UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS

FERNANDA MARCONDES DE REZENDE

Incorporação de informações de marcadores genéticos em programas de melhoramento genético de bovinos de corte

Pirassununga – SP
2012

FERNANDA MARCONDES DE REZENDE

Incorporação de informações de marcadores genéticos em programas de melhoramento genético de bovinos de corte

(Versão corrigida)

Tese apresentada à Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo, como parte dos requisitos para a obtenção do Título de Doutor em Zootecnia.

Área de concentração: Qualidade e Produtividade Animal

Orientador: Prof. Dr. Flávio Vieira Meirelles
Co-orientadora: Dra. Noelia Ibáñez-Escriche

Pirassununga – SP

2012

Dados Internacionais de Catalogação na Publicação

Serviço de Biblioteca e Informação da Faculdade de Zootecnia e Engenharia de Alimentos
da Universidade de São Paulo

R467i

Rezende, Fernanda Marcondes de
Incorporação de informações de marcadores genéticos em programas de melhoramento genético de bovinos de corte / Fernanda Marcondes de Rezende. -- Pirassununga, 2012.
88 f.
Tese (Doutorado) -- Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo.
Departamento de Ciências Básicas.
Área de Concentração: Qualidade e Produtividade Animal.
Orientador: Prof. Dr. Flávio Vieira Meirelles.

1. Seleção assistida por marcadores 2. Inferência bayesiana 3. Valor genético molecular. I. Título.

A minha tão amada família, dedico este trabalho.

Ao meu mestre, meu cúmplice, meu amante e meu eterno incentivador,

José Bento, agradeço por fazer parte da minha história.

Agradecimentos

A Deus por guiar meu caminho e iluminar minhas escolhas e a Nossa Senhora por me proteger;

Aos meus pais e irmãs por serem meu porto seguro, sem vocês nada disso seria possível;

À minha família, avós, tios e primos pela torcida constante;

Ao Grupo de Melhoramento Animal e Biotecnologia, pela formação pessoal e profissional nesses 11 anos de convivência. Ao Prof. Bento por acreditar em mim, quando até eu duvidei. Ao Prof. Flávio pelo incentivo constante e por ter assumido minha orientação na reta final. Aos Profs. Joanir e Cesinha pelo convívio. À Elis, pela amizade, paciência e toda ajuda nesses anos. A todos os amigos da família GMAB, Adalfredo, Prof. Caju, Cucco, Gersinho, Prof. Heidge, Jane, Luis Fernando, Marina, Manicardi, Martini, Miguel, Mirele, Paulinho, Preto, Pri, Rodrigo, Roulber e Tércio;

À Noelia: "Gracias por todo que me enseñaste y por la amistad";

Aos amigos do IRTA e do Club Majestic em Lleida: Ana, Anália, Angela, Carlos, Cláudia, Davi, Esther, Hossein, Juan Pablo, Jules, Maestro, Noguera, Pedro, Raquel, Rosa, Toni, pelos bons momentos vividos;

À Teka, a "tia" Cecília e ao Fron por serem muito mais que bons amigos;

Ao "tio" Gerson e "tia" Lu por serem parceiros em qualquer situação;

Ao "tio" San e a Paula pela amizade e pelo apoio sempre;

Ao Neto, Nessa, Laura, Arthur, Davi, Mário e Caio pelo carinho;

À FAPESP pelo apoio financeiro que propiciou minha manutenção e dedicação exclusiva a essa pesquisa;

À CAPES e CNPq, pelo apoio ao Programa de Pós-Graduação em Zootecnia, área de Qualidade e Produtividade Animal, da FZEA/USP;

Aos funcionários FZEA/USP, em especial a Layla, Alecsandra, Gilson, Arnaldo, Renan, Kleber e Pedro pelo apoio constante;

Aos amigos da Merial/Igenity: Henry, Guilherme, Fernanda, Stewart e Nick;

À FUNPEC-RP que geriu os recursos destinados à pesquisa;

À Agro-Pecuária CFM Ltda. por disponibilizar os dados aqui analisados;

À Universidade de São Paulo, que por meio da FZEA proporcionou a realização desse trabalho;

A todos aqueles que por ventura não tenham sido citados e que de alguma maneira contribuíram para a realização desse trabalho fica aqui meu sincero agradecimento.

Pedras no caminho? Guardo todas, um dia vou construir um castelo...

Fernando Pessoa

APRESENTAÇÃO

Fernanda Marcondes de Rezende é Zootecnista formada pela Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo (FZEA/USP) em 2005. Após experiência na iniciativa privada junto à Sadia, em Uberlândia-MG, e ao Sebrae, em Piracicaba-SP, dedicou-se a obtenção do título de Mestre em Zootecnia - Qualidade e Produtividade Animal pela FZEA/USP, no período de 2007 a 2009. Desde 2007, como mestranda e, posteriormente, como doutoranda do Grupo de Melhoramento Animal e Biotecnologia, estudou a utilização de marcadores genéticos como ferramenta auxiliar na seleção de reprodutores. Em 2011, permaneceu por seis meses no Instituto de Investigación y Tecnología Agroalimentarias da Cataluña (IRTA) sob a supervisão da Dra. Noelia Ibáñez-Escriche desenvolvendo parte do presente trabalho. Durante todos esses anos de dedicação à pesquisa, acumulou conhecimentos relacionados ao gerenciamento e análise de dados, avaliação genética de animais de produção, ao estado da arte das análises de efeito de associação de marcadores em características de relevância econômica e noções de programação em R e Fortran. Tem experiência na área de Zootecnia, com ênfase em Produção Animal e Melhoramento Genético Animal, atuando principalmente nos seguintes temas: produção de frangos de corte, estimação de parâmetros genéticos, tendências genética e endogâmica em aves, ovinos e bovinos e utilização da informação molecular em programas de melhoramento genético animal.

RESUMO

REZENDE, F.M. **Incorporação de informações de marcadores genéticos em programas de melhoramento genético de bovinos de corte.** 2012. 88 f. Tese (Doutorado) – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2012.

A disponibilidade de informações baseadas nos marcadores genéticos surgiu como oportunidade de aprimorar os programas de melhoramento animal pela incorporação desses efeitos nas avaliações genéticas. Nesse contexto, o presente estudo teve como objetivos comparar modelos que consideraram ou não os efeitos dos marcadores para a estimação dos valores genéticos dos animais, bem como estimar os efeitos de substituição alélica dos marcadores por seis metodologias distintas (regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano) e avaliar o impacto da inclusão desses efeitos na acurácia das estimativas dos valores genéticos e os conflitos de seleção existentes aos serem comparadas as classificações dos animais com base nos valores genéticos clássicos e nos valores genéticos assistidos por marcadores. Dados de 83.404 animais pertencentes a um programa de seleção de animas da raça Nelore, mensurados para peso na desmama, ganho de peso pós-desmama, perímetro escrotal e escore de musculosidade, que corresponderam a 116.652 animais na matriz de parentesco, foram utilizados. Do total de animais com informações fenotípicas e genealógicas disponíveis, apenas 3.160 foram genotipados para 106 marcadores do tipo SNP. Os resultados obtidos para a comparação de modelos não demonstraram vantagens claras da inclusão conjunta dos efeitos poligênicos e dos marcadores nos modelos de avaliação genética, entretanto, os modelos que incluíram apenas o efeito dos marcadores tiveram os piores ajustes e desempenhos preditivos. As diferenças observadas entre as estimativas dos efeitos de substituição alélica dos marcadores pelas diferentes metodologias analisadas se devem à maneira como cada método regulariza esses efeitos. A incorporação das informações dos marcadores nas avaliações genéticas proporcionou, no geral, um aumento na acurácia das estimativas dos valores genéticos, especialmente para os tourinhos de reposição. Ao serem comparados os 20% melhores animais classificados com base no valor genético clássico e no valor genético assistido por marcadores, os maiores conflitos de seleção foram observados para os touros e tourinhos genotipados. Em suma, o presente projeto demonstrou que, embora a utilização de painéis de marcadores de muito baixa densidade não altere a capacidade preditiva dos modelos de avaliação genética, esses têm impacto na acurácia das estimativas dos valores genéticos.

Palavras-chave: seleção assistida por marcadores, inferência bayesiana, valor genético molecular.

ABSTRACT

REZENDE, F.M. **Incorporation of genetic markers information in beef cattle breeding programs.** 2012. 88 f. PhD Thesis – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2012.

The availability of molecular markers information turned out to be an opportunity to improve animal breeding programs, by the inclusion of those effects in the estimation of breeding values. Under that perspective, the aims of present research were to compare genetic evaluation models that assumed or not markers effects on the estimation of breeding values, as well estimate the allelic substitution effects of SNP markers applying six different methodologies (Bayesian multiple regression, Bayesian ridge regression, Bayes A, Bayes B, Bayes Cπ and Bayesian Lasso) and evaluate the impact of these effects on the reliability of breeding values and the divergences on animals classification based on classical breeding values and marker assisted breeding values. Data of 83,404 animals belonging to a Nellore beef cattle (*Bos indicus*) selection program, measured for post-weaning gain, scrotal circumference and muscle score, corresponding to 116,562 animals on the relationship matrix, were used. From those animals, a set of 3,160 animals with phenotypic and genealogy data available, was genotyped for a panel of 106 SNP markers. Model comparison results did not demonstrate clearly the advantage of assuming polygenic and markers effects together in genetic evaluation models, however, models that considered only markers effects presented the worst global fit and predictive ability. Differences observed on the markers effects estimates were due the shrinkage process applied by each method. The incorporation of markers information on genetic evaluations provided, in general, increases on the reliability of breeding values, mainly for replacement young animals. Comparing the 20% best animals classified by classical breeding value and marker assisted breeding value, the highest divergences were observed to sires and young bulls that were genotyped. Summarizing, although this research showed that the inclusion of very low density SNP chip information was not able to improve the predictive ability of genetic evaluation models, they increased the reliability of breeding values estimates.

Key-words: marker assisted selection, bayesian inference, molecular breeding value.

SUMÁRIO

1. Introdução

O foco principal dos programas de melhoramento é o acúmulo de ganho genético a cada geração, o qual é resultante da seleção e acasalamento de candidatos que são melhores que a média dos animais disponíveis na população. O grande desafio é avaliar corretamente o mérito genético dos animais superiores, que são os que de fato contribuirão para o maior desempenho da próxima geração.

No melhoramento genético clássico, as informações de pedigree e de desempenho dos animais candidatos, de seus ascendentes e colaterais são analisadas para obter a melhor predição linear não viesada (BLUP – *Best Linear Unbiased Prediction*; HENDERSON, 1975) do valor genético. Essa tem sido uma abordagem de sucesso, apesar de não fornecer estimativas de mérito genético de alta acurácia para animais jovens ou com poucos descendentes avaliados, exceto para características de alta herdabilidade.

O desenvolvimento da genética molecular permitiu a identificação de regiões polimórficas do DNA. Algumas dessas regiões foram associadas às alterações no desempenho dos animais para características de interesse econômico. A disponibilidade dessa informação molecular surgiu como oportunidade de aprimorar os programas de melhoramento genético, por meio da seleção assistida por marcadores e da introgressão de genes nas populações.

Inicialmente, o número de marcadores disponíveis era limitado e as estratégias para a incorporação da informação molecular nos programas de melhoramento genético consistiam em estimar previamente o efeito dos marcadores para, então, aplicar a seleção *in tandem* ou a teoria de índices de seleção. Para a estimação dos efeitos dos marcadores foram propostos os métodos dos mínimos quadrados, regressão múltipla, BLUP para o efeito de um único *locus* e regressão de cumeeira.

A seleção assistida por marcadores tinha como grande vantagem a promessa de aumento da acurácia das estimativas de valores genéticos para animais jovens e o encurtamento do intervalo de gerações pela avaliação precoce dos animais, o que resultaria em incremento no ganho genético. Muitos estudos teóricos foram conduzidos demonstrando essas vantagens. Entretanto, a aplicação na prática foi limitada, devido ao reduzido número de marcadores disponíveis que não explicava parcela significativa da variância aditiva das características quantitativas e o elevado custo por unidade genotipada.

O desenvolvimento de tecnologias para o sequenciamento do genoma possibilitou a identificação de polimorfismos nucleotídeo único (SNP – *Single Nucleotide Polimorphism*)

distribuídos ao longo do genoma. Em seguida, a tecnologia evoluiu para a confecção de *chips* que permitem a genotipagem simultânea para milhares de SNP, disponibilizando um volume nunca visto de informações para um único animal. Nesse contexto, a seleção assistida por marcadores em escala ampla passa a ser denominada seleção genômica, na qual o valor genético dos candidatos à seleção é estimado usando apenas a informação molecular.

Na última década, com a drástica redução nos custos por unidade genotipada e um crescimento exponencial do número de SNP genotipados simultaneamente, tem sido possível a aplicação prática da seleção genômica em alguns programas de melhoramento animal, mais especificamente em gado leiteiro. Diversos estudos têm demonstrado a potencialidade da seleção genômica em aprimorar os programas de melhoramento genético, com incrementos no ganho genético pela redução do intervalo de gerações e aumento na acurácia das estimativas de valor genético, além de permitir a seleção para características de baixa herdabilidade, alto custo de mensuração, ligadas ao sexo e medidas tardiamente na vida dos animais. No Brasil, os estudos envolvendo marcadores genéticos focaram principalmente na estimação do efeito individual de cada marcador e na aplicação de *chips* comerciais de baixa densidade. Recentemente, alguns grupos de pesquisa estão genotipando alguns milhares de animais para *chips* de alta densidade, mas os resultados ainda não foram publicados.

Muitos são os desafios para a incorporação da informação molecular nos programas de melhoramento genético animal, dentre eles, a obtenção de milhares de dados de animais avaliados para as mais diversas características de interesse, a existência de amostras de material biológico seguro para a extração de DNA e a definição da metodologia mais adequada para a estimação dos efeitos dos marcadores, bem como a estratégia a ser utilizada para a incorporação desses dados nas avaliações genéticas tradicionais. Em bovinos de corte, os desafios são ainda maiores em função da estrutura da cadeia, das diferenças entre os programas de melhoramento e o baixo valor agregado do produto final que não permite grandes aumentos nos custos de produção, fazendo com que cada programa busque a estratégia mais adequada à realidade na qual está inserido.

2. Revisão bibliográfica

2.1. Panorama da pecuária de corte brasileira

O Brasil já é considerado a sétima maior economia mundial, com o produto interno bruto (PIB) tendo alcançado R$ 4.143 trilhões em 2011, segundo o IBGE (2012a). Desse montante, o agronegócio representou 22,74%, com faturamento bruto de R$ 942 bilhões (CEPEA-USP, 2011), dos quais R$ 94,59 bilhões foram em produtos exportados (AGENCIA BRASIL, 2012). O superávit comercial do agronegócio brasileiro foi de R$ 77,51 bilhões, enquanto o superávit geral do país atingiu somente R$ 29,80 bilhões, evidenciando a importância do agronegócio brasileiro, que tem coberto déficits comerciais de outras áreas da economia (AGENCIA BRASIL, 2012).

A pecuária de corte brasileira é a maior cadeia produtiva do agronegócio brasileiro e ocupa posição de destaque na composição do PIB do país, conforme revelado no relatório do CNA/CEPEA (2012): "A cadeia da bovinocultura se destacou por apresentar a maior renda no período em estudo. Em 2009, ela gerou renda de R$ 77,6 bilhões, a preços reais de 2007. Na sequência, veio a da cana-de-açúcar, com R$ 65,8 bilhões, seguida pelas cadeias da soja (R$ 36,2 bilhões), do leite (R$ 34,5 bilhões) e do algodão (R$ 14 bilhões)".

O rebanho bovino brasileiro é o maior rebanho comercial do mundo, com 209,5 milhões de cabeças (IBGE, 2011), sendo que 80% desse total são constituídos por animais de raças zebuínas ou azebuadas, 90% dos quais são da raça Nelore. Isso perfaz um total aproximado de 150 milhões de cabeças da raça Nelore, uma raça muito bem adaptada ao sistema de produção brasileiro, que é predominantemente a pasto, com suplementação mineral. Dos 28,8 milhões animais abatidos no Brasil em 2011(IBGE, 2012b), apenas cerca de 3,39 milhões de animais (11,77%) foram terminados em confinamento (NOGUEIRA, 2012), o que demonstra a importância dos sistemas de produção a pasto no Brasil.

Atualmente, o país conta com cerca de 70 milhões de vacas e insemina pouco mais de 10 milhões delas, ou seja, pouco mais de 14%. As matrizes restantes são submetidas à monta natural, com o uso de cerca de 2,5 milhões de touros, o que leva a uma demanda de tourinhos de reposição da ordem de 450 a 500 mil animais/ano, se considerarmos a taxa de reposição anual ao redor de 20%. Embora os diversos programas de melhoramento genético da raça Nelore tenham atingido ganhos genéticos muito expressivos, alguns deles ultrapassando os

10% da média das características em ganho nos últimos 15 anos, o número de tourinhos de reposição avaliados geneticamente não ultrapassa 5% da demanda anual.

Esses números demonstram a potencialidade de produção da pecuária brasileira, que devido ao tamanho do rebanho e aos baixos custos de produção, disputa a liderança do mercado internacional de carne bovina há muitos anos. Por outro lado, seus índices de produtividade são muito inferiores aos dos principais concorrentes, Austrália e EUA, indicando a existência de oportunidades para o aumento da eficiência produtiva pela melhoria do potencial genético do rebanho, a qual se dá pela seleção de animais de mérito genético superior à média da população e pela disseminação desse material genético por meio de acasalamentos.

2.2. A informação molecular nos processos de seleção

Tradicionalmente a seleção de animais de elevado potencial genético é realizada com base em dois tipos de dados, pedigree e fenótipo, que são combinados para a estimação da melhor predição linear não viesada do valor genético. Com o desenvolvimento da genética molecular, uma terceira fonte de informação tornou-se disponível, a obtida com base nos marcadores genéticos. Com isso, um grande esforço tem sido feito para a identificação, localização, estimação do efeito e a inclusão dessa informação em programas de melhoramento genético.

Vários tipos de marcadores têm sido usados para a identificação da variabilidade existente no DNA. Os mais comuns são RFLP (*restriction fragment length polymorphism*), RAPD (*random amplification of polymorphism*), AFLP (*amplified fragment length polymorphism*), microssatélites (VNTR – *variable number of tandem repeats*) e, com o desenvolvimento de equipamentos que permitem o sequenciamento do DNA, os polimorfismos nucleotídeo único (SNP – *Single Nucleotide Polymorphism*). A disponibilidade dessa informação molecular surgiu como oportunidade de aprimorar os programas de melhoramento genético, pela seleção direta de genes ou regiões genômicas que afetam as características de interesse econômico por meio da seleção assistida por marcadores e pela introgressão de genes (DEKKERS; HOSPITAL, 2002).

