17=rbind(t(matrix(c("chr17",rep(c("    "),(ncol(snp_c17t_two2))))),
as.matrix(cbind( colnames(snp),snp_c17t_two2)))

chr17_1=chr17[chr17[,1]!=643&chr17[,1]!=851&chr17[,1]!=922&chr17[,1]!=
1214,] #IDs not specified in pedigree

write.table(chr17_1,"chr17.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


#chr4

snp_c4=snp[82:135,]

snp_c4t=t(snp_c4)

snp_c4t_two=matrix(0,nrow(snp_c4t),ncol(snp_c4t))

snp_c4t_two[snp_c4t==0]<-c("1 1")

snp_c4t_two[snp_c4t==1]<-c("1 2")

snp_c4t_two[snp_c4t==2]<-c("2 2")

write.table(snp_c4t_two,"snp_c4t_two1.txt",col.names=FALSE,row.names=F
ALSE, quote=FALSE)

snp_c4t_two2=read.table("snp_c4t_two1.txt")

chr4=rbind(t(matrix(c("chr4",rep(c("      "),(ncol(snp_c4t_two2))))),
as.matrix(cbind( colnames(snp),snp_c4t_two2)))

chr4_1=chr4[chr4[,1]!=643&chr4[,1]!=851&chr4[,1]!=922&chr4[,1]!=1214,]
#IDs not specified in pedigree

write.table(chr4_1,"chr4.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


#chr7

snp_c7=snp[136:194,]

snp_c7t=t(snp_c7)

snp_c7t_two=matrix(0,nrow(snp_c7t),ncol(snp_c7t))

snp_c7t_two[snp_c7t==0]<-c("1 1")

snp_c7t_two[snp_c7t==1]<-c("1 2")

snp_c7t_two[snp_c7t==2]<-c("2 2")

write.table(snp_c7t_two,"snp_c7t_two1.txt",col.names=FALSE,row.names=F
ALSE, quote=FALSE)

snp_c7t_two2=read.table("snp_c7t_two1.txt")

chr7=rbind(t(matrix(c("chr7",rep(c("      "),(ncol(snp_c7t_two2))))),
as.matrix(cbind( colnames(snp),snp_c7t_two2)))

chr7_1=chr7[chr7[,1]!=643&chr7[,1]!=851&chr7[,1]!=922&chr7[,1]!=1214,]
#IDs not specified in pedigree

write.table(chr7_1,"chr7.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


#chr8

snp_c8=snp[195:225,]

snp_c8t=t(snp_c8)
```

```
snp_c8t_two=matrix(0,nrow(snp_c8t),ncol(snp_c8t))

snp_c8t_two[snp_c8t==0]<-c("1 1")

snp_c8t_two[snp_c8t==1]<-c("1 2")

snp_c8t_two[snp_c8t==2]<-c("2 2")

write.table(snp_c8t_two,"snp_c8t_two1.txt",col.names=FALSE,row.names=F
ALSE, quote=FALSE)

snp_c8t_two2=read.table("snp_c8t_two1.txt")

chr8=rbind(t(matrix(c("chr8",rep(c("       "),(ncol(snp_c8t_two2))))))),
as.matrix(cbind( colnames(snp),snp_c8t_two2)))

chr8_1=chr8[chr8[,1]!=643&chr8[,1]!=851&chr8[,1]!=922&chr8[,1]!=1214,]
#IDs not specified in pedigree

write.table(chr8_1,"chr8.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


#chr18

snp_c18=snp[226:237,]

snp_c18t=t(snp_c18)

snp_c18t_two=matrix(0,nrow(snp_c18t),ncol(snp_c18t))

snp_c18t_two[snp_c18t==0]<-c("1 1")

snp_c18t_two[snp_c18t==1]<-c("1 2")

snp_c18t_two[snp_c18t==2]<-c("2 2")

write.table(snp_c18t_two,"snp_c18t_two1.txt",col.names=FALSE,row.names
=FALSE, quote=FALSE)