Neimann-Sorensen e Robertson (1961) foram os primeiros a questionar a importância da detecção de genes e de marcadores genéticos para os programas de seleção, fazendo um paralelo com os métodos clássicos de genética quantitativa, baseados apenas nas medidas das características de interesse. Soller (1978) e Soller e Beckmann (1983) estudaram o ganho

genético adicional esperado com a seleção assistida por marcadores, devido ao aumento na acurácia das estimativas de valor genético e redução no intervalo de gerações.

Desde então, diversos estudos têm sido desenvolvidos no sentido de avaliar metodologias para o uso da informação dos marcadores genéticos nos programas de seleção. A primeira proposta para estimação do efeito dos marcadores em características quantitativas foi pelo método dos mínimos quadrados (GELDERMANN, 1975). Em seguida, baseado na teoria de índices de seleção, Soller (1978) sugeriu combinar as informações fenotípicas e de marcadores nas avaliações genéticas. Em 1989, Fernando e Grossman apresentaram uma metodologia para a estimação do efeito de um único marcador sob a teoria de modelos mistos, assumindo efeitos aditivos para os alelos. Posteriormente, esse modelo foi expandido por Goddard (1992) para a análise de diversos marcadores.

A estimação do efeito de muitos marcadores pelo método da regressão múltipla usando um conjunto de dados restrito é uma dificuldade estatística, uma vez que o número de coeficientes de regressão a ser estimado é muito maior que o número de dados disponíveis. Com o intuito de minimizar o problema de dimensionalidade e de superestimação dos efeitos, Lande e Thompson (1990) propuseram utilizar metade dos dados para a seleção dos marcadores de maiores efeitos e a outra metade para a reestimação conjunta dos efeitos dos marcadores selecionados para compor o índice de seleção. Whittaker, Thompson e Dehan (2000) demonstraram que o uso da regressão de cumeeira, como alternativa à pré-seleção dos marcadores, aumentou, em média, a resposta à seleção e reduziu a variabilidade das estimativas dos coeficientes de regressão.

Diversas estratégias e metodologias foram publicadas na literatura para a implantação e avaliação do impacto da seleção assistida por marcadores em programas de melhoramento animal (GODDARD, 1992; VAN ARENDONK; TIER; KINGHORN, 1994; MEUWISSEN; GODDARD, 1996; HALEY; VISSCHER, 1998).

Na última década, o sequenciamento do genoma de diversas espécies de interesse produtivo e o desenvolvimento de *chips* que permitiram a genotipagem para milhares de SNP ao longo do genoma culminaram na seleção assistida por marcadores em escala genômica, denominada como seleção genômica.

A seleção genômica enfatiza a predição simultânea dos efeitos de milhares de marcadores distribuídos ao longo do genoma de forma a explicar toda variação genética aditiva de uma característica quantitativa pela captura dos efeitos de todos os locos que a condicionam. Nesse contexto, vários métodos de predição dos efeitos dos marcadores foram propostos: quadrados mínimos, BLUP, Bayes A e Bayes B (MEUWISSEN; HAYES;

GODDARD, 2001), regressão de cumeeira bayesiana (GIANOLA; PEREZ-ENCISO, TORO, 2003), regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA; FERNANDO; STELLA, 2006), aprendizado de máquina (LONG et al., 2007), regressão *stepwise* (HABIER; FERNANDO; DEKKERS, 2007), regressão RKHS (*Reproducing Kernel Hilbert Spaces*, GIANOLA; VAN KAAM, 2008), LASSO bayesiano (PARK; CASELLA, 2008; DE LOS CAMPOS et al., 2009), Bayes B acelerado (MEUWISSEN et al., 2009), regressão via quadrados mínimos parciais e via componentes principais (SOLBERG et al., 2009) e Bayes C, Bayes Cπ, Bayes D e Bayes Dπ (HABIER et al., 2011).

O efeito de cada marcador estimado pelas metodologias retrocitadas representa a contribuição individual de cada um para o mérito genético do animal. A soma dos efeitos de cada marcador ponderada pelo número de alelos presentes (0, 1 ou 2) em cada *locus*, fornece uma estimativa do valor genético molecular de cada animal. A predição dos efeitos dos marcadores é realizada com base em dados fenotípicos e genotípicos de uma amostra da população e esses efeitos são aplicados para a estimação dos valores genéticos de indivíduos candidatos à seleção, que foram apenas genotipados. Por não ser possível a genotipagem de todos os animais da população, procedimentos *two*, *three* e *single-step* foram propostos para a incorporação dessa informação molecular nos programas de melhoramento (GUILLAUME et al., 2008; VAN RADEN et al., 2009; LEGARRA; AGUILAR; MISZTAL, 2009; WOLC et al., 2011).

Atualmente, existe um consenso muito grande das vantagens da seleção genômica no aumento da acurácia de predição do valor genético, na redução do intervalo de gerações e, por consequência, no aumento no ganho genético. Ainda não está tão evidente qual a melhor metodologia para a estimação dos efeitos dos marcadores ou estratégia para a incorporação dessa informação nos programas de melhoramento animal, e é provável que, talvez, não haja uma metodologia e uma estratégia ideais, mas sim metodologias e estratégias mais adequadas a cada sistema de produção.

3. Hipóteses

A incorporação da informação de marcadores genéticos, juntamente com as informações fenotípicas e de pedigree nas avaliações genéticas, aumenta a acurácia das estimativas de valores genéticos para características produtivas avaliadas em programas de melhoramento genético de bovinos de corte.

4. Objetivos

4.1. Objetivo geral

Avaliar o impacto da incorporação das informações de marcadores genéticos, juntamente com as informações fenotípicas e genealógicas, nas estimativas de valores genéticos para características produtivas em um programa de melhoramento genético de bovinos de corte.

4.2. Objetivos específicos

1. Comparar as capacidades de ajuste e preditiva de modelos de avaliação genética que consideraram ou não os efeitos dos marcadores genéticos na estimação do valor genético dos animais.

2. Estimar os efeitos de substituição alélica dos marcadores genéticos pelos métodos de regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano.

3. Avaliar o impacto das diferentes metodologias de estimação dos efeitos de substituição alélica dos marcadores genéticos na acurácia das estimativas de valores genéticos, comparando-as com as obtidas pelo melhoramento genético clássico.

4. Estudar o impacto da inclusão dos efeitos dos marcadores nos modelos de avaliação genética no *ranking* dos reprodutores e avaliar os possíveis conflitos de seleção, ou seja, comparar a proporção de animais que seriam selecionados pela avaliação genética clássica com a de animais que seriam selecionados pela avaliação genética assistida por marcadores.

CAPITULO I

Avaliação da qualidade de ajuste global e da capacidade preditiva de diferentes modelos de avaliação genética

1. Introdução

A seleção de animais de elevado potencial genético para características de interesse produtivo é, tradicionalmente, realizada por meio dos valores genéticos estimados com base no fenótipo do indivíduo e/ou de seus parentes considerando a proporção de alelos em comum por descendência e a herdabilidade da característica (VAN ARENDONK; TIER; KINGHORN, 1994). Essa tem sido uma abordagem de sucesso pelos programas de seleção de bovinos de corte, mas o progresso genético obtido tem sido relativamente lento para características mensuradas em apenas um dos sexos, em idade adulta ou após o abate, além das características de baixa herdabilidade e de difícil ou alto custo de mensuração (GODDARD; HAYES, 2009).

A inclusão de uma terceira fonte de informação, baseada nos marcadores genéticos, nos programas de seleção foi denominada seleção assistida por marcadores (MAS – *Marker Assisted Selection*). A idéia básica da MAS é explorar a dependência estatística (desequilíbrio de ligação) existente na distribuição conjunta dos marcadores e das regiões cromossômicas associadas às características quantitativas (GIANOLA; PEREZ-ENCISO; TORO, 2003) permitindo aprimorar as predições do mérito genético dos candidatos a seleção (FERNANDO; GROSSMAN, 1989).

O domínio da técnica de sequenciamento total do genoma e o desenvolvimento da tecnologia dos SNP *chips* promoveram um aumento exponencial do número de marcadores disponíveis, bem como o interesse pela MAS aplicada em ampla escala, denominada seleção genômica. Entretanto, os custos para a aplicação dessa tecnologia e a organização dos programas de melhoramento animal fazem com que sua aplicação prática esteja limitada à pecuária de leite, em especial aos animais da raça Holandesa. Adicionalmente, não existe um consenso sobre qual a melhor metodologia ou estratégia para aplicar a seleção genômica. A primeira metodologia proposta para estimar os efeitos dos marcadores foi pela aplicação do método dos mínimos quadrados (GELDERMANN, 1975). Em 1989, Fernando e Grossman estimaram o efeito de um *locus* pela teoria de modelos mistos. A metodologia de regressão de cumeeira foi sugerida para a estimação dos efeitos dos marcadores para superar o problema de

dimensionalidade dos dados e de colinearidade entre efeitos (WHITTAKER; THOMPSON; DEHAN, 2000). Na última década, os métodos bayesianos de encurtamento e os métodos não paramétricos também têm sido aplicados (MEUWISSEN; HAYES; GODDARD, 2001; GIANOLA; FERNANDO; STELLA, 2006; DE LOS CAMPOS et al., 2009; HABIER et al., 2011). Nesse ínterim, diversas estratégias para a incorporação da informação molecular nos programas de melhoramento foram sugeridas, como os índices de seleção, as análises *two* e *single-step* e os métodos não paramétricos (SOLLER, 1978; GONZÁLEZ-RECIO et al., 2008; VAN RADEN, 2008; LEGARRA; AGUILAR; MISZTAL, 2009).

Em animais de origem Zebu, especialmente nos da raça Nelore, a mais representativa no rebanho bovino brasileiro, o número de estudos sobre a aplicação dos marcadores genéticos nos programas de seleção é muito escasso. De fato, os estudos existentes focaram principalmente na avaliação do efeito de polimorfismos individuais nas características quantitativas (FERRAZ et al., 2009; AYRES et al., 2010; PINTO et al., 2010). O objetivo desse estudo foi avaliar a aplicação de diferentes modelos de avaliação genética, que consideraram ou não os efeitos dos marcadores para a estimação dos valores genéticos, comparando-os em termos de alterações no *ranking* dos animais, qualidade de ajuste e habilidade preditiva, quando um número reduzido de animais possui informações fenotípicas, genotípica e genealógica disponível.

2. Material e Métodos

2.1. Dados fenotípicos

Os dados analisados nesse projeto são pertencentes a Agro-Pecuária CFM Ltda. e fazem parte do seu programa de avaliação genética que é realizado pelo Grupo de Melhoramento Animal e Biotecnologia da Faculdade de Zootecnia e Engenharia de Alimentos (GMAB-FZEA).

Foram utilizados registros de animais nascidos entre 1984 e 2009, filhos de touros selecionados para características de desempenho e reprodutivas avaliados a pasto. As características analisadas foram peso na desmama (PD), ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM), as quais compõem o índice de seleção utilizado na seleção dos animais por essa empresa. O PD foi mensurado ao redor dos 205 dias de idade por meio de pesagem em balança eletrônica acopladas ao tronco de

contenção após jejum total dos animais de, no mínimo, 12 horas. O GPD foi calculado pela diferença de pesos observada entre os 550 dias (18 meses) e os 205 dias de idade, ajustada pelo número de dias entre as pesagens e padronizada para 345 dias. A mensuração do PE foi realizada na região do maior diâmetro do escroto por uma fita métrica metálica ao redor dos 550 dias de idade. A avaliação visual do escore de musculosidade foi realizada por um número ímpar de avaliadores treinados, na qual o escore de cada animal é resultante do consenso entre os avaliadores. A escala adotada variou de 1 a 6, sendo a maior nota dada ao animal que melhor expressa o perfil desejável para a característica.

2.2. Obtenção do fenótipo corrigido

Análises unicaracterística foram realizadas pelo método da máxima verossimilhança restrita por meio do conjunto de programas MTDFREML – *Multiple Trait Derivative Free Restricted Maximum Likelihood* (BOLDMAN et al., 1995) para estimar as soluções para os efeitos fixos e aleatórios para as características PD, GPD, PE e EM, sob modelo animal, usando o banco de dados completo disponível para o rebanho em estudo. Os efeitos fixos considerados foram o grupo de contemporâneos composto por fazenda, safra, estação, sexo e grupo de manejo na mensuração, bem como a idade a mensuração como efeito linear e a idade da mãe ao parto como efeito linear e quadrático. Além do efeito aditivo direto, o efeito genético materno e o efeito de ambiente permanente foram assumidos como efeitos aleatórios para o PD e o efeito do grupo de manejo na desmama como efeito aleatório para GPD, PE e EM, conforme proposto por Pedrosa (2011).

Nesse estudo, definiram-se como animais pertencentes a uma mesma safra todos aqueles nascidos até o mês de maio do ano seguinte à estação de monta, a qual é realizada durante os meses de novembro a fevereiro. Foram consideradas como sendo quatro as estações do ano, divididas em trimestres a partir do mês de janeiro. Além disso, para a característica escore de musculosidade, as mensurações anteriores a 1999, realizadas utilizando uma escala de 1 a 9, foram convertidas para a escala atual de 1 a 6. Os dados de animais faltando a identificação da mãe, nascidos de grupos de reprodutor múltiplo e pertencentes a grupos de contemporâneos com menos de cinco animais foram desconsiderados. Após todas essas verificações o banco de dados analisado apresentou a estrutura descrita na Tabela 1.

Tabela 1. Estrutura dos dados analisados para a estimação das soluções dos efeitos fixos e predição das soluções para os aleatórios para as características peso na desmama (PD), ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM).

DESCRIÇÃO	PD	GPD	PE	EM
Nº de pais	1.355	1.326	1.276	1.300
Nº médio de filhos/pai	61,5	51,6	27,7	49,1
Nº máximo de filhos/pai	5.440	3.987	2.027	4067
Nº mínimo de filhos/pai	1	1	1	1
Nº de mães	23.576	22.507	18.130	21.642
Nº médio de filhos/mãe	3,5	3,0	2,0	3,0
Média de idd_mae (meses)	82,7	81,8	82,3	81,7
Nº de gc	2.874	1.046	518	1.111
Nº médio de animais por gc	29,0	65,4	68,3	57,5
Média de idd (dias)	200,8	546,7	542,2	541,5
Nº de fazendas	13	12	12	12
Nº de safras	26	26	25	20

Nº: número; idd_mãe: idade da mãe ao parto; gc: grupo de contemporâneos; idd: idade à mensuração.

As estatísticas descritivas dos dados utilizados para a estimação das soluções dos efeitos fixos e predição das soluções dos efeitos aleatórios estão apresentadas na Tabela 2. A matriz de parentesco correspondente era formada por um total de 116.652 animais.

Tabela 2. Estatísticas descritivas dos dados analisados para a estimação das soluções para os efeitos fixos e predição das soluções dos efeitos aleatórios para as características peso na desmama (PD), ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM).

Característica	N	MÉDIA	DP	CV	MIN	MAX
PD (kg)	83.404	188,93	27,35	14,48	68,00	297,00
GPD (kg)	68.424	114,47	32,25	28,18	1,01	294,54
PE (cm)	35.401	27,36	3,43	12,53	16,00	44,00
EM (unid)	63.854	3,61	1,00	27,84	1,00	6,00

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor fenotípico observado; MAX: máximo valor fenotípico observado.

Para todas as características analisadas, o fenótipo observado de cada indivíduo avaliado na análise unicaracterística foi corrigido para os mesmos efeitos fixos e aleatórios assumidos no modelo usado para a estimação da solução desses efeitos, exceto para o efeito aditivo direto. O fenótipo corrigido representa a soma das porções aditiva e residual do fenótipo observado. O efeito genético aditivo direto estimado foi assumido como sendo o valor genético de referência para os animais genotipados e não genotipados. O fenótipo

corrigido e o valor genético serão usados posteriormente nas análises para a comparação de modelos.

2.3. Dados moleculares

Foram genotipados 3.549 animais desse rebanho, dos quais 3.160 tinham pai e mãe conhecidos e por isso compuseram o banco de dados genotípico analisado no presente projeto. Esses dados consistiram de 377 fêmeas e 2.783 machos. No total, 3.010 mães e 752 pais estavam representados no banco de dados genotípico, dos quais 104 mães e 278 pais também foram genotipados. A maioria dessas mães tinha apenas uma progênie genotipada e 46 pais tinham mais de 10 progênies genotipadas, dos quais 21 eram pais genotipados. As estatísticas descritivas para PD, GPD, PE e EM dos animais genotipados estão apresentadas na Tabela 3.

Tabela 3. Estatísticas descritivas das características peso na desmama (PD), ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM) mensuradas nos animais genotipados.

Característica	N	MÉDIA	DP	CV	MIN	MAX
PD (kg)	3.042	207,05	22,90	11,06	89,00	297,00
GPD (kg)	3.033	125,19	30,50	24,36	10,63	272,86
PE (cm)	2.664	27,18	3,36	12,36	17,00	38,50
EM (unid)	3.149	3,58	1,17	32,68	1,00	6,00

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor fenotípico observado; MAX: máximo valor fenotípico observado.

Os animais foram genotipados para 222 marcadores genéticos do tipo polimorfismo de nucleotídeo único (SNP – *Single Nucleotide Polymorphism*) identificados em raças de origem *Bos taurus*, representados por 123 SNP descritos na literatura como associados com alguma função biológica que afeta a expressão de características produtivas, dos quais 85-90% estão localizados na região de transcrição do gene e de 10-15% estão na região promotora do gene, e 99 SNP são marcadores usados em testes de paternidade, que embora não tenham sido associados com nenhuma função biológica, estão distribuídos ao longo do genoma, podendo por isso estar em desequilíbrio de ligação com genes de interesse. Todo processo de genotipagem foi realizado em laboratórios licenciados pela empresa Merial/Igenity, que é a companhia que detém o direito de uso dos marcadores utilizados nesse projeto.

As frequências alélicas e genotípicas de cada marcador foram estimadas por contagem simples dos diferentes alelos e genótipos por meio do procedimento PROC FREQ do pacote estatístico SAS. Os SNP que apresentaram menor frequência alélica (MAF) abaixo de 5%

foram retirados das análises, com isso, 106 marcadores foram mantidos nas análises seguintes. Essa redução no número de marcadores mantido nas análises comparado com o número de marcadores genotipados deve-se ao fato desses marcadores terem sido descobertos em raças *Bos taurus* e a raça Nelore ser uma raça *Bos indicus*, o que resulta num grande número de marcadores fixados ou com a MAF inferior a 5%.

2.4. Modelos estatísticos

Para os animais genotipados, que representam menos de 8% do total de animais mensurados no rebanho em estudo, três modelos mistos lineares foram propostos para estimar a porção aditiva contida no fenótipo corrigido para cada uma das características analisadas.

Modelo 1: Esse modelo considerou apenas o efeito poligênico e pode ser expresso, em notação matricial, como:

$$\mathbf{y} = \mu + \mathbf{Za} + \mathbf{e}$$

em que **y** é um vetor $n \times 1$ dos fenótipos corrigidos (n = número de observações); μ é a media geral; **a** é um vetor $q \times 1$ dos efeitos genéticos aditivos poligênicos (q = número de animais na matriz de parentesco) com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2_a)$; **Z** é a matriz de incidência dos efeitos genéticos aditivos de ordem $n \times q$; **e** é o vetor dos efeitos residuais, assumidos como sendo independentes e com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$. O efeito aditivo poligênico estimado por esse modelo foi chamado de valor genético dos animais.

Modelo 2: Esse modelo incluiu apenas os efeitos dos marcadores e pode ser descrito como:

$$\mathbf{y} = \mu + \mathbf{Xg} + \mathbf{e}$$

em que **y** é um vetor $n \times 1$ dos fenótipos corrigidos (n = número de observações); μ é a media geral; **g** é um vetor $p \times 1$ do efeito de substituição alélica dos marcadores (p = número de marcadores analisados) com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2_g)$; **X** é a matriz de incidência dos genótipos de ordem $n \times p$, para cujos elementos foi estabelecido um modelo aditivo, com valores 1, 2 ou 3 para aa, Aa e AA, respectivamente; **e** é o vetor dos efeitos residuais, assumidos como sendo independentes e com distribuição $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$. Nesse caso, o valor genético dos animais foi

chamado de valor genético molecular e foi calculado pela soma dos efeitos aditivos de cada marcador.

Modelo 3: Esse modelo incluiu tanto o efeito dos marcadores quanto o efeito poligênico:

$$\mathbf{y} = \mu + \mathbf{Xg} + \mathbf{Za} + \mathbf{e}$$

em que **y** é um vetor $n \times 1$ dos fenótipos corrigidos (n = número de observações); μ é a media geral; **g** é um vetor $p \times 1$ do efeito de substituição alélica dos marcadores (p = número de marcadores analisados) com distribuição N(**0**, $\mathbf{I}\sigma^2_g$); **X** é a matriz de incidência dos genótipos de ordem $n \times p$, para cujos elementos foi estabelecido um modelo aditivo, com valores 1, 2 ou 3 para aa, Aa e AA, respectivamente; **a** é um vetor $q \times 1$ dos efeitos genéticos aditivos poligênicos (q = número de animais na matriz de parentesco) com distribuição N(**0**, $\mathbf{I}\sigma^2_a$); **Z** é a matriz de incidência dos efeitos genéticos aditivos de ordem $n \times q$; **e** é o vetor dos efeitos residuais, assumidos como sendo independentes e com distribuição N(**0**, $\mathbf{I}\sigma^2_e$). O valor genético estimado por esse modelo foi chamado valor genético assistido por marcadores e consiste na soma do efeito aditivo poligênico com a soma dos efeitos aditivos de cada marcador.

O número de observações utilizadas corresponde ao número de animais genotipados para cada característica analisada, conforme apresentado na Tabela 3. A matriz de parentesco utilizada nos modelos 1 e 3 foi a mesma descrita anteriormente, composta por 116.652 animais.

2.5. Análises estatísticas

A inferência Bayesiana pelo método Monte Carlo usando as cadeias de Markov (MCMC) foi aplicada na análise dos dados assumindo distribuição normal para os dados:

$$\mathbf{y}|\boldsymbol{\theta}_j \sim \mathrm{N}\,(\mu_j, \mathbf{I}\sigma^2_e\,)$$

em que o subscrito j = 1, 2 e 3, corresponde ao modelo 1, modelo 2 e modelo 3, respectivamente; $\boldsymbol{\theta}_j$ são os parâmetros desconhecidos para cada modelo; σ^2_e é a variância residual. Note que $\mu_1=\mu+\mathbf{Za}$, $\mu_2=\mu+\mathbf{Xg}$ e $\mu_3=\mu+\mathbf{Xg}+\mathbf{Za}$.

Em todos os modelos, uma priori uniforme foi usada para μ e, para os efeitos residuais assumiu-se distribuição N(**0**, $\mathbf{I}\sigma_e^2$) com σ_e^2 apresentando distribuição χ^{-2} (ν, s^2) com ν = -2 e s^2 = 0. Nos modelos 1 e 3, o efeito genético aditivo foi assumido como tendo distribuição N(**0**, $\mathbf{A}\sigma_a^2$), em que **A** é a matriz de parentesco e σ_a^2 é a variância aditiva com distribuição χ^{-2} (ν, s^2), em que ν = -2 e s^2 = 0. Duas priores foram assumidas para os efeitos de substituição alélica dos marcadores nos modelo 2 e 3. A primeira priori empregada foi uma distribuição uniforme, que consiste na mesma análise de regressão descrita por Meuwissen, Hayes e Goddard (2001), na qual os efeitos dos marcadores são estimados pelo método dos mínimos quadrados. A segunda priori assumida considerou que os efeitos dos marcadores apresentavam uma distribuição N(**0**, $\mathbf{I}\sigma_g^2$), em que σ_g^2 é a variância aditiva dos marcadores, considerada como tendo distribuição χ^{-2} (ν, s^2), com ν = -2 e s^2 = 0. Essa abordagem, chamada de regressão de cumeeira bayesiana (*Bayesian ridge regression*), foi proposta por Gianola, Perez-Enciso e Toro (2003) e, é similar a regressão de cumeeira frequentista proposta por Whittaker, Thompson e Dehan (2000). Deste ponto em diante, neste trabalho, os modelos 2 e 3 foram identificados de acordo com a distribuição a priori assumida para os efeitos dos marcadores, 2_U e 3_U para a priori uniforme e como 2_N e 3_N para a priori normal. Com isso, para a comparação de modelos foram considerados como havendo cinco diferentes modelos. A distribuição condicional completa a posteriori para todos os modelos são distribuições normais para os parâmetros μ, **g** e **a** e distribuição Qui-quadrado escalonada invertida para os parâmetros σ_a^2, σ_g^2 e σ_e^2.

Análises unicaracterística foram realizadas utilizando o *software* Threshold Model - TM (LEGARRA; VARONA; MATURANA, 2008) para estimar os efeitos de substituição alélica dos marcadores e os efeitos genéticos aditivos poligênicos para as características PD, GPD, PE e EM. O *software* TM é um programa escrito em Fortran 90 para a estimação de componentes de variância, valores genéticos e efeitos fixos em modelos de limiar, lineares e censurados lineares para múltiplas características em programas de melhoramento genético. No presente estudo o *software* TM foi modificado para incluir os efeitos de substituição alélica dos marcadores.

Com base na teoria das cadeias de Markov, a aplicação do algoritmo de Gibbs permite gerar amostras da distribuição posteriori conjunta de duas ou mais variáveis aleatórias. Por ser um processo iterativo de amostragem, as amostras iniciais são altamente influenciadas pelos valores iniciais e são necessários alguns ciclos para atingir a distribuição posteriori conjunta, da qual as amostras serão usadas para inferência. Além disso, as amostras sequenciais são

correlacionadas e pouco informativas. Nesse contexto, duas práticas adotadas são o descarte das amostras iniciais, chamado de período de aquecimento (*burn in*), e salvar as amostras em intervalos que minimizem a autocorrelação entre as amostras sucessivas, chamado de *thinning*. Para a obtenção das cadeias de Markov dos parâmetros desconhecidos nos modelos 1, 2_U, 2_N, 3_U e 3_N foi gerada uma cadeia única de 1.000.000 iterações usando o amostrador de Gibbs com *burn in* de 80.000 iterações e *thinning* de 100 amostras. Portanto, 10.000 amostras compuseram as cadeias usadas para inferência.

A convergência das cadeias de Markov foi testada separadamente para cada um dos parâmetros desconhecidos usando o programa de diagnóstico de convergência BOA (*Bayesian Output Analysis*) do pacote estatístico R. Foi realizada uma análise gráfica das cadeias, bem como foram avaliados o critério Z de Geweke (1992) e o erro de amostragem de Monte Carlo calculado pelo procedimento de séries-temporais descrito por Geyer (1992).

2.6. Comparação de modelos

O *ranking* dos animais baseado no valor genético de referência foi usado para avaliar as mudanças no *ranking* quando o valor genético estimado pelos modelos 1, 2_U, 2_N, 3_U e 3_N foi usado para classificar os animais.

A associação entre o *ranking* dos animais baseado no valor genético de referência e os *rankings* baseados no valor genético poligênico estimado no modelo 1 para todos os animais da matriz de parentesco, no valor genético molecular estimado nos modelos 2_U e 2_N apenas para os animais genotipados e no valor genético assistido por marcadores estimado nos modelos 3_U e 3_N apenas para os animais genotipados, foi calculada usando duas medidas de correlação de postos por meio do procedimento PROC CORR do pacote estatístico SAS.

Coeficiente de correlação de postos de Spearman's (ρ_s). O coeficiente de correlação de postos de Spearman é obtido pela ordenação dos dados e utilização dos postos na equação de correlação produto-momento de Pearson, demonstrada a seguir. No caso de empate, utiliza-se a média dos postos.

$$\rho_s = \frac{\sum_i((R_i - \bar{R})(S_i - \bar{S}))}{\sqrt{\sum_i(R_i - \bar{R})^2 \sum(S_i - \bar{S})^2}}$$

em que R_i é o posto de cada observação na classificação pelo valor genético de referência; S_i é o posto de cada observação pelos valores genéticos fornecidos pelos modelos 1, 2_U, 2_N, 3_U e 3_N; $\overline{R}$ é a média dos valores de R_i; $\overline{S}$ é a média dos valores de S_i.

Coeficiente de correlação de postos de Kendall (τ_b). Os dados são duplamente ordenados, primeiro com base nos valores genéticos "verdadeiros" e, em seguida, com base nos valores genéticos obtidos nos modelos 1, 2_U, 2_N, 3_U e 3_N. O coeficiente de correlação de postos de Kendall (KENDALL, 1947) é calculado com base no número de pares de observações concordantes e discordantes, como:

$$\tau_b = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

em que n_c é o número de pares concordantes; n_d é o número de pares discordantes; n é o número total de animais ordenados.

Os modelos 1, 2_U, 2_N, 3_U e 3_N também foram comparados em termos da qualidade de ajuste global e da capacidade preditiva pelos métodos apresentados a seguir.

Critério da informação da "deviance" (DIC – *Deviance Information Criteria*). O DIC compara a qualidade de ajuste global de dois ou mais modelos, considerando a capacidade de ajuste dos dados ao modelo e a complexidade de cada um (SPIEGELHALTER et al., 2002). Para um modelo M, o DIC é definido como:

$$DIC = 2\overline{D} - D(\overline{\theta}_M)$$

em que $\overline{D} = -2\int[\log p(y|\theta_M)]p(\theta_M|\mathbf{y}, M)d\theta_M = E_{\theta_M|y}[D(\theta_M)]$ é a esperança posterior da deviance $D(\theta_M)$, e $D(\overline{\theta}_M) = -2\log p(y|\overline{\theta}_M)$ é a deviance dada pela média posterior do vetor de parâmetros θ_M. A expressão do DIC é resultante da combinação de ambos os termos, onde $\overline{D}$ é uma medida de ajuste e $\overline{D} - D(\overline{\theta}_M)$ está relacionado ao número efetivo de parâmetros (p_d). Esse critério é uma generalização do AIC (*Akaike information criterion*) e do BIC (*Bayesian information criterion*) muito útil quando as distribuições posteriores dos modelos bayesianos são obtidas pelo método Monte Carlo usando cadeias de Markov (MCMC). Os modelos que apresentam os menores valores de DIC exibem um melhor ajuste global considerando sua complexidade. Diferenças nos valores de DIC maiores que 7 são

consideradas importantes, de acordo com Spiegelhalter et al. (2002). Originalmente, o cálculo do DIC não é realizado pelo *software* TM, tendo sido implementado no presente estudo.

Validação cruzada k-vezes (*k-fold*)**.** A validação cruzada k-vezes é uma abordagem usada para avaliar os modelos quanto sua capacidade preditiva. O conjunto de dados é dividido em um subconjunto de treinamento usado para a estimação dos parâmetros dos modelos e um subconjunto de validação no qual se avalia a capacidade preditiva dos modelos. Duas estratégias distintas de validação cruzada foram empregadas: 1) validação cruzada 1-vez (1-*fold*), o subconjunto de treinamento foi composto pelos animais mais velhos e o subconjunto de validação pelos animais mais jovens, nascidos na safra de 2009, os quais representaram cerca de 25% dos animais genotipados. Essa estratégia de divisão temporal dos dados tem forte relação com programas de melhoramento, no qual as informações fenotípicas dos antecessores e as relações de parentesco são usadas para a estimação dos valores genéticos dos animais mais jovens; 2) validação cruzada 4-vezes (4-*fold*), na qual o conjunto de dados foi dividido aleatoriamente em quatro subconjuntos distintos, cada um contendo ao redor de ¼ dos dados. Para a validação cruzada, três dos quatro subconjuntos foram utilizados para a estimação dos parâmetros dos modelos (subconjunto de treinamento) e o subconjunto restante foi usado para avaliar a capacidade preditiva dos modelos (subconjunto de validação). Esse procedimento foi realizado quatro vezes, mudando o subconjunto usado para avaliar a capacidade preditiva dos modelos. Com essa estratégia, atemporal, buscou-se avaliar a capacidade preditiva dos modelos quando um quarto dos animais, definido aleatoriamente, não foi mensurado.

Para avaliar a capacidade preditiva dos modelos, dois critérios foram adotados:

Erro quadrático médio. O erro quadrático médio (MSE – *Mean Squared Error*) foi calculado como:

$$\mathrm{EQM} = \frac{1}{\mathrm{n_{data}}} \sum_{\mathrm{i=1}}^{\mathrm{n_{data}}} (\mathrm{y} - \hat{\mathrm{y}})^2$$

em que y e $\hat{y}$ correspondem aos fenótipos corrigidos observados e preditos, respectivamente; n_{data} é o número de observações no subconjunto de validação. Os modelos que apresentaram os menores valores de EQM foram considerados como aqueles com a melhor capacidade preditiva.

Correlação de Pearson. A correlação de Pearson (ρ) entre os valores observados e preditos foi calculada por:

$$\rho_{y,\hat{y}} = \frac{1}{n_{data}} \sum_{i=1}^{n_{data}} \frac{cov(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}}$$

em que $cov(y,\hat{y})$ é a covariância entre os fenótipos corrigidos observados e preditos; σ_y e $\sigma_{\hat{y}}$ são as estimativas de desvio-padrão dos fenótipos corrigidos observados e preditos; n_{data} é o número de observações no subconjunto de validação. O modelo que apresentou a maior correlação foi considerado como aquele com a melhor capacidade preditiva.

3. Resultados e discussão

3.1. Análise de convergência das cadeias de Markov

Para avaliar a convergência das cadeias, inicialmente analisou-se a variação das cadeias no espaço paramétrico. Essa avaliação foi feita com base no gráfico de traços no qual são plotados os valores das amostras salvas a cada iteração e no gráfico de médias correntes, no qual a média da distribuição posterior é calculada a cada amostra salva. Na Figura 1 estão apresentados os gráficos de traços dos valores das amostras salvas para todos os parâmetros desconhecidos para o escore de musculosidade pelo modelo 3_N e na Figura 2 estão apresentados os gráficos de médias correntes das amostras salvas para todos os parâmetros desconhecidos para o escore de musculosidade pelo modelo 3_N.

Os gráficos de traços (Figura 1) indicaram não ciclicidade das cadeias e a estabilização dos valores das amostras após o período de aquecimento, que foi confirmada pelos gráficos de médias correntes (Figura 2). Sendo assim, o comportamento das cadeias ao longo do processo iterativo permitiu concluir que as cadeias convergiram e apresentaram uma boa mistura.

Em ambos os gráficos foi possível verificar a alta dependência das primeiras amostras em relação aos valores iniciais. A análise dos gráficos de médias correntes confirmou que o período de aquecimento de 80.000 iterações, ou seja, as 800 primeiras iterações salvas, foi suficiente para romper a relação dos valores amostrados com os valores iniciais. Desse modo, todas as inferências foram realizadas com as 10.000 amostras salvas após o período de aquecimento.

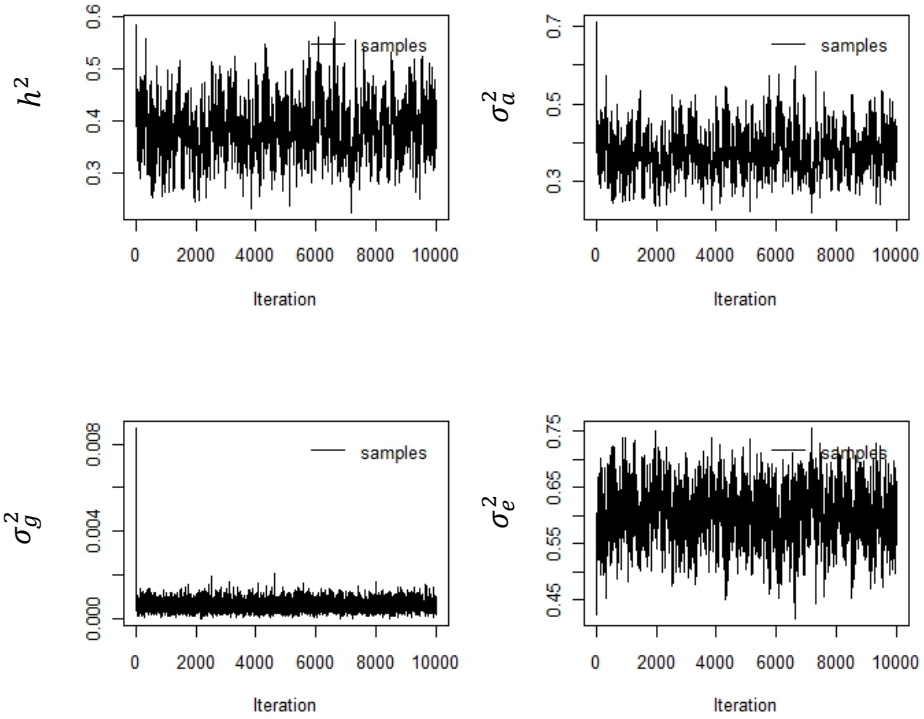


Figura 1. Gráficos de traços dos valores das amostras de herdabilidade (h^2), variância aditiva (σ_a^2), variância dos marcadores (σ_g^2) e variância residual (σ_e^2) para o escore de musculosidade pelo modelo 3_N.