snp_c18t_two2=read.table("snp_c18t_two1.txt")

chr18=rbind(t(matrix(c("chr18",rep(c("    "),(ncol(snp_c18t_two2))))))),
as.matrix(cbind( colnames(snp),snp_c18t_two2)))

chr18_1=chr18[chr18[,1]!=643&chr18[,1]!=851&chr18[,1]!=922&chr18[,1]!=
1214,] #IDs not specified in pedigree

write.table(chr18_1,"chr18.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


##Phenotypes

pheno0=data[,43:45]

pheno0[is.na(pheno0)] <- c("0")

pheno=data.frame(data[,c(1,4:6)],rep(1,nrow(pheno0)),pheno0)

pheno1=pheno[pheno[,1]!=643&pheno[,1]!=851&pheno[,1]!=922&pheno[,1]!=1
214,] #IDs not specified in pedigree

write.table(pheno1,"pheno_end.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)


###pedigree

ped0=read.table("pedig.txt")

ped1=ped0[order(ped0[ ,1]), ]

write.table(ped1,"pedig_end.txt",col.names=FALSE,row.names=FALSE,
quote=FALSE)
```

Input de Arquivo de parâmetros do software QXPAK para implementar o modelo com efeito aleatório genotípico de QTL via matriz IBD

```
ML_OPTION

Y

DATAFILE

pheno_end.dat

OUTFILE

qxpak_elcer_NEW.out append

MARKERFILE

chr1.mkr

PEDIGREEFILE

pedig_end.ped

MARKER_POSITIONS

chr1 0.209568 290.184165  !primeira e última posição dos marcadores no cromossomo

QTL

qtl_1 ran_1 global

#gmol1 ran_mol global

EFFECT

mean cross 4

sex  cross 2

hal  cross 3

LOTE cross 6

id   cross 1 pedigree pedig_end.ped

TRAIT

t1 7 mean sex hal id qtl_1

TEST

qtl_1
```

Input de Programa R para implementação dos modelos LALDA Bayesianos (BRR, BA, BB, BC e BL) – análise de validação cruzada

```
setwd("C:\\Users\\Usuario\\Desktop\\programa_elcer_LALDA_janeo")

##################
#reading pedigree#
##################

ped=as.matrix(read.table("pedig_end.ped")[,-4])
#calculating A matrix
library(gap)
A=2*kin.morgan(ped)$kin.matrix

################################################
#reading IBD matrix at significant QTL positions#
################################################

tmp=read.table("zran.10000",skip=2)
id1<-tmp[,1]
id2<-tmp[,2]
 x<-(tmp[,3])
 IBD<-matrix(nrow=502,ncol=502,0)

 for(i in 1:length(id1))
{
        tmp1<-id1[i]
        tmp2<-id2[i]
        IBD[tmp1,tmp2]<-IBD[tmp2,tmp1]<-x[i]
 }

####################
#reading phenotypes#
####################

pheno=read.table("pheno_end.dat")
colnames(pheno)=c("id_geno", "sex", "lote","hal", "mean", "cr", "gdp", "ca")
id_geno=pheno[,1] #only genotyped IDs

################################################
##selecting genotyped animals in A and IBD matrices#
################################################

id=ped[,1] #All IDs from pedigree

#A matrix
```

```
colnames(A)=id    #adding ID names to columns of A matrix
rownames(A)=id    #adding ID names to rows of A matrix
A[1:10,1:10]

A1=A[intersect(id_geno,rownames(A)),]
dim(A1)
A2=t(A1)
A3=A2[intersect(id_geno,rownames(A2)),] # A3= A matrix considering only genotyped IDs
dim(A3)
A3[1:10,1:10]

#IBD matrix

colnames(IBD)=id      #adding ID names to columns of IBD matrix
rownames(IBD)=id      #adding ID names to rows of IBD matrix
IBD[1:10,1:10]