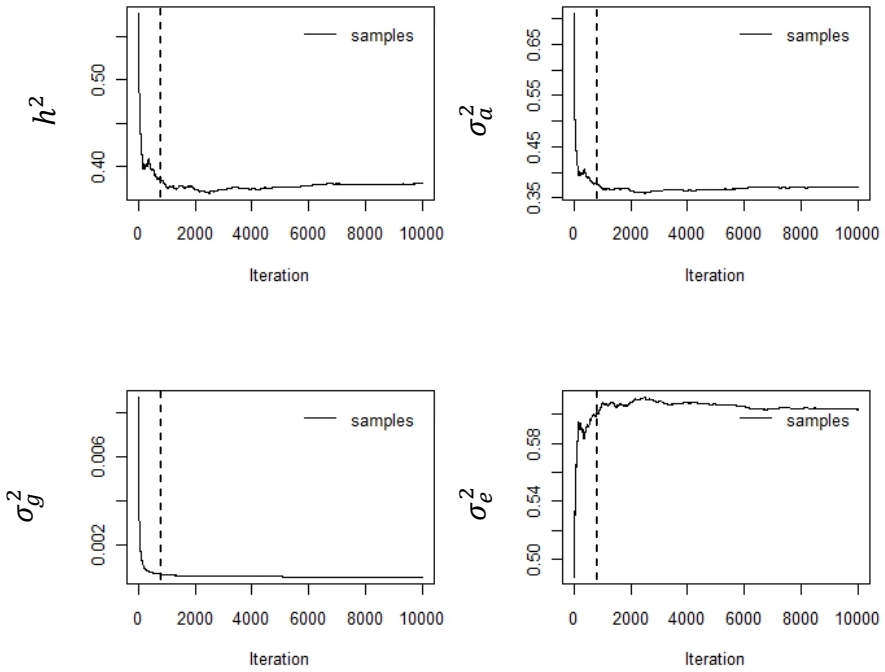


Figura 2. Gráficos de médias correntes dos valores das amostras de herdabilidade (h^2), variância aditiva (σ_a^2), variância dos marcadores (σ_g^2) e variância residual (σ_e^2) para o escore de musculosidade pelo modelo 3_N, com *burn in* delimitado pela linha pontilhada.

O critério Z de Geweke (1992) avalia a média de duas partes não sobrepostas das cadeias, os 10% inicial e os 50% final, e testa se essas duas partes pertencem à mesma distribuição. No presente estudo todas as cadeias apresentaram valores de Z inferiores a 1,96, o que indica convergência (BLASCO; PILES; VARONA, 2003) e estacionariedade das cadeias geradas (GEWEKE, 1992). Outro indicativo da convergência das cadeias também avaliado foi o erro de amostragem de Monte Carlo calculado pelo procedimento de séries-temporais descrito por Geyer (1992), cujos valores foram pelo menos 20 vezes menores que o desvio padrão das distribuições posteriores. Essa baixa magnitude do erro de Monte Carlo também sugere convergência da cadeia, de acordo com Blasco, Piles e Varona (2003).

3.2. Medidas de correlação de postos (ranking)

A superioridade genética de um animal não é dada pela magnitude do seu valor genético aditivo isoladamente, mas sim pela superioridade do seu valor genético em comparação aos valores genéticos de outros animais da mesma avaliação genética. Isso porque, as magnitudes dos valores genéticos são afetadas pela base genética assumida e pela variabilidade e estrutura dos dados utilizados em cada análise. Sendo assim, o *ranking* dos animais baseado nos valores genéticos aditivos é uma importante estratégia para a seleção de animais geneticamente superiores em programas de melhoramento genético.

Nas Figuras de 3 a 6 estão apresentados os gráficos dos coeficientes de correlação de Spearman e de Kendall calculados para comparar o *ranking* dos animais com base no valor genético de referência e os *rankings* baseados no valor genético poligênico estimado no modelo 1 para todos os animais da matriz de parentesco, no valor genético molecular estimado nos modelos 2_U e 2_N apenas para os animais genotipados e no valor genético assistido por marcadores estimado nos modelos 3_U e 3_N apenas para os animais genotipados, para as características PD, GPD, PE e EM, respectivamente.

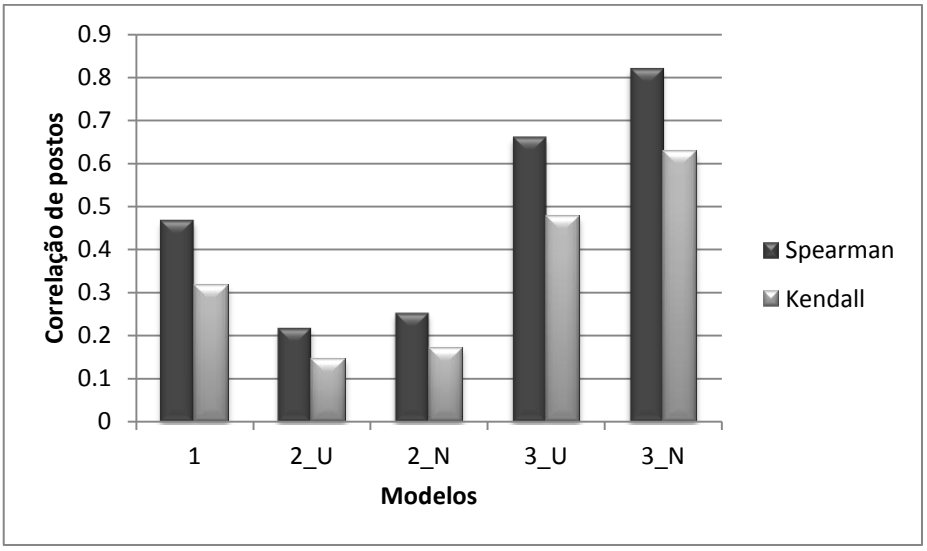


Figura 3. Coeficientes de correlação de postos de Spearman e de Kendall entre o *ranking* dos animais com base no valor genético de referência e o *ranking* fornecido pelos valores genéticos estimados pelos modelos 1, 2_U, 2_N, 3_U e 3_N para o peso na desmama.

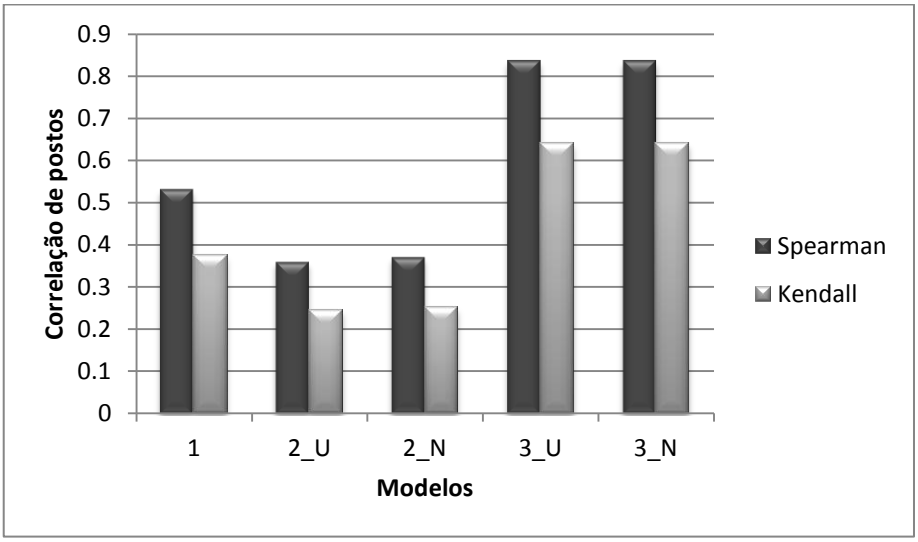


Figura 4. Coeficientes de correlação de postos de Spearman e de Kendall entre o *ranking* dos animais com base no valor genético de referência e o *ranking* fornecido pelos valores genéticos estimados pelos modelos 1, 2_U, 2_N, 3_U e 3_N para o ganho de peso pós-desmama.

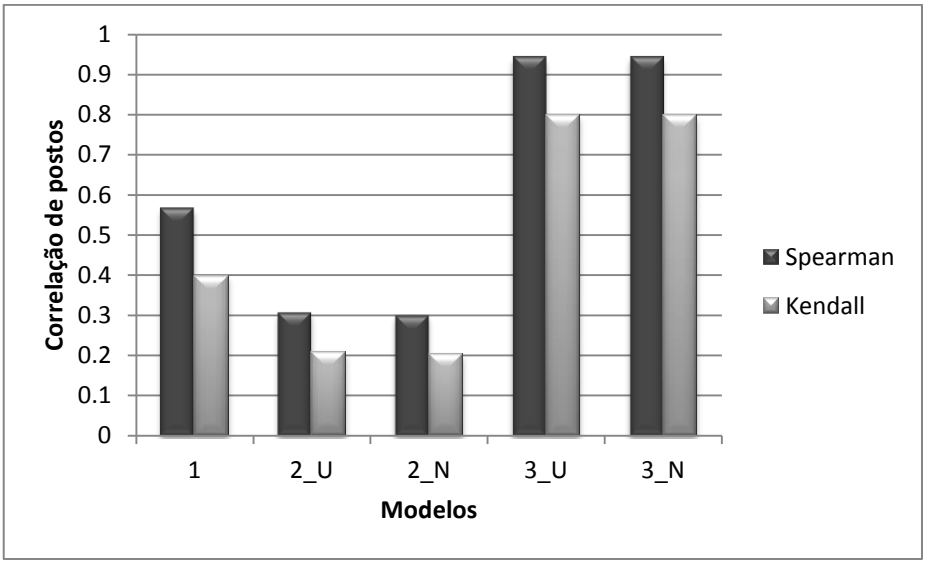


Figura 5. Coeficientes de correlação de postos de Spearman e de Kendall entre o *ranking* dos animais com base no valor genético de referência e o *ranking* fornecido pelos valores genéticos estimados pelos modelos 1, 2_U, 2_N, 3_U e 3_N para o perímetro escrotal.

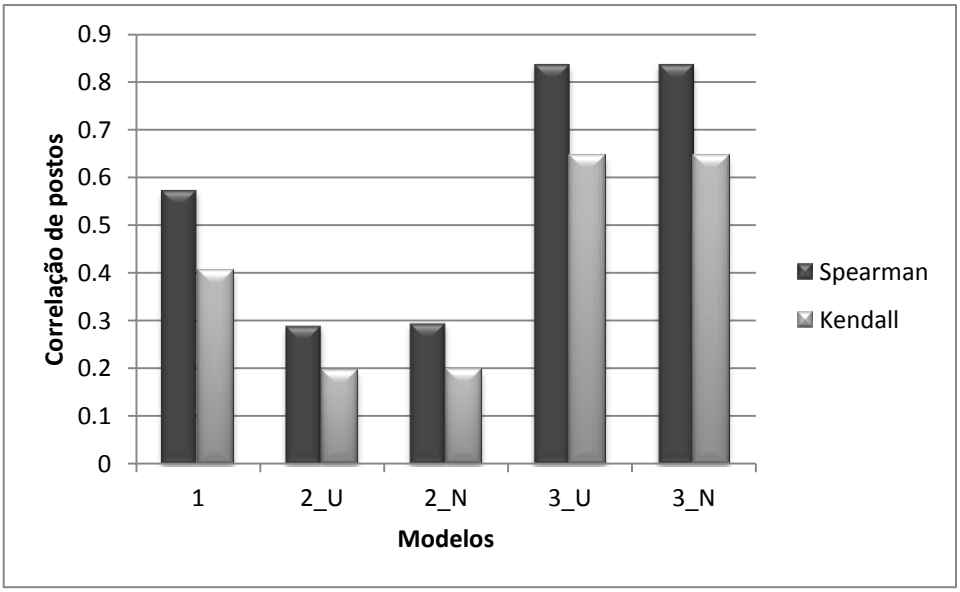


Figura 6. Coeficientes de correlação de postos de Spearman e de Kendall entre o *ranking* dos animais com base no valor genético de referência e o *ranking* fornecido pelos valores genéticos estimados pelos modelos 1, 2_U, 2_N, 3_U e 3_N para o escore de musculosidade.

A variação dos coeficientes de correlação de postos exibidos nas figuras acima seguiu um mesmo padrão para todas as características analisadas. Os valores estimados para o coeficiente de Kendall (τ_b) foram menores que os valores estimados para o coeficiente de

Spearman (ρ_s), o que está de acordo com Kendall (1947) que descreveu que para um grande número de observações, os valores de τ_b são aproximadamente dois terços dos valores de ρ_s. Embora τ_b seja uma medida de correlação de postos mais conservadora, a interpretação de ambas as medidas conduziram às mesmas inferências. As maiores divergências no *ranking* dos animais foram observadas pelos modelos 2_U e 2_N, uma vez que apresentaram os menores valores de correlação com o *ranking* baseado nos valores genéticos de referência. Essa divergência pode ser consequência do pequeno conjunto de marcadores analisados, que não foram suficientes para explicar todo efeito aditivo contido no fenótipo corrigido. Por outro lado, o modelo 1 apresentou valores intermediários de coeficientes de correlação de postos, sugerindo uma maior habilidade desse modelo em reter a porção aditiva do fenótipo corrigido em comparação aos modelos 2_U e 2_N. Essa melhor performance do modelo 1 era esperada ao se considerar a densa matriz de parentesco utilizada. Além disso, a inclusão dos efeitos dos marcadores e poligênico conjuntamente nos modelos 3_U e 3_N levou a uma maior semelhança entre o *ranking* dos animais com base no valor genético assistido por marcadores estimado por esses modelos e o *ranking* usando o valor genético de referência. A combinação de ambos os efeitos melhorou a eficiência desses modelos para captar a porção aditiva contida no fenótipo corrigido, quando um reduzido conjunto de dados foi utilizado. A priori assumida para o efeito dos marcadores afetou ligeiramente as estimativas dos efeitos genéticos aditivos diretos pelos métodos de mínimos quadrados e regressão de cumeeira, conforme observado pelas diferenças entre as estimativas de correlação entre os modelos 2_U e 2_N e 3_U e 3_N. Os maiores valores de coeficientes de correlação de *ranking* observados para os modelos que assumiram distribuições a priori normais para os efeitos dos marcadores podem ser atribuídos à habilidade desses em lidar com a existência de efeitos colineares (WHITTAKER; THOMPSON; DEHAN, 2000).

3.4. Comparação de modelos

As médias das distribuições posteriores e os intervalos de maior densidade a posteriori (HPD95% – *highest 95% posterior density interval*) das variâncias aditiva, dos marcadores e residual e dos coeficientes de herdabilidade para os modelos 1, 2_U, 2_N, 3_U e 3_N estão apresentadas na Tabela 4. As estimativas das médias a posteriori dos coeficientes de herdabilidade pelos modelos 1, 3_U e 3_N para PD (0,16; 0.16; 0,16), GPD (0,33; 0,30; 0,32), PE (0,52; 0,51; 0,52) e EM (0,39; 0,36; 0,38) estão de acordo com as obtidas nas análises unicaracterística realizadas por meio do *software* MTDFREML utilizando o banco de dados

completo, que foram 0,19, 0,21, 0,50 e 0,21, respectivamente. Além disso, as estimativas de herdabilidade apresentadas foram condizentes com as observadas na última avaliação genética realizada pelo GMAB-FZEA para esse rebanho, 0,22, 0,32, 0,45 e 0,25 para PD, GPD, PE e EM, respectivamente (comunicação pessoal). Horimoto et al. (2007) e Van Melis et al. (2010) analisando outras amostras dessa população relataram estimativas de herdabilidade para PD (0,28 e 0,55) fora dos HPD95% apresentados para essa característica pelos modelos 1, 3_U e 3_N. Esses autores descreveram estimativas para GPD (0,32 e 0,25) e PE (0,55 e 0,48) no mesmo intervalo das obtidas pelo presente estudo. Adicionalmente, os coeficientes de herdabilidade estimados por ambos os autores para EM (0,16 e 0,23) foram inferiores aos HPD95% descritos para essa característica pelos modelos 1, 3_U e 3_N. Como os HPD95% para os coeficientes de herdabilidade estimados pelos modelos 1, 3_U e 3_N estão sobrepostos, essas estimativas não podem ser consideradas diferentes.

Comparando as estimativas das variâncias residuais apresentadas pelos modelos 1, 2_U, 2_N, 3_U e 3_N nota-se que o modelo 3_N apresentou as menores estimativas para todas as características estudadas, com pequenas diferenças em relação aos modelos 1 e 3_U. Isso indica que os modelos 1, 3_U e 3_N possuem maior capacidade de explicação da variação dos dados. Por outro lado, os modelos 2_U e 2_N apresentaram a pior performance, dado os altos valores estimados para a variância residual, sugerindo que a inclusão apenas dos efeitos dos marcadores no modelo não é suficiente para a retenção de toda porção aditiva da variância. A priori assumida para os efeitos dos marcadores não proporcionou diferenças nas estimativas obtidas pelos modelos 2_U e 2_N, o que pode ser atribuído a não existência do problema de dimensionalidade para a estimação dos efeitos dos marcadores, já que o número de marcadores é menor que o número de observações analisadas. A questão da dimensionalidade foi discutida por Meuwissen, Hayes e Goddard (2001) como uma das maiores limitações ao uso do método de mínimos quadrados, uma vez que proporciona estimativas viesadas dos efeitos dos marcadores. Embora esse problema dimensionalidade não exista nos dados analisados pelo presente estudo, o uso da distribuição a priori uniforme para o efeito dos marcadores no modelo 3_U conduziu a ligeiro confundimento na estimação dos efeitos aditivos poligênicos e dos marcadores, conforme observado ao serem comparadas as variâncias aditivas e residuais obtidas para os modelos 1 e 3_U. Esperava-se que a inclusão dos efeitos dos marcadores no modelo 3_U reduziria a variância residual em relação ao modelo 1 sem alterar a estimativa da variância aditiva, mas houve tênue redução do valor estimado para as características GPD, PE e EM.

Tabela 4. Estimativas das médias a posteriori (na primeira linha) e os intervalos de maior densidade a posteriori – HPD95% (na segunda linha) para as variâncias aditiva ($\hat{\sigma}_a^2$), dos marcadores ($\hat{\sigma}_g^2$) e residual ($\hat{\sigma}_e^2$) e os coeficientes de herdabilidade (h^2) para o peso na desmama (PD), ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM) para os modelos 1, 2_U, 2_N, 3_U e 3_N.

Modelo	Parâmetro	PD	GPD	PE	EM
1	$\hat{\sigma}_a^2$	25,47 (15,10; 37,28)	127,01 (92,80; 160,11)	3,66 (2,85; 4,43)	0,38 (0,28; 0,50)
	$\hat{\sigma}_e^2$	133,21 (121,33; 144,11)	259,34 (229,36; 289,12)	3,31 (2,69; 3,93)	0,61 (0,51; 0,69)
	h^2	0,16 (0,09; 0,23)	0,33 (0,25; 0,41)	0,52 (0,43; 0,62)	0,39 (0,29; 0,48)
2_U	$\hat{\sigma}_e^2$	156,24 (148,53; 164,46)	367,53 (349,20; 386,36)	6,47 (6,12; 6,83)	0,93 (0,89; 0,98)
2_N	$\hat{\sigma}_g^2$	0,02 (0,00; 0,05)	0,50 (0,25; 0,76)	0,01 (0,005; 0,02)	0,001 (0,0004; 0,002)
	$\hat{\sigma}_e^2$	156,24 (148,22; 164,06)	367,44 (348,73; 386,07)	6,50 (6,16; 6,88)	0,93 (0,89; 0,98)
3_U	$\hat{\sigma}_a^2$	25,35 (14,67; 36,82)	112,41 (79,13; 148,18)	3,48 (2,72; 4,33)	0,36 (0,25; 0,46)
	$\hat{\sigma}_e^2$	133,97 (122,16; 145,04)	264,68 (233,21; 294,66)	3,31 (2,67; 3,94)	0,62 (0,52; 0,71)
	h^2	0,16 (0,09; 0,23)	0,30 (0,21; 0,38)	0,51 (0,41; 0,61)	0,36 (0,26; 0,47)
3_N	$\hat{\sigma}_a^2$	25,50 (15,22; 35,46)	120,98 (87,21; 153,11)	3,55 (2,78; 4,40)	0,37 (0,27; 0,48)
	$\hat{\sigma}_g^2$	0,007 (0,00; 0,03)	0,24 (0,07; 0,44)	0,005 (0,00; 0,01)	0,0006 (0,00; 0,001)
	$\hat{\sigma}_e^2$	133,09 (122,03; 143,62)	257,54 (228,98, 287,30)	3,28 (2,66; 3,91)	0,60 (0,51; 0,69)
	h^2	0,16 (0,10; 0,22)	0,32 (0,24; 0,40)	0,52 (0,42; 0,62)	0,38 (0,28; 0,48)

Na Tabela 5 estão apresentadas as estimativas para o critério da informação da "deviance" (DIC), a "deviance" ($\overline{D}$) e o número efetivo de parâmetros (p_d) para os modelos analisados. O modelo 3_N foi o que apresentou melhor ajuste global para todas as características analisadas, como indicado pelos menores valores de DIC, exceto para PD, em que o modelo 1 apresentou o menor valor. Entretanto, diferenças entre os valores de DIC menores que 7 não são diferenças importantes, segundo Spiegelhalter et al. (2002), sendo assim, os modelos 1 e 3_N apresentaram a mesma qualidade de ajuste para PD.

Tabela 5. Estimativas de Monte Carlo para critério da informação da "deviance" (DIC), a "deviance" ($\bar{D}$) e o número efetivo de parâmetros (p_d) para os modelos 1, 2_U, 2_N, 3_U e 3_N.

Modelo	Parâmetro	PD	GPD	PE	EM
1	DIC	23.934,89	26.316,04	11.965,48	8.408,07
	$\bar{D}$	23.507,92	25.457,25	10.738,52	7.359,02
	p_d	426,97	858,80	1226,96	1.049,04
2_U	DIC	24.105,70	26.626,99	12.640,97	8.820,40
	$\bar{D}$	23.997,42	26.518,86	12.532,96	8.712,36
	p_d	108,28	108,12	108,01	108,05
2_N	DIC	24.005,91	26.573,64	12.598,50	8.766,28
	$\bar{D}$	23.995,46	26.518,48	12.544,31	8.716,90
	p_d	10,45	55,16	54,20	49,38
3_U	DIC	24.037,92	26.377,91	11.984,66	8.455,03
	$\bar{D}$	23.526,44	25.521,65	10.731,66	7.395,34
	p_d	511,47	856,26	1253,00	1.059,69
3_N	DIC	23.936,58	26.297,91	11.941,03	8.390,72
	$\bar{D}$	23.505,37	25.437,93	10.709,36	7.337,42
	p_d	431,21	859,98	1231,67	1.053,30

Duas estratégias de validação cruzada, 1-*fold* e 4-*fold*, e dois critérios, erro quadrático médio (EQM) e correlação de Pearson, foram adotados para a avaliação da capacidade preditiva dos modelos 1, 2_U, 2_N, 3_U e 3_N. Nas Tabelas 6 e 7 estão demonstrados os resultados de EQM e da correlação de Pearson para a validação cruzada 1-*fold*, nessa ordem. Os menores valores de EQM foram exibidos pelo modelo 1 para PD e GPD e pelo modelo 3_N para PE e EM, sugerindo uma melhor habilidade preditiva desses modelos para essas características. As estimativas das correlações de Pearson para os modelos 1 e 3_N foram as iguais para PD e GPD e, para PE e EM, o modelo 3_N apresentou estimativas ligeiramente superiores as do modelo 1.

Tabela 6. Estimativas do erro quadrático médio dos modelos 1, 2_U, 2_N, 3_U e 3_N, pela validação cruzada 1-*fold*.

Característica	1	2_U	2_N	3_U	3_N
PD	148,50	158,13	149,33	157,76	148,71
GPD	362,05	390,62	369,14	390,13	363,85
PE	6,06	6,72	6,35	6,36	6,03
EM	1,18	1,23	1,20	1,20	1,17

PD: peso na desmama; GPD: ganho de peso pós-desmama; PE: perímetro escrotal; EM: escore de musculosidade.

Tabela 7. Estimativas da correlação de Pearson dos modelos 1, 2_U, 2_N, 3_U e 3_N, pela validação cruzada 1-*fold*.

Característica	1	2_U	2_N	3_U	3_N
PD	0,06	0,00	-0,01	0,02	0,06
GPD	0,11	0,04	0,06	0,07	0,11
PE	0,22	0,09	0,10	0,20	0,24
EM	0,14	0,08	0,09	0,13	0,15

PD: peso na desmama; GPD: ganho de peso pós-desmama; PE: perímetro escrotal; EM: escore de musculosidade.

Os resultados de EQM e da correlação de Pearson para a validação cruzada 4-*fold*, estão apresentados nas Tabelas 8 e 9, respectivamente. Esses resultados corroboram com os observados para a validação cruzada 1-*fold*, na qual os modelos com melhores habilidades preditivas foram os modelos 1 e 3_N, com pequenas diferenças entre eles. Para PD, o modelo 1 exibiu a menor estimativa de EQM e o maior valor de correlação de Pearson, portanto, esse modelo possui uma melhor habilidade de predição para essa característica. O modelo 3_N teve melhor performance para GPD, PE e EM, conforme indicado pelas baixas estimativas de EQM e altos valores de correlação de Pearson.

Tabela 8. Estimativas do erro quadrático médio dos modelos 1, 2_U, 2_N, 3_U e 3_N, para os subconjuntos 1, 2, 3 e 4, pela validação cruzada 4-*fold*.

Característica	Subconjunto	1	2_U	2_N	3_U	3_N
PD	1	152,58	162,39	155,22	159,27	152,57
	2	147,51	155,77	148,61	155,39	147,56
	3	155,01	169,12	158,28	167,72	155,44
	4	160,25	175,33	164,35	172,58	160,68
	Média	153,84	165,65	156,62	163,74	154,06
GPD	1	389,62	407,87	403,55	399,50	386,51
	2	323,37	344,98	327,42	339,77	322,81
	3	352,40	386,80	371,39	369,19	351,29
	4	380,73	401,76	396,74	385,47	376,95
	Média	361,53	385,35	374,78	373,48	359,39
PE	1	7,21	7,62	7,57	7,27	7,15
	2	5,41	6,00	5,85	5,57	5,38
	3	5,98	6,66	6,42	6,23	5,96
	4	6,18	7,04	6,75	6,42	6,16
	Média	6,20	6,83	6,65	6,37	6,16
EM	1	1,01	1,03	1,01	1,02	1,00
	2	0,87	0,92	0,89	0,89	0,86
	3	0,85	0,89	0,87	0,85	0,84
	4	1,00	1,07	1,03	1,06	1,00
	Média	0,93	0,98	0,95	0,96	0,93

PD: peso na desmama; GPD: ganho de peso pós-desmama; PE: perímetro escrotal; EM: escore de musculosidade.

Tabela 9. Estimativas da correlação de Pearson dos modelos 1, 2_U, 2_N, 3_U e 3_N, para os subconjuntos 1, 2, 3 e 4, pela validação cruzada 4-*fold*.

Característica	Subconjunto	1	2_U	2_N	3_U	3_N
PD	1	0,12	0,00	-0,02	0,06	0,12
	2	0,10	0,01	0,01	0,04	0,10
	3	0,14	-0,02	-0,03	0,02	0,12
	4	0,16	-0,03	-0,02	0,02	0,15
	Média	0,13	-0,01	-0,02	0,04	0,12
GPD	1	0,25	0,18	0,19	0,24	0,27
	2	0,17	0,09	0,09	0,15	0,17
	3	0,26	0,10	0,13	0,20	0,26
	4	0,26	0,17	0,17	0,24	0,28
	Média	0,24	0,14	0,15	0,21	0,25
PE	1	0,27	0,18	0,18	0,27	0,29
	2	0,30	0,16	0,14	0,29	0,31
	3	0,28	0,12	0,12	0,25	0,29
	4	0,31	0,12	0,12	0,27	0,31
	Média	0,29	0,15	0,14	0,27	0,30
EM	1	0,16	0,12	0,12	0,17	0,17
	2	0,19	0,11	0,12	0,18	0,21
	3	0,22	0,13	0,13	0,22	0,24
	4	0,18	0,07	0,06	0,13	0,17
	Média	0,19	0,11	0,11	0,18	0,20

PD: peso na desmama; GPD: ganho de peso pós-desmama; PE: perímetro escrotal; EM: escore de musculosidade.

A habilidade preditiva dos modelos avaliada pelas estimativas de EQM e correlação de Pearson foi consistente entre as estratégias de validação cruzada. Os modelos 1 e 3_N foram equivalentes em termos de capacidade de predição das características analisadas. Diferenças entre os modelos 3_U e 3_N devido à distribuição a priori assumida para o efeito dos marcadores ficou evidente, com pior desempenho preditivo do modelo 3_U. Isso pode ser atribuído à dificuldade do modelo 3_U em lidar com a existência de colinearidade entre os efeitos dos marcadores, o que não ocorre com o modelo 3_N, conforme discutido por Whittaker, Thompson e Dehan (2000).

Adicionalmente, as elevadas estimativas de EQM e as baixas correlações de Pearson exibidas pelos modelos 2_U e 2_N, demonstraram que a habilidade desses modelos em predizer o desempenho da próxima geração ou de dados retirados aleatoriamente é muito limitada. Esse resultado está de acordo com Lande e Thompson (1990) que relataram que a seleção baseada apenas na informação dos marcadores é mais eficiente quando a porção da variância genética aditiva explicada pelos marcadores é superior a herdabilidade da característica, o que não ocorre no presente estudo para o pequeno conjunto de marcadores analisado.

4. Conclusão

Em termos da correlação de *ranking*, uma tênue vantagem foi observada pelos modelos que assumiram os efeitos poligênicos e dos marcadores conjuntamente (modelos 3_U e 3_N). Os modelos 3_U e 3_N apresentaram os *rankings* dos animais mais similares ao *ranking* baseado nos valores genéticos de referência.

A maior qualidade de ajuste global foi exibida pelo modelo 3_N, exceto para PD, seguido pelo modelo 1, que considerou apenas os efeitos poligênicos. Em relação à habilidade preditiva, não foram observadas diferenças entre os modelos 1 e 3_N. Por outro lado, os modelos que incluíram apenas o efeito dos marcadores (modelos 2_U e 2_N) tiveram os piores ajustes e desempenhos preditivo.

Embora os resultados não demonstraram vantagens claras da inclusão conjunta dos efeitos poligênicos e dos marcadores nos modelos de avaliação genética, em termos da qualidade de ajuste e capacidade preditiva, estudos adicionais devem ser conduzidos para avaliar o impacto dessa informação na acurácia das estimativas dos valores genéticos.

CAPITULO II

Comparação de metodologias de estimação do efeito de substituição alélica sob a inferência Bayesiana

1. Introdução

O rápido desenvolvimento das técnicas de genotipagem simultânea para milhares de polimorfismos do tipo SNP aumentou o interesse pela seleção assistida por marcadores em larga escala, denominada seleção genômica, na qual se aproveita o desequilíbrio de ligação entre regiões cromossômicas associadas às características quantitativas e os marcadores dispersos ao longo do genoma para estimar o valor genético dos candidatos à seleção. As predições genômicas permitem a estimativa dos valores genéticos dos candidatos à seleção pela soma dos efeitos de substituição alélica de cada marcador ponderados pelo número de alelos presentes (0, 1 ou 2) em cada *locus*. Portanto, as estimativas dos efeitos de todos os marcadores devem ser obtidas simultaneamente, caso contrário, dois marcadores que estejam em desequilíbrio de ligação com a mutação causal teriam efeitos similares ou efeitos superestimados (GARRICK; SAATCHI, 2011).

O grande desafio da seleção genômica é a estrutura dos dados na qual o número de efeitos no modelo (p) excede largamente o número de observações fenotípicas disponíveis (n) usadas para a sua estimação, na chamada situação de "n pequeno, p grande" (GIANOLA et al., 2009). No geral existem duas maneiras de lidar com modelos de regressão superparametrizados: seleção de variáveis ou encurtamento (*shrinkage*) dos parâmetros do modelo (XU, 2007). Métodos como regressão *stepwise* (HABIER; FERNANDO; DEKKERS, 2007), aprendizado de máquina (LONG et al., 2007) e regressão via quadrados mínimos parciais e via componentes principais (SOLBERG et al., 2009) são exemplos de métodos que promovem a seleção de variáveis. Os métodos de encurtamento, ao invés da deleção de variáveis "não significativas", propõem a manutenção de todas as variáveis no modelo com penalização via encolhimento das estimativas dos coeficientes de regressão para zero (XU, 2007), dentre eles destacam-se: regressão de cumeeira bayesiana (GIANOLA; PEREZ-ENCISO, TORO, 2003), métodos bayesianos (MEUWISSEN; HAYES; GODDARD, 2001; HABIER et al., 2011) e LASSO bayesiano (PARK; CASELLA, 2008; DE LOS CAMPOS et al., 2009).

Os métodos de encurtamento bayesianos requerem o conhecimento ou a estimação dos parâmetros de regularização para cada *locus*, os quais consistem em uma razão entre a variância residual e a variância de cada marcador que podem ser estimadas a partir dos dados (GARRICK; SAATCHI, 2011). Esses métodos diferem segundo duas pressuposições que os influenciam na maneira e no grau de encurtamento dos modelos: se assumem que todos os marcadores ou se apenas uma fração dos marcadores tem efeito diferente de zero e se a razão entre a variância residual e a variância de cada marcador é constante ou variável de um marcador para outro (GARRICK; SAATCHI, 2011). Na regressão múltipla bayesiana, em que se assume distribuição a priori uniforme para os efeitos dos marcadores, não ocorre encurtamento do número de efeitos a ser estimado. Embora nos métodos regressão de cumeeira bayesiana e Bayes A, a pressuposição de que todos os marcadores têm efeito sobre a característica quantitativa de interesse é considerada, a maneira como ocorre a penalização é distinta. Na regressão de cumeeira bayesiana, a penalização é dada pela razão constante entre a variância residual e a variância de cada marcador e, no Bayes A, a distribuição a priori normal centrada em zero com razão variável entre a variância residual e a variância de cada marcador promove penalização específica sobre cada efeito estimado. Por outro lado, os métodos Bayes B, Bayes Cπ e LASSO bayesiano supõem que apenas uma fração (π) dos marcadores tem efeito e o processo para a determinação do grau de penalização é o que difere entre eles. No Bayes Cπ e no LASSO bayesiano o grau de encurtamento é definido pelos parâmetros π e λ, respectivamente, que são estimados com base no conjunto de dados, diferentemente do Bayes B, no qual o parâmetro π é definido de forma arbitrária. Adicionalmente, os métodos B e Cπ diferem quanto à razão entre a variância residual e a variância de cada marcador, que para o Bayes B é variável e para o Bayes Cπ é fixa. Embora a aplicação dos métodos de encurtamento bayesianos para a estimação dos efeitos de marcadores genéticos esteja relacionada com a necessidade de lidar com a questão da dimensionalidade dos dados, sua utilização no presente estudo é justificada pela possível existência de efeitos colineares entre os marcadores.

O objetivo do presente trabalho foi comparar seis metodologias de estimação dos efeitos dos marcadores (regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano) e avaliar o impacto na acurácia das estimativas dos valores genéticos e os conflitos de seleção quando essa informação foi incorporada em um programa de melhoramento genético de bovinos de corte, analisando, para tal, o ganho de peso pós-desmama, perímetro escrotal e escore de musculosidade, medidos ao redor dos 18 meses de idade.

2. Material e Métodos

2.1. Dados fenotípicos

Os dados analisados nesse projeto são pertencentes a Agro-Pecuária CFM Ltda. e fazem parte do seu programa de avaliação genética que é realizado pelo Grupo de Melhoramento Animal e Biotecnologia da Faculdade de Zootecnia e Engenharia de Alimentos (GMAB-FZEA).

Foram utilizados registros de animais nascidos entre 1984 e 2009, filhos de touros selecionados para características de desempenho e reprodutivas avaliados a pasto. As características analisadas foram ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM). O GPD foi calculado pela diferença de pesos observada entre os 550 dias (18 meses) e os 205 dias de idade, ajustada pelo número de dias entre as pesagens e padronizada para 345 dias. Os pesos aos 550 dias e os 205 dias de idade foram obtidos por meio de pesagem individual em balança eletrônica acopladas ao tronco de contenção após jejum total dos animais de, no mínimo, 12 horas. A mensuração do PE foi realizada na região do maior diâmetro do escroto por uma fita métrica metálica ao redor dos 550 dias de idade. A avaliação visual do escore de musculosidade foi realizada por um número ímpar de avaliadores treinados, na qual o escore de cada animal é resultante do consenso entre os avaliadores. A escala adotada variou de 1 a 6, sendo a maior nota dada ao animal que melhor expressa o perfil desejável para a característica.