IBD1=IBD[intersect(id_geno,rownames(IBD)),]
dim(IBD1)
IBD2=t(IBD1)
IBD3=IBD2[intersect(id_geno,rownames(IBD2)),]  #  IBD3=  IBD  matrix  considering  only
genotyped IDs
dim(IBD3)
IBD3[1:10,1:10]

##########################
##reading genotype file #
##########################

geno=read.table("GWS.txt",h=T)
geno1=geno[,c(1,65:301)]  #subseting ID and colunms of SNPs
pheno_geno=data.frame(merge(pheno,geno1,by=intersect("id_geno","id_geno"))) #correct IDs
order in geno and pheno files
geno2=as.matrix(pheno_geno[,-c(1:8)]) #final genotype file

############################################################
##pre-adjusting phenotypic values for sistematic effects##
############################################################

##Daily feed intake - DFI
dfi_adj=     as.matrix(mean(pheno_geno$cr)     + lm(cr   ~ factor(sex) + factor(lote) +
factor(hal), data = pheno_geno)$residuals)

#Average daily gain - ADG
gdp_adj=as.matrix(mean(pheno_geno$gdp)      +  lm(gdp  ~  factor(sex)  +  factor(lote)  +
factor(hal), data = pheno_geno)$residuals)

#Feed conversion rate - FCR
```

```
fcr_adj=as.matrix(mean(pheno_geno$ca)       + lm(ca    ~  factor(sex)  +  factor(lote)  +
factor(hal), data = pheno_geno)$residuals)

#############LD model#############################

####################################################################

##Bayesian alphabet (LD) including pedigree based polygenic effect##

####################################################################



library(BGLR) #using BGLR package

y=fcr_adj     #analysis for trait "feed conversion rate"

#five groups cross-valid (68 68 68 68 69 IDs/group = 341

nf=5

groups=c(rep(1,68),rep(2,68),rep(3,68),rep(4,68),rep(5,69))

#creating matrices to store correlations

ryg_BRR=matrix(0,ncol=1,nrow=nf)

ryg_BA=matrix(0,ncol=1,nrow=nf)

ryg_BB=matrix(0,ncol=1,nrow=nf)

ryg_BC=matrix(0,ncol=1,nrow=nf)

ryg_BL=matrix(0,ncol=1,nrow=nf)



for(i in 1:nf) #loop to run all folds

{

  y2<-y

  y2[i==groups]<-NA  #Omitting known values

  yout<-y[i==groups] #Taking the omitted known values

  #Fitting Bayes RR, A, B, C and LASSO with polygenic effect

  fit_BRR=BGLR(y=y2,ETA=list(list(X=geno2,model='BRR'     ),  A=list(K=A3,model='RKHS')),
nIter=100,burnIn=3,thin=5,verbose=F)

  fit_BA=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesA'),  A=list(K=A3,model='RKHS')),
nIter=100,burnIn=3,thin=5,verbose=F)

  fit_BB=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesB'),  A=list(K=A3,model='RKHS')),
nIter=100,burnIn=3,thin=5,verbose=F)

  fit_BC=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesC'),  A=list(K=A3,model='RKHS')),
nIter=100,burnIn=3,thin=5,verbose=F)

  fit_BL=BGLR(y=y2,ETA=list(list(X=geno2,  model='BL'     ),  A=list(K=A3,model='RKHS')),
nIter=100,burnIn=3,thin=5,verbose=F)



  #calculating GEBVs + EBVs = Total prediction by using $yHat



  yhat_out_BRR=fit_BRR$yHat[i==groups]

  yhat_out_BA= fit_BA$yHat[i==groups]
```

```
  yhat_out_BB= fit_BB$yHat[i==groups]
  yhat_out_BC= fit_BC$yHat[i==groups]
  yhat_out_BL= fit_BL$yHat[i==groups]

  #calculating between Total prediction and adjusted phenotypes

  ryg_BRR[i,1]=cor(yout,yhat_out_BRR)
  ryg_BA[i,1]= cor(yout,yhat_out_BA)
  ryg_BB[i,1]= cor(yout,yhat_out_BB)
  ryg_BC[i,1]= cor(yout,yhat_out_BC)
  ryg_BL[i,1]= cor(yout,yhat_out_BL)
}