2.2. Obtenção do fenótipo corrigido

Análises unicaracterística foram realizadas pelo método da máxima verossimilhança restrita por meio do conjunto de programas MTDFREML – *Multiple Trait Derivative Free Restricted Maximum Likelihood* (BOLDMAN et al., 1995) para estimar as soluções para os efeitos fixos e aleatórios para as características GPD, PE e EM, sob modelo animal, usando o banco de dados completo disponível para o rebanho em estudo. Os efeitos fixos considerados foram o grupo de contemporâneos composto por fazenda, safra, estação, sexo e grupo de manejo na mensuração, bem como a idade a mensuração como efeito linear e a idade da mãe ao parto como efeito linear e quadrático. Além do efeito aditivo direto, o efeito do grupo de

manejo na desmama foi assumido como efeito aleatório para GPD, PE e EM, conforme proposto por Pedrosa (2011).

Nesse estudo, definiram-se como animais pertencentes a uma mesma safra todos aqueles nascidos até o mês de maio do ano seguinte à estação de monta, a qual é realizada durante os meses de novembro a fevereiro. Foram consideradas como sendo quatro as estações do ano, divididas em trimestres a partir do mês de janeiro. Além disso, para a característica escore de musculosidade, as mensurações anteriores a 1999, realizadas utilizando uma escala de 1 a 9, foram convertidas para a escala atual de 1 a 6. Os dados de animais faltando a identificação da mãe, nascidos de grupos de reprodutor múltiplo e pertencentes a grupos de contemporâneos com menos de 5 animais foram desconsiderados. Após todas essas checagens o banco de dados analisado apresentou a estrutura descrita na Tabela 1.

Tabela 1. Estrutura dos dados analisados para a estimação das soluções dos efeitos fixos e predição das soluções dos efeitos aleatórios para o ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM).

DESCRIÇÃO	GPD	PE	EM
Nº de pais	1.326	1.276	1.300
Nº médio de filhos/pai	51,6	27,7	49,1
Nº máximo de filhos/pai	3.987	2.027	4067
Nº mínimo de filhos/pai	1	1	1
Nº de mães	22.507	18.130	21.642
Nº médio de filhos/mãe	3,0	2,0	3,0
Média de idd_mae (meses)	81,8	82,3	81,7
Nº de gc	1.046	518	1.111
Nº médio de animais por gc	65,4	68,3	57,5
Média de idd (dias)	546,7	542,2	541,5
Nº de fazendas	12	12	12
Nº de safras	26	25	20

Nº: número; idd_mãe: idade da mãe ao parto; gc: grupo de contemporâneos; idd: idade à mensuração.

As estatísticas descritivas dos dados utilizados para a estimação das soluções dos efeitos fixos e predição das soluções dos efeitos aleatórios estão apresentadas na Tabela 2. A matriz de parentesco correspondente era formada por um total de 116.652 animais.

Tabela 2. Estatísticas descritivas dos dados analisados para a estimação das soluções para os efeitos fixos e predição das soluções dos efeitos aleatórios para o ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM).

Característica	N	MÉDIA	DP	CV	MIN	MAX
GPD (kg)	68.424	114,47	32,25	28,18	1,01	294,54
PE (cm)	35.401	27,36	3,43	12,53	16,00	44,00
EM (unid)	63.854	3,61	1,00	27,84	1,00	6,00

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor fenotípico observado; MAX: máximo valor fenotípico observado.

Para todas as características analisadas, o fenótipo observado de cada indivíduo avaliado na análise unicaracterística foi corrigido para os mesmos efeitos fixos e aleatórios assumidos no modelo usado para a estimação da solução desses efeitos, exceto para o efeito aditivo direto. O fenótipo corrigido representa a soma das porções aditiva e residual do fenótipo observado. O efeito genético aditivo direto estimado foi assumido como sendo o valor genético de referência para os animais genotipados e não genotipados. O fenótipo corrigido e o valor genético serão usados posteriormente nas análises para a estimação dos efeitos de substituição alélica dos marcadores pelas diferentes metodologias avaliadas.

2.3. Dados moleculares

Foram genotipados 3.549 animais desse rebanho, dos quais 3.160 tinham pai e mãe conhecidos e por isso compuseram o banco de dados genotípico analisado no presente projeto. Esses dados consistiram de 377 fêmeas e 2.783 machos. No total, 3.010 mães e 752 pais estavam representados no banco de dados genotípico, dos quais 104 mães e 278 pais também foram genotipados. A maioria dessas mães tinha apenas uma progênie genotipada e 46 pais tinham mais de 10 progênies genotipadas, dos quais 21 eram pais genotipados. As estatísticas descritivas para GPD, PE e EM dos animais genotipados estão apresentadas na Tabela 3.

Tabela 3. Estatísticas descritivas para o ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM) mensuradas nos animais genotipados.

Característica	N	MÉDIA	DP	CV	MIN	MAX
GPD (kg)	3.033	125,19	30,50	24,36	10,63	272,86
PE (cm)	2.664	27,18	3,36	12,36	17,00	38,50
EM (unid)	3.149	3,58	1,17	32,68	1,00	6,00

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor fenotípico observado; MAX: máximo valor fenotípico observado.

Os animais foram genotipados para 222 marcadores genéticos do tipo polimorfismo de nucleotídeo único (SNP – *Single Nucleotide Polymorphism*) identificados em raças de origem

Bos taurus, representados por 123 SNP descritos na literatura como associados com alguma função biológica que afeta a expressão de características produtivas, dos quais 85-90% estão localizados na região de transcrição do gene e de 10-15% estão na região promotora do gene, e 99 SNP são marcadores usados em testes de paternidade, que embora não tenham sido associados com nenhuma função biológica, estão distribuídos ao longo do genoma, podendo por isso estar em desequilíbrio de ligação com genes de interesse. Todo processo de genotipagem foi realizado em laboratórios licenciados pela empresa Merial/Igenity, que é a companhia que detém o direito de uso dos marcadores utilizados nesse projeto.

As frequências alélicas e genotípicas de cada marcador foram estimadas por contagem simples dos diferentes alelos e genótipos por meio do procedimento PROC FREQ do pacote estatístico SAS. Os SNP que apresentaram menor frequência alélica (MAF) abaixo de 5% foram retirados das análises, com isso, 106 marcadores foram mantidos nas análises seguintes. Essa redução no número de marcadores mantido nas análises comparado com o número de marcadores genotipados deve-se ao fato desses marcadores terem sido descobertos em raças *Bos taurus* e a raça Nelore ser uma raça *Bos indicus*, o que resulta num grande número de marcadores fixados ou com a MAF inferior a 5%.

2.4. Método bayesiano

A inferência estatística tem como foco obter conclusões, a partir de dados observados **y**, de quantidades não observadas **θ**. As duas principais escolas de inferência são a inferência frequentista (ou clássica) e a inferência bayesiana. A inferência bayesiana é uma forma de inferência que considera os parâmetros do modelo como variáveis aleatórias e, expressa o conhecimento acumulado sobre eles por meio de distribuições a priori (GIANOLA; FERNANDO, 1986). Essa escola descreve as incertezas sobre **θ** de forma probabilística e baseia-se no Teorema de Bayes para obter conclusões sobre os parâmetros do modelo por meio da probabilidade condicional aos dados observados **y**.

Considere um vetor de parâmetros a serem estimados **θ**, um conjunto de dados observados **y** e sua densidade conjunta p(**θ**, **y**), pela teoria de probabilidades, segundo Gianola e Fernando (1986), tem-se:

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

e

$$p(\boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}) * p(\mathbf{y})$$

em que $p(\boldsymbol{\theta})$ e $p(\mathbf{y})$ são as densidades marginais de $\boldsymbol{\theta}$ e $\mathbf{y}$, respectivamente.

Combinando as duas expressões, tem-se:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

Como $p(\mathbf{y})$ não é uma função de $\boldsymbol{\theta}$, a expressão reescrita é o Teorema de Bayes:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}) * p(\boldsymbol{\theta})$$

em que $p(\boldsymbol{\theta})$ é a densidade a priori de $\boldsymbol{\theta}$, a qual reflete a incerteza sobre os possíveis valores de $\boldsymbol{\theta}$ antes dos dados y serem observados; $p(\mathbf{y}|\boldsymbol{\theta})$ é a função de verossimilhança, que representa a contribuição dos dados $\mathbf{y}$ para o conhecimento de $\boldsymbol{\theta}$**;** $p(\boldsymbol{\theta}|\mathbf{y})$ é a densidade posterior conjunta de $\boldsymbol{\theta}$ usada para fazer inferências sobre os parâmetros do modelo**.**

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são uma família de métodos iterativos que permite obter amostras da densidade posterior conjunta de $\boldsymbol{\theta}$, a partir da qual é possível fazer inferências sobre toda a função densidade. Para obter as amostras da densidade posterior conjunta, o algoritmo de MCMC mais usado é o amostrador de Gibbs, o qual necessita do conhecimento das distribuições condicionais completas a posteriori de todos os parâmetros desconhecidos para a construção das cadeias de Markov. Entretanto, quando as distribuições condicionais completas a posteriori não são conhecidas, a aplicação do algoritmo Metropolis-Hastings possibilita a amostragem da densidade posterior conjunta de $\boldsymbol{\theta}$ utilizando uma distribuição candidata como distribuição auxiliar para atualizar o algoritmo durante o processo iterativo.

2.5. Estimação dos valores genéticos moleculares

Para estimação dos efeitos de substituição alélica dos marcadores realizada por meio de análises de regressão linear múltipla, duas pré-suposições foram adotadas, os efeitos dos marcadores se combinam aditivamente entre e dentro de *loci* e a relação entre a variável

dependente **y** e o valor genético molecular $\left(\sum_{k=1}^{106} \mathbf{x}_k g_k\right)$ é dada pelo seguinte modelo estatístico:

$$\mathbf{y} = \mu + \sum_{k=1}^{106} \mathbf{x}_k g_k + \mathbf{e}$$

em que **y** é um vetor $n \times 1$ dos fenótipos corrigidos (n = número de observações); μ é a media geral; $\mathbf{x_k}$ é um vetor coluna de ordem $n \times 1$ dos genótipos para o marcador k (k = número de marcadores analisados); $g_{\mathbf{k}}$ é o efeito aditivo do marcador k com distribuição N(**0**, $\mathbf{I}\sigma_g^2$); **e** é o vetor dos efeitos residuais, assumidos como sendo independentes e com distribuição N(**0**, $\mathbf{I}\sigma_e^2$).

O modelo é chamado de regressão linear múltipla, pois envolve mais de um coeficiente de regressão e é linear em relação aos parâmetros $g = (g_1, g_2, \ldots, g_k)$.

O número de observações utilizadas nessas análises foi o mesmo apresentado anteriormente na Tabela 3, que corresponde ao número de animais genotipados por característica analisada. Os métodos adotados para a estimação dos efeitos aditivos dos marcadores foram regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano, todos sob a perspectiva bayesiana de análise de dados. Essas análises foram realizadas por meio de programas elaborados em linguagem R, durante estágio realizado no Instituto de Investigación y Tecnología Agroalimentarias da Cataluña (IRTA) sob a supervisão da Dra. Noelia Ibáñez-Escriche, no período de fevereiro a agosto de 2011.

A inferência bayesiana pelo método Monte Carlo usando as cadeias de Markov (MCMC) foi aplicada nas análises assumindo distribuição normal para os dados:

$$\mathbf{y}|\boldsymbol{\theta}_j \sim \mathrm{N}\left(\mu + \mathbf{Xg}, \mathbf{I}\sigma_e^2\right)$$

em que o subscrito j = 1, 2, 3, 4, 5 e 6, corresponde aos métodos de regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano, respectivamente; $\boldsymbol{\theta}_j$ são os parâmetros desconhecidos para cada modelo; σ_e^2 é a variância residual.

2.5.1. Regressão múltipla bayesiana

Para as análises de regressão múltipla sob a inferência bayesiana, as distribuições a priori consideradas foram uniforme para μ e $\mathbf{g} = (g_1, g_2, ..., g_{106})$ e $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais. Para a σ^2_e assumiu-se a distribuição χ^{-2} (ν, s^2) com ν = -2 e $s^2 = 0$, que a torna uma priori plana e imprópria. Na implementação do amostrador de Gibbs, os parâmetros do modelo de regressão foram amostrados das distribuições condicionais completas a posteriori descritas a seguir.

$$f(\mu|\mathbf{g}, \sigma^2_e, \mathbf{y}) \propto N\left(\hat{\mu}, \frac{\sigma^2_e}{n}\right)$$

em que $\hat{\mu} = (\mathbf{1}'_n\mathbf{1}_n)^{-1}\mathbf{1}_n(\mathbf{y} - \mathbf{X}'\mathbf{g})$

$$f\left(g_k\middle|\mu, \mathbf{g}_{\_k}, \sigma^2_e, \mathbf{y}\right) \propto N(\hat{g}_k, (\mathbf{x}'_k\mathbf{x}_k)^{-1}\sigma^2_e)$$

em que $\hat{g}_k = (\mathbf{x}'_k\mathbf{x}_k)^{-1}\mathbf{x}'_k(\mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}_{\_k}\mathbf{g}_{\_k})$

$$f(\sigma^2_e|\mu, \mathbf{g}, \mathbf{y}) \propto \chi^{-2}\left(n + \nu; \frac{\mathbf{e}'\mathbf{e} + \nu s^2}{n + \nu}\right)$$

em que $\mathbf{e} = \mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}'\mathbf{g}$, ν = -2 e $s^2 = 0$

em que n é o número de observações; k é o número de marcadores.

2.5.2. Regressão de cumeeira bayesiana

No contexto frequentista, a estimação dos efeitos dos marcadores pelo método dos mínimos quadrados trata os marcadores como efeitos fixos contínuos e, os coeficientes de regressão associados a eles são obtidos pela minimização da soma de quadrados dos resíduos. Nesse caso, os estimadores de mínimos quadrados para os coeficientes de regressão linear associados aos marcadores são obtidos pela equação:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} * (\mathbf{X}'\mathbf{y})$$

Whittaker, Thompson e Dehan (2000) propuseram empregar a regressão de cumeeira para a estimação dos efeitos dos marcadores, como uma maneira de solucionar o problema de colinearidade causado pela inclusão de muitos marcadores no modelo de regressão. Esse

procedimento regulariza (*shrink*) as estimativas de mínimos quadrados a zero prevenindo que a matriz $\mathbf{X'X}$ seja singular ou próxima a isso, geralmente promove uma redução nas estimativas dos erros quadráticos médios e assume que os efeitos dos marcadores $\mathbf{g}$ são aleatórios com uma variância comum σ_g^2. Nesse caso, os estimadores de mínimos quadrados para os coeficientes de regressão linear associados aos marcadores são substituídos por:

$$\hat{\beta} = (\mathbf{X'X} + \lambda\mathbf{I})^{-1} * (\mathbf{X'y})$$

em que $\mathbf{I}$ é uma matriz identidade e λ é o parâmetro que determina o grau de encurtamento (*shrinkage*).

No contexto bayesiano, a regressão de cumeeira é equivalente a adotar, para o vetor dos coeficientes de regressão, uma distribuição a priori normal centrada em zero com uma estrutura de covariância igual a uma matriz identidade multiplicada pela variância dessa distribuição (GIANOLA; PEREZ-ENCISO; TORO, 2003). Essa abordagem implica que os coeficientes de regressão individuais dos fenótipos nos marcadores sejam independentes e identicamente distribuídos a priori.

Uma dificuldade desse método é a definição do parâmetro λ. Whittaker, Thompson e Dehan (2000) escolheram o valor de λ que minimizasse a soma de quadrados dos erros obtidas pela diferença entre os valores genéticos verdadeiros dos animais e os valores genéticos moleculares estimados pela regressão de cumeeira. Se $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ na equação da regressão de cumeeira, esse método é igual ao BLUP usado por Meuwissen, Hayes e Goddard (2001) e ao método bayesiano proposto por Gianola, Perez-Enciso e Toro (2003).

Outro aspecto importante é a estimação da variância dos efeitos dos marcadores, σ_g^2. Meuwissen, Hayes e Goddard (2001) assumiram como sendo a variância genética esperada em um modelo de deriva e mutação genética baseado em QTLs (*Quantitative Trait Loci*), a qual não representa a variância dos marcadores. Gianola, Perez-Enciso e Toro (2003) demonstraram como estimar a variância dos efeitos dos marcadores por meio de um "BLUP bayesiano", que foi a abordagem aplicada no presente estudo.

As distribuições a priori assumidas para os parâmetros desconhecidos foram uniforme para μ, $N(\mathbf{0}, \mathbf{I}\sigma^2_g)$ para $\mathbf{g} = (g_1, g_2, ..., g_{106})$ e $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais. Para as variâncias σ^2_g e σ^2_e assumiu-se a distribuição χ^{-2} (ν, s^2) com $\nu = -2$ e $s^2 = 0$, que a torna uma priori plana. Nesse caso, as distribuições condicionais completas a posteriori das quais foram feitas as amostragens por Gibbs dos parâmetros desconhecidos, foram:

$$f\left(\mu\middle|\mathbf{g},\sigma_g^2,\sigma_e^2,\mathbf{y}\right)\propto N\left(\hat{\mu},\frac{\sigma_e^2}{n}\right)$$

em que $\hat{\mu}=(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n(\mathbf{y}-\mathbf{X}'\mathbf{g})$

$$f\left(g_k\middle|\mu,\mathbf{g}_{\_k},\sigma_g^2,\sigma_e^2,\mathbf{y}\right)\propto N(\hat{g}_k,(\mathbf{x}_k'\mathbf{x}_k+\lambda)^{-1}\sigma_e^2)$$

em que $\hat{g}_k=(\mathbf{x}_k'\mathbf{x}_k+\lambda)^{-1}\mathbf{x}_k'(\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}_{\_k}\mathbf{g}_{\_k})$ e $\lambda=\frac{\sigma_e^2}{\sigma_g^2}$

$$f\left(\sigma_g^2\middle|\mu,\mathbf{g},\sigma_e^2,\mathbf{y}\right)\propto\chi^{-2}\left(k+\nu;\frac{\mathbf{g}'\mathbf{g}+\nu s^2}{k+\nu}\right)$$

em que $\nu = -2$ e $s^2 = 0$

$$f\left(\sigma_e^2\middle|\mu,\mathbf{g},\sigma_g^2,\mathbf{y}\right)\propto\chi^{-2}\left(n+\nu;\frac{\mathbf{e}'\mathbf{e}+\nu s^2}{n+\nu}\right)$$

em que $\mathbf{e}=\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}'\mathbf{g}$, $\nu = -2$ e $s^2 = 0$

em que n é o número de observações; k é o número de marcadores.