 #saving calculated correlations

write.table(ryg_BRR,"ryg_BRR.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BA,"ryg_BA.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BB,"ryg_BB.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BC,"ryg_BC.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BL,"ryg_BL.txt",quote=F,row.names=F,col.names=F)



#####################LALDA model############################
##############################################################################################
##Bayesian alphabet (LD) including pedigree based polygenic effect plus IBD based qtl
effect (LA)##
##############################################################################################



library(BGLR) #using BGLR package
y=fcr_adj     #analysis for trait "feed conversion rate"

#five groups cross-valid (68 68 68 68 69 IDs/group = 341
nf=5
groups=c(rep(1,68),rep(2,68),rep(3,68),rep(4,68),rep(5,69))

#creating matrices to store correlations

ryg_BRR1=matrix(0,ncol=1,nrow=nf)
ryg_BA1=matrix(0,ncol=1,nrow=nf)
ryg_BB1=matrix(0,ncol=1,nrow=nf)
ryg_BC1=matrix(0,ncol=1,nrow=nf)
ryg_BL1=matrix(0,ncol=1,nrow=nf)



for(i in 1:nf) #loop to run all folds
{
```

```
  y2<-y
  y2[i==groups]<-NA  #Omitting known values
  yout<-y[i==groups] #Taking the omitted known values

  #Fitting Bayes RR, A, B, C and LASSO with polygenic effect

  fit_BRR1=BGLR(y=y2,ETA=list(list(X=geno2,model='BRR'     ),  A=list(K=A3,model='RKHS'),
IBD=list(K=IBD3,model='RKHS')), nIter=100,burnIn=3,thin=5,verbose=F)
  fit_BA1=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesA'),  A=list(K=A3,model='RKHS'),
IBD=list(K=IBD3,model='RKHS')), nIter=100,burnIn=3,thin=5,verbose=F)
  fit_BB1=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesB'),  A=list(K=A3,model='RKHS'),
IBD=list(K=IBD3,model='RKHS')), nIter=100,burnIn=3,thin=5,verbose=F)
  fit_BC1=BGLR(y=y2,ETA=list(list(X=geno2,  model='BayesC'),  A=list(K=A3,model='RKHS'),
IBD=list(K=IBD3,model='RKHS')), nIter=100,burnIn=3,thin=5,verbose=F)
  fit_BL1=BGLR(y=y2,ETA=list(list(X=geno2,  model='BL'     ),  A=list(K=A3,model='RKHS'),
IBD=list(K=IBD3,model='RKHS')), nIter=100,burnIn=3,thin=5,verbose=F)

  #calculating GEBVs + EBVs = Total prediction by using $yHat

  yhat_out_BRR1=fit_BRR1$yHat[i==groups]
  yhat_out_BA1= fit_BA1$yHat[i==groups]
  yhat_out_BB1= fit_BB1$yHat[i==groups]
  yhat_out_BC1= fit_BC1$yHat[i==groups]
  yhat_out_BL1= fit_BL1$yHat[i==groups]

  #calculating between Total prediction and adjusted phenotypes

  ryg_BRR1[i,1]=cor(yout,yhat_out_BRR1)
  ryg_BA1[i,1]= cor(yout,yhat_out_BA1)
  ryg_BB1[i,1]= cor(yout,yhat_out_BB1)
  ryg_BC1[i,1]= cor(yout,yhat_out_BC1)
  ryg_BL1[i,1]= cor(yout,yhat_out_BL1)
}

 #saving calculated correlations

write.table(ryg_BRR1,"ryg_BRR1.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BA1,"ryg_BA1.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BB1,"ryg_BB1.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BC1,"ryg_BC1.txt",quote=F,row.names=F,col.names=F)
write.table(ryg_BL1,"ryg_BL1.txt",quote=F,row.names=F,col.names=F)
```