2.5.3. Bayes A

A estimação dos efeitos dos marcadores pela metodologia Bayes A (MEUWISSEN; HAYES; GODDARD, 2001) difere da anterior no que diz respeito à variância dos efeitos dos marcadores. Nesse caso, assume-se que os efeitos dos marcadores **g** são aleatórios com uma variância diferente para cada efeito, $\sigma_{g_k}^2$.

As distribuições a priori assumidas para os parâmetros desconhecidos foram uniforme para μ, $N(0, \sigma_{g_k}^2)$ para g_k (k = 1, 2, ..., 106) e $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais. A distribuição a priori assumida para $\sigma_{g_k}^2$ foi a distribuição χ^{-2} (ν, s^2) com ν = 4,012 e s^2 = 0,002. Esses valores de ν e s^2 foram os mesmos propostos por Meuwissen, Hayes e Goddard (2001) e fazem com que a distribuição a priori não seja plana. Para σ^2_e, a distribuição χ^{-2} (ν, s^2) com ν = -2 e s^2 = 0, que a torna uma priori plana, foi assumida como distribuição a priori.

As distribuições condicionais completas a posteriori, para os parâmetros desconhecidos, usadas na amostragem de Gibbs estão apresentadas a seguir.

$$f\left(\mu\middle|\mathbf{g},\sigma_g^2,\sigma_e^2,\mathbf{y}\right)\propto N\left(\hat{\mu},\frac{\sigma_e^2}{n}\right)$$

em que $\hat{\mu}=(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n(\mathbf{y}-\mathbf{X}'\mathbf{g})$

$$f\left(g_k\middle|\mu,\mathbf{g}_{\_k},\sigma_{g_k}^2,\sigma_e^2,\mathbf{y}\right)\propto N(\hat{g}_k,(\mathbf{x}_k'\mathbf{x}_k+\lambda)^{-1}\sigma_e^2)$$

em que $\hat{g}_k=(\mathbf{x}_k'\mathbf{x}_k+\lambda_k)^{-1}\mathbf{x}_k'(\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}_{\_k}\mathbf{g}_{\_k})$ e $\lambda_k=\frac{\sigma_e^2}{\sigma_{g_k}^2}$

$$f\left(\sigma_{g_k}^2\middle|\mu,\mathbf{g},\sigma_e^2,\mathbf{y}\right)\propto\chi^{-2}\left(1+\nu;\frac{g_k{}'g_k+\nu s^2}{1+\nu}\right)$$

em que $\nu = 4{,}012$ e $s^2 = 0{,}002$

$$f\left(\sigma_e^2\middle|\mu,\mathbf{g},\sigma_g^2,\mathbf{y}\right)\propto\chi^{-2}\left(n+\nu;\frac{\mathbf{e}'\mathbf{e}+\nu s^2}{n+\nu}\right)$$

em que $\mathbf{e}=\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}'\mathbf{g}$, $\nu = -2$ e $s^2 = 0$

em que n é o número de observações; k é o número de marcadores.

2.5.4. Bayes B

A metodologia Bayes B baseia-se na teoria de que muitos *loci* não têm efeito nem contribuem para a variância genética e que poucos *loci* têm efeito e variância, portanto estão segregando. Ao contrário da metodologia Bayes A, em que a probabilidade de $\sigma_{g_k}^2=0$ é infinitesimal, a metodologia Bayes B assume uma distribuição a priori que tem uma alta densidade, π, de $\sigma_{g_k}^2=0$ e tem uma distribuição qui-quadrado escalonada invertida para $\sigma_{g_k}^2>0$ (MEUWISSEN; HAYES; GODDARD, 2001). Para tanto, as distribuições a priori assumidas para g_k e $\sigma_{g_k}^2$ foram:

$$(g_k|\pi,\sigma_{g_k}^2)\begin{cases}\sim N\left(0,\sigma_{g_k}^2\right)\text{ com probabilidade }(1-\pi)\\ =0\text{ com probabilidade }(\pi)\end{cases}$$

$$\sigma_{g_k}^2\sim\chi^{-2}\,(\nu,s^2)$$

em que $\nu = 4{,}234$ e $s^2 = 0{,}0429$

Com essas distribuições a priori o amostrador de Gibbs não seria capaz de cobrir todo o espaço de amostragem do método Bayes B. Isso porque a amostragem de $\sigma^2_{g_k} = 0$ não seria possível se $g_k' g_k > 0$. Por outro lado, a probabilidade de amostrar $g_k = 0$ quando $\sigma^2_{g_k} > 0$ é zero. Assim, uma maneira proposta por Meuwissen, Hayes e Goddard (2001) foi amostrar g_k e $\sigma^2_{g_k}$ simultaneamente a partir da distribuição:

$$p\left(\sigma^2_{g_k}, g_k | y^*\right) = p\left(\sigma^2_{g_k} | y^*\right) * p(g_k | \sigma^2_{g_k}, y^*)$$

em que y* representa os dados corrigidos para os efeitos da média e de todos os outros marcadores, exceto para g_k.

Nesse caso a amostragem por Gibbs deveria ser feita primeiro em $p\left(\sigma^2_{g_k} | y^*\right)$ e, em seguida, em $p(g_k | \sigma^2_{g_k}, y^*)$. Entretanto, a distribuição $p\left(\sigma^2_{g_k} | y^*\right)$ não tem forma conhecida e, portanto, o amostrador de Gibbs não poderia ser usado. Dessa maneira, o algoritmo de Metropolis-Hastings foi usado para amostrar $p\left(\sigma^2_{g_k} | y^*\right)$, em que a distribuição a priori de $\sigma^2_{g_k}$ foi usada como distribuição auxiliar para atualizar o algoritmo.

As distribuições a priori assumidas para os demais parâmetros desconhecidos foram uniforme para μ e $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais, com distribuição χ^{-2} (ν, s^2) com ν = -2 e s^2 = 0 para σ^2_e. Com isso, as distribuições condicionais completas a posteriori usadas para amostragem pelo algoritmo de Gibbs foram:

$$f\left(\mu | \mathbf{g}, \sigma^2_g, \sigma^2_e, \mathbf{y}\right) \propto N\left(\hat{\mu}, \frac{\sigma^2_e}{n}\right)$$

$$\text{em que } \hat{\mu} = (\mathbf{1}_n' \mathbf{1}_n)^{-1} \mathbf{1}_n (\mathbf{y} - \mathbf{X}'\mathbf{g})$$

$$f\left(g_k | \mu, \mathbf{g}_{\_k}, \sigma^2_{g_k}, \sigma^2_e, \mathbf{y}\right) \propto N(\hat{g}_k, (\mathbf{x}_k' \mathbf{x}_k + \lambda)^{-1} \sigma^2_e) \text{ para } \sigma^2_{g_k} > 0$$

$$\text{em que } \hat{g}_k = (\mathbf{x}_k' \mathbf{x}_k + \lambda)^{-1} \mathbf{x}_k' (\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}_{\_k} \mathbf{g}_{\_k}) \text{ e } \lambda = \frac{\sigma^2_e}{\sigma^2_{g_k}}$$

$$f\left(\sigma^2_e | \mu, \mathbf{g}, \sigma^2_g, \mathbf{y}\right) \propto \chi^{-2}\left(n + \nu; \frac{\mathbf{e}'\mathbf{e} + \nu s^2}{n + \nu}\right)$$

$$\text{em que } \mathbf{e} = \mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}'\mathbf{g}, \ \nu = -2 \text{ e } s^2 = 0$$

em que n é o número de observações; k é o número de marcadores.

2.5.5. Bayes Cπ

Na metodologia Bayes B demonstrada no item anterior, a assinação da porcentagem, π, de marcadores com efeitos e a variâncias iguais a zero é realizada de maneira arbitrária. Além disso, tanto o método Bayes A quanto o Bayes B apresentam grande dependência das variâncias da distribuição a priori. No método Bayes Cπ, proposto por Habier et al. (2011), o valor de π é estimado conjuntamente com os outros parâmetros desconhecidos do modelo. Esse método considera uma mesma variância para todos os efeitos de g_k diferentes de zero, o que faz reduzir a dependência da distribuição a priori. Nesse caso, as distribuições a priori assumidas para g_k, σ_g^2 e π foram:

$$(g_k|\pi,\sigma_g^2)\begin{cases}\sim N(0,\sigma_g^2) \text{ com probabilidade } (1-\pi)\\ = 0 \text{ com probabilidade } (\pi)\end{cases}$$

$$\sigma_g^2 \sim \chi^{-2}\,(\nu, s^2)$$

em que $\nu = 4{,}234$ e $s^2 = 0{,}0429$

$$\pi \sim \text{Uniforme}\,(0,1)$$

As distribuições a priori assumidas para os demais parâmetros desconhecidos foram uniforme para μ e $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais, com distribuição χ^{-2} (ν, s^2) com ν = -2 e $s^2 = 0$ para σ^2_e. Para a amostragem dos parâmetros desconhecidos no modelo foi utilizado o algoritmo de Gibbs, uma vez que todas as distribuições condicionais completas a posteriori tinham forma conhecida:

$$f(\mu|\pi,\mathbf{g},\sigma_g^2,\sigma_e^2,\mathbf{y}) \propto N\left(\hat{\mu},\frac{\sigma_e^2}{n}\right)$$

em que $\hat{\mu} = (\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n(\mathbf{y}-\mathbf{X}'\mathbf{g})$

$$f(g_k|\mu,\pi,\mathbf{g}_{\_k},\sigma_g^2,\sigma_e^2,\mathbf{y}) \propto N(\hat{g}_k,(\mathbf{x}_k'\mathbf{x}_k+\lambda)^{-1}\sigma_e^2) \text{ para } \sigma_g^2 > 0$$

em que $\hat{g}_k = (\mathbf{x}_k'\mathbf{x}_k+\lambda)^{-1}\mathbf{x}_k'(\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}_{\_k}\mathbf{g}_{\_k})$ e $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$

$$f\left(\sigma_g^2 \middle| \mu, \pi, \mathbf{g}, \sigma_e^2, \mathbf{y}\right) \propto \chi^{-2}\left(k + \nu; \frac{\mathbf{g}'\mathbf{g} + \nu s^2}{1 + \nu}\right)$$

em que $\nu = 4{,}234$ e $s^2 = 0{,}0429$

$$f\left(\pi \middle| \mu, \mathbf{g}, \sigma_g^2, \sigma_e^2, \mathbf{y}\right) \propto \text{Beta}\ (k - \mathbf{g}'\mathbf{g} + 1, \mathbf{g}'\mathbf{g} + 1)$$

$$f(\sigma_e^2 | \mu, \pi, \mathbf{g}, \mathbf{y}) \propto \chi^{-2}\left(n + \nu; \frac{\mathbf{e}'\mathbf{e} + \nu s^2}{n + \nu}\right)$$

em que $\mathbf{e} = \mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}'\mathbf{g}$, $\nu = -2$ e $s^2 = 0$

em que n é o número de observações; k é o número de marcadores.

2.5.6. LASSO bayesiano

O método LASSO (*Least Absolute Shrinkage and Selection Operator* - TIBSHIRANI, 1996) combina o encurtamento (*shrinkage*) das estimativas dos coeficientes de regressão com a seleção de variáveis, minimizando a soma de quadrados residual e regularizando via encurtamento as somas de quadrados dos valores absolutos dos coeficientes de regressão. De los Campos et al. (2009) propuseram a aplicação desse método na seleção genômica, de modo que os efeitos dos marcadores fossem estimados pela otimização do seguinte problema:

$$\min_g \left\{ (\mathbf{y} - \mu - \mathbf{X}\mathbf{g})'(\mathbf{y} - \mu - \mathbf{X}\mathbf{g}) + \lambda \sum_{j=1}^{k} |g_j| \right\}$$

em que λ é o parâmetro que determina o grau de encurtamento (*shrinkage*), se $\lambda=0$ não há encurtamento; $\sum_{j=1}^{k} |g_j|$ é a soma dos valores absolutos dos coeficientes de regressão associados aos marcadores.

Na implementação desse método, impõe-se que os parâmetros de regressão tenham distribuições a priori de Laplace idênticas e independentes. Com isso, a distribuição a priori para os coeficientes de regressão dos marcadores apresentou-se como sendo um produto de densidades exponenciais duplas:

$$p(\mathbf{g}|\lambda) = \prod_{j=i}^{k} \left(\frac{\lambda}{2}\right) e^{(-\lambda|g_j|)}$$

A diferença entre o método LASSO bayesiano e as metodologias de regressão de cumeeira bayesiana, Bayes A e Bayes B está no fato dessas assumirem um produto de densidades normais independentes como distribuição a priori para os coeficientes de regressão dos marcadores:

$$p(\mathbf{g}|\sigma_g^2) = \prod_{j=i}^{k} \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{{g_j}^2}{2\sigma_g^2}\right)$$

As demais distribuições a priori assumidas foram uniforme para μ, $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$ para os efeitos residuais, Exp(λ) para τ^2 e G(α, β) com α = 5 e β = 1 (comunicação pessoal de Oscar González-Recio) para λ^2. Para a variância σ^2_e assumiu-se a distribuição χ^{-2} (ν, s^2) com ν = 3 e s^2 = 0,7 (comunicação pessoal de Oscar González-Recio).

Park and Casella (2008), considerando que a distribuição densidade exponencial a priori para os coeficientes de regressão dos marcadores pode ser vista como uma mistura de distribuições normais, implementaram o algoritmo de Gibbs assumindo as seguintes distribuições condicionais completas a posteriori para os parâmetros desconhecidos:

$$f(\mu|\mathbf{g}, \sigma_e^2, \tau^2, \lambda, \mathbf{y}) \propto N\left(\hat{\mu}, \frac{\sigma_e^2}{n}\right)$$

$$\text{em que } \hat{\mu} = (\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n(\mathbf{y} - \mathbf{X}'\mathbf{g})$$

$$f(g_k|\mu, \mathbf{g}_{\_k}, \sigma_e^2, \tau^2, \lambda, \mathbf{y}) \propto N\left(\hat{g}_k, \left(\mathbf{x}_k'\mathbf{x}_k + \tau^{-2}\right)^{-1}\sigma_e^2\right)$$

$$\text{em que } \hat{g}_k = \left(\mathbf{x}_k'\mathbf{x}_k + \tau^{-2}\right)^{-1}\mathbf{x}_k'(\mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}_{\_k}\mathbf{g}_{\_k})$$

$$f(\tau^2|\mu, \mathbf{g}, \sigma_e^2, \lambda, \mathbf{y}) \propto IG\left(\frac{\sigma_e\lambda}{|g_k|}, \lambda^2\right)$$

$$f\left(\lambda^2\middle|\mu,\mathbf{g},\sigma_e^2,\tau^2,\mathbf{y}\right)\propto G\left(k+\alpha,\frac{1}{2}\sum_{j=1}^{k}\tau_j^2+\beta\right)$$

em que $\alpha = 5$ e $\beta = 1$

$$f(\sigma_e^2|\mu,\pi,\mathbf{g},\mathbf{y})\propto\chi^{-2}\left(n+\nu;\frac{\mathbf{e}'\mathbf{e}+\nu s^2}{n+\nu}\right)$$

em que $\mathbf{e}=\mathbf{y}-\mu\mathbf{1}_n-\mathbf{X}'\mathbf{g}$, $\nu = 3$ e $s^2 = 0{,}7$

em que n é o número de observações; k é o número de marcadores.

Para a estimação dos efeitos aditivos dos marcadores pelas metodologias descritas foram realizadas 1.000.000 de iterações para a obtenção das cadeias de Markov para os parâmetros desconhecidos em cada metodologia para todas as características estudadas. Para o período de aquecimento da cadeia assumiu-se um descarte inicial (*burn in*) de 80.000 iterações e, para assegurar a independência entre amostras consecutivas, considerou-se um intervalo de retirada (*thinning*) de 100 amostras. Desse modo, todas as inferências foram realizadas com as 9.200 amostras salvas após o período de aquecimento.

As estimativas dos efeitos dos marcadores foram dispostas em gráficos do tipo linha apresentados nos Apêndices A, B e C para cada uma das características analisadas. O efeito de cada marcador estimado pelas metodologias regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano foi ponderado pelo número de alelos favoráveis presentes (0, 1 ou 2) em cada *locus* e somado para a obtenção do valor genético molecular (VGM) dos animais, conforme proposto por Lande e Thompson (1990), sob o modelo:

$$\widehat{\mathbf{VGM}}=\sum_{k=1}^{106}\mathbf{x}_k\hat{g}_k$$

em que $\widehat{\mathbf{VGM}}$ é um n × 1 vetor das estimativas dos valores genéticos moleculares (*n* = número de observações); $\mathbf{x_k}$ é um vetor coluna de ordem *n* × *1* dos genótipos para o marcador *k* (*k* = número de marcadores analisados); $\hat{g}_k$ é a estimativa do efeito aditivo do marcador *k*.

Para analisar a relação entre as estimativas dos VGM dos animais entre as diferentes metodologias aplicadas para a estimação do efeito dos marcadores foram construídos gráficos de dispersão por meio do pacote estatístico SAS. Adicionalmente, foram calculados os

coeficientes de correlação de Pearson e de Spearman entre os VGM estimados pelas seis metodologias por meio do procedimento PROC CORR do pacote estatístico SAS para cada uma das características analisadas, os quais estão apresentados nos Apêndices D, E e F.

2.6. Inclusão da informação molecular na avaliação genética

A informação dos marcadores genéticos foi incorporada na avaliação genética tradicional combinando os fenótipos observados para as características GPD, PE e EM com os respectivos VGM em análises bicaracterísticas utilizando o conjunto de programas MTDFREML – *Multiple Trait Derivative Free Restricted Maximum Likelihood* (BOLDMAN et al., 1995), sob modelo animal. O *software* MTDFREML usa técnicas de matrizes esparsas do pacote de público domínio FSPAK desenvolvido por Perez-Enciso, Misztal e Elzo (1994), baseado no pacote SPARSPAK (GEORGE; LUI, 1980) e fatoração de Choleski para obter o *log* do determinante da matriz dos coeficientes e a soma de quadrados generalizada dos resíduos. Adicionalmente, esse pacote calcula a inversa da matriz de parentesco segundo as regras definidas por Quaas (1976) e os componentes de (co) variância e os parâmetros genéticos pelo procedimento da máxima verossimilhança restrita, usando um algoritmo livre de derivadas (SMITH; GRASER, 1986).

A estrutura dos dados analisados é a mesma apresentada na Tabela 1, sendo que o número de observações com VGM corresponde ao número de animais genotipados para cada característica, conforme apresentado na Tabela 3.

Os efeitos fixos considerados para o fenótipo observado para GPD, PE e EM foram o grupo de contemporâneos composto por fazenda, safra, estação, sexo e grupo de manejo na mensuração, bem como a idade a mensuração como efeito linear e a idade da mãe ao parto como efeito linear e quadrático. Como efeitos aleatórios foram assumidos o efeito genético aditivo direto e o efeito do grupo de manejo na desmama para GPD, PE e EM. Para os VGM, apenas o efeito aleatório genético aditivo direto foi considerado.

A análise do fenótipo observado e do VGM simultaneamente é justificada pela existência de vínculo entre eles advindo da matriz de parentesco, de modo análogo ao que ocorre em análises de características limitadas pelo sexo (Boldman et al., 1995). As covariâncias residuais entre o fenótipo observado e VGM foram assumidas como sendo iguais a zero. O critério de convergência assumido foi quando a variância do simplex alcançou 10^{-9}, isto ocorre quando a variação do valor do *log* da função do ciclo anterior e do presente for menor que 10^{-9}.

Duas estratégias foram adotadas para avaliar as alterações causadas pela incorporação das informações dos marcadores genéticos na avaliação genética clássica: as variações na acurácia das estimativas dos valores genéticos e os conflitos de seleção entre as classificações dos animais baseadas no valor genético clássico estimado em análises unicaracterística e no valor genético assistido por marcadores estimado em análises bicaracterísticas. Essas estratégias foram aplicadas dentro de categorias de animais definidas por sua importância dentro do rebanho, dado o potencial para transmitir seus alelos para as gerações futuras. As categorias consideradas foram:

- ✓ GERAL: todos os animais que compuseram a matriz de parentesco;
- ✓ TOURO: pai de animais que apresentaram medições em qualquer uma das características analisadas;
- ✓ CEIP: animais jovens que receberam o certificado CEIP ou que estavam aptos a receber e foram aprovados em avaliação visual, considerados como tourinhos de reposição para o programa.

Adicionalmente, as categorias TOUROS e CEIP foram subdivididas em animais não genotipados e genotipados, totalizando cinco categorias analisadas.

Em 1995, o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) instituiu o Certificado Especial de Identificação e Produção – CEIP para bovinos resultantes de cruzamentos planificados e para animais de raças puras com aptidão para corte, que ao serem avaliados geneticamente obtivessem classificação superior em relação ao grupo de animais submetidos à seleção, iniciando com os 20% melhores animais e chegando ao limite de 30%. Dentre as empresas que possuem programas de melhoramento animal autorizados a emitir o certificado, destacam-se a Agro-Pecuária CFM Ltda., a Agropecuária Jacarezinho Ltda. e o Programa PAI*N*T da Central de Inseminação Artificial Lagoa da Serra Ltda.

2.6.1. Análise do impacto na acurácia

As acurácias ($r_{T,I}$) das estimativas dos valores genéticos clássicos e dos valores genéticos assistidos por marcadores para as características GPD, PE e EM foram obtidas em análises uni e bicaracterísticas, respectivamente, por meio da expressão:

$$r_{T,I} = \sqrt{1 - \frac{VEP}{\sigma_a^2}}$$

em que VEP é a variância do erro de predição, estimada por:

$$VEP = diag\left(Z'Z + A^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1} * \sigma_e^2$$

em que σ_a^2 é a variância genética aditiva; σ_e^2 é a variância residual.

O impacto na acurácia das estimativas dos valores genéticos pela inclusão da informação molecular na avaliação genética do rebanho analisado foi calculado como a variação percentual média observada entre a acurácia média das estimativas dos valores genéticos assistidos por marcadores para cada metodologia e a acurácia média das estimativas dos valores genéticos clássicos, para as cinco classes de animais: GERAL, CEIP_ng, CEIP_g, TOURO_ng e TOURO_g.

2.6.2. Avaliação dos conflitos de seleção

Os conflitos de seleção são divergências na categorização dos animais em função do critério adotado, ou seja, quando um animal é classificado de maneira distinta dependendo do critério pelo qual está sendo avaliado. Visando demonstrar que a incorporação da informação molecular nas avaliações genéticas, não somente, influencia a acurácia das estimavas dos valores genéticos, bem como pode causar mudanças na classificação dos animais, dois critérios de seleção foram assumidos: o valor genético estimado em análises unicaracterística, denominado valor genético clássico e o valor genético assistido por marcadores estimado em análises bicaracterísticas, que consideraram, como segunda característica, os valores genéticos moleculares obtidos pela metodologia Lasso bayesiano.

Os conflitos de seleção foram avaliados pelo cálculo da proporção de animais com classificação divergente ao serem selecionados 20% dos animais de maiores valores genéticos na população. Gráficos de dispersão entre as estimativas dos valores genéticos clássicos e assistidos por marcadores foram construídos por meio do procedimento PROC GPLOT do pacote estatístico SAS.

3. Resultados e discussão

3.1. Estimativas dos valores genéticos moleculares

As estatísticas descritivas das estimativas dos valores genéticos moleculares (VGM) obtidos pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para as características ganho de peso pós-desmama (GPD), perímetro escrotal (PE) e escore de musculosidade (EM) estão apresentadas nas Tabelas 4 a 6.

Tabela 4. Estatísticas descritivas dos valores genéticos moleculares, estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para característica ganho de peso pós-desmama.

Método	N	MÉDIA	DP	CV	MIN	MAX
RMB	3.033	17,91	5,36	29,96	-1,00	34,27
RCB	3.033	6,63	2,92	44,07	-3,06	15,08
Bayes A	3.033	0,15	0,13	83,16	-0,29	0,49
Bayes B	3.033	0,01	0,01	162,00	-0,02	0,02
Bayes Cπ	3.033	6,64	2,89	43,61	-3,41	14,96
LASSO	3.033	15,28	4,86	31,81	-1,18	29,87

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor genético molecular; MAX: máximo valor genético molecular.

Tabela 5. Estatísticas descritivas dos valores genéticos moleculares, estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para característica perímetro escrotal.

Método	N	MÉDIA	DP	CV	MIN	MAX
RMB	2.664	2,45	0,76	30,96	-0,11	5,15
RCB	2.664	1,15	0,38	33,18	-0,13	2,48
Bayes A	2.664	0,83	0,28	33,58	-0,16	1,73
Bayes B	2.664	0,24	0,12	48,19	-0,09	0,45
Bayes Cπ	2.664	1,19	0,37	31,06	-0,09	2,47
LASSO	2.664	2,31	0,67	29,16	0,07	4,60

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor genético molecular; MAX: máximo valor genético molecular.

Tabela 6. Estatísticas descritivas dos valores genéticos moleculares, estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para característica escore de musculosidade.

Método	N	MÉDIA	DP	CV	MIN	MAX
RMB	3.149	1,15	0,25	21,50	0,35	1,97
RCB	3.149	0,44	0,13	29,03	0,01	0,87
Bayes A	3.149	0,68	0,17	25,26	0,11	1,24
Bayes B	3.149	0,14	0,04	31,46	-0,01	0,30
Bayes Cπ	3.149	0,44	0,12	27,29	0,04	0,83
LASSO	3.149	1,01	0,22	21,77	0,29	1,75

N: número de observações; DP: desvio padrão; CV: coeficiente de variação; MIN: mínimo valor genético molecular; MAX: máximo valor genético molecular.

Com base nas médias das estimativas dos valores genéticos moleculares, grandes similaridades foram observadas entre as estimativas fornecidas pelas metodologias RMB e LASSO, entre RCB e Bayes Cπ e entre Bayes A e Bayes B. Para melhor análise das semelhanças e divergências existentes entre as estimativas dos diferentes métodos, gráficos de dispersão entre os VGM fornecidos pelas metodologias RMB, RCB, Bayes A, Bayes B, Bayes Cπ e LASSO para as características GPD, PE e EM foram construídos e estão apresentados nas Figuras 1 a 3.

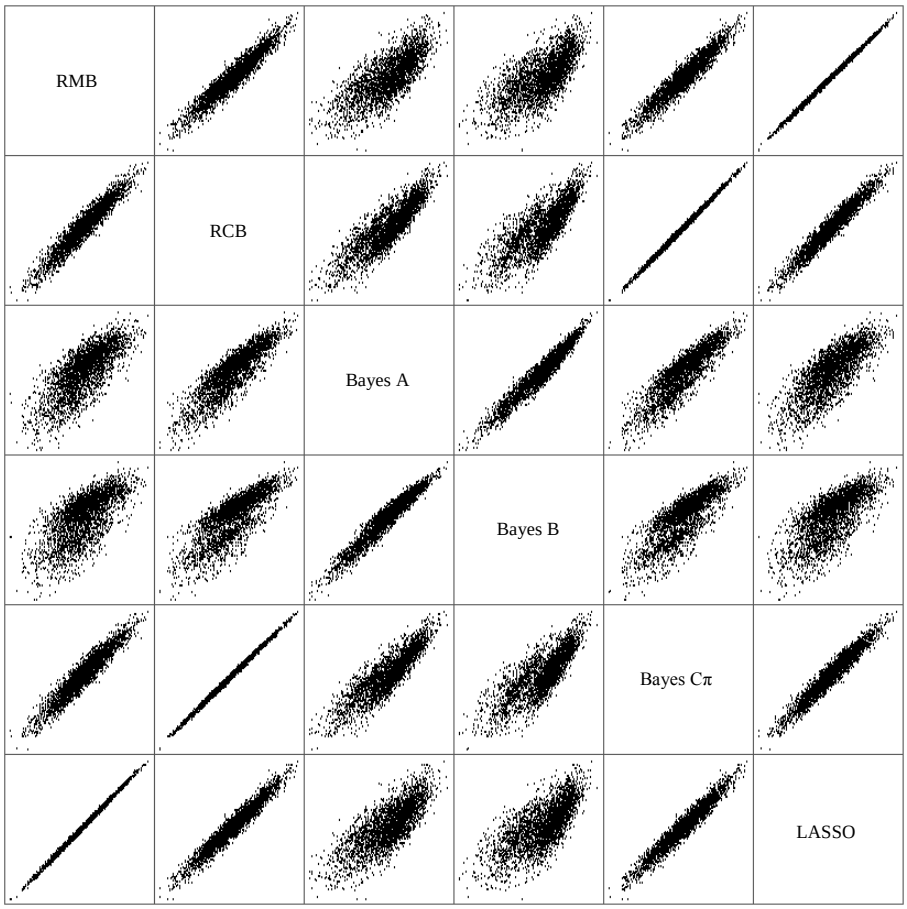


Figura 1. Gráficos de dispersão entre as estimativas de valor genético molecular, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o ganho de peso pós-desmama.

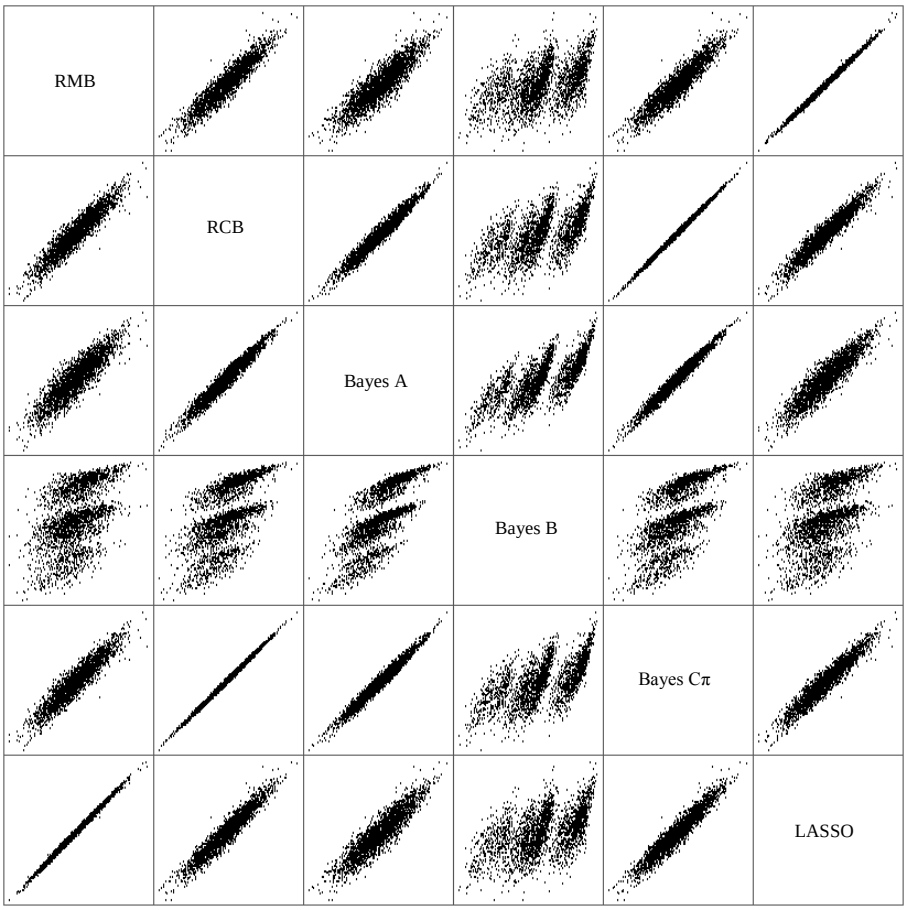


Figura 2. Gráficos de dispersão entre as estimativas de valor genético molecular, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o perímetro escrotal.

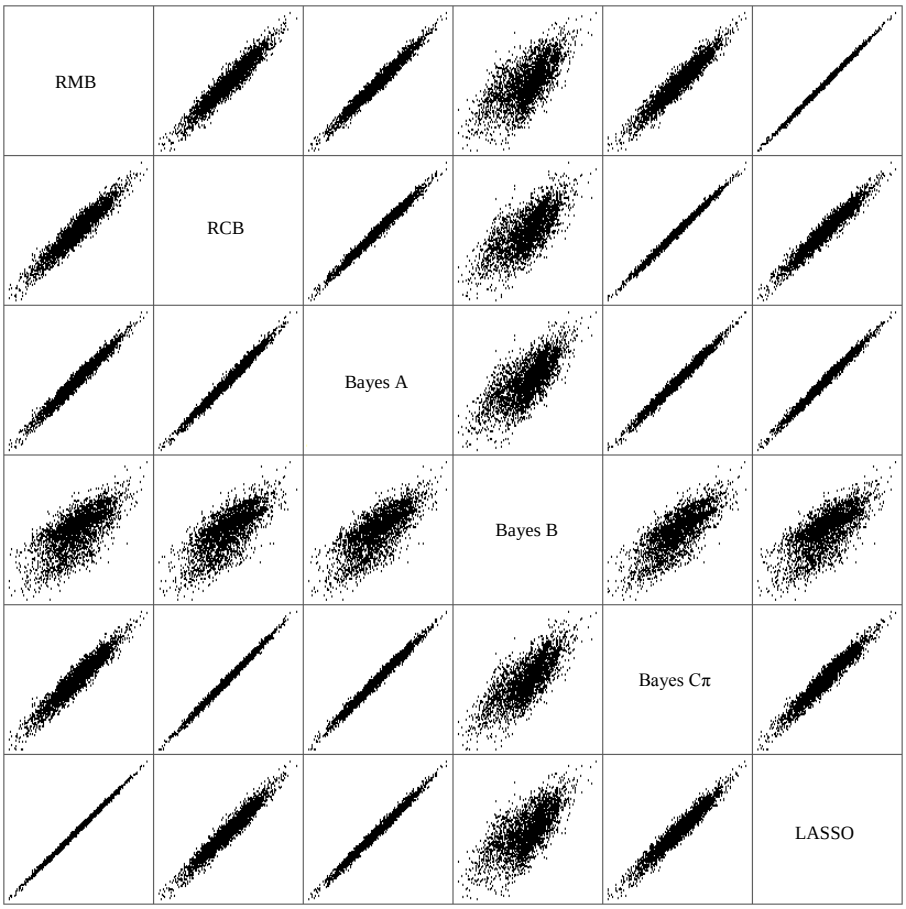


Figura 3. Gráficos de dispersão entre as estimativas de valor genético molecular, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o escore de musculosidade.

A análise dos gráficos de dispersão evidencia a alta similaridade entre as estimativas de VGM das metodologias RMB e LASSO e, RCB e Bayes Cπ pelas baixas dispersões dos pontos observadas entre elas, indicando elevada correlação linear entre os VGM fornecidos por essas metodologias, o que corrobora com os coeficientes de correlação apresentados nos Apêndices E, F e G. Tal fato pode ser explicado considerando a maneira como cada método regulariza as estimativas dos efeitos dos marcadores. Segundo De los Campos et al. (2009), a medida que o parâmetro λ, responsável pelo encurtamento promovida pelo método LASSO, se aproxima de zero, as estimativas fornecidas por esse método tendem às estimativas de mínimos quadrados, o que justificaria a semelhança observada entre RMB e LASSO, uma vez que a média a posteriori do λ, estimada no presente estudo, foi igual a 0,06 para as três características analisadas. Esse resultado sugere que todos os 106 marcadores analisados estão associados a expressão das características GP, PE e EM e que pequenos efeitos colineares existam entre eles, conforme indicado pelo $\lambda > 0$. Os mesmos autores relatam ainda que o

LASSO, em comparação à RCB, promove maior encolhimento nas estimativas dos efeitos que são próximos de zero e menor encolhimento nas estimativas dos efeitos com valores absolutos elevados. Analisando conjuntamente esse comportamento do método LASSO e as estatísticas descritivas dos VGM para RCB e LASSO, sugere-se que a maioria dos marcadores analisados apresenta importantes efeitos sobre a expressão do GPD, PE e EM, a ponto da metodologia não encolhê-los em direção à zero, como indicado pela superioridade das médias dos VGM do método LASSO. Isso pode ser explicado pelo fato dos marcadores analisados, em sua maioria, estarem localizados em regiões de transcrição ou promotoras de genes descritos na literatura por afetarem a expressão de características de interesse econômico.

Os métodos RCB e Bayes Cπ divergem quanto à pressuposição em relação à proporção de marcadores com efeito diferente de zero e na maneira como promovem o encurtamento das estimativas. Na RCB assume-se que todos os marcadores têm efeito e o encurtamento, por meio da razão constante entre a variância residual e a variância de cada marcador, leva à seleção automática dos marcadores com base nos dados, tendendo a zero, os que apresentam efeitos colineares. No método Bayes Cπ considera-se que uma fração (π) dos marcadores tem efeito, a qual é determinada com base nos dados, e para esses é assumida razão constante entre a variância residual e a variância de cada marcador. Entretanto, a correlação linear entre os VGM estimados por esses métodos, demonstrada nos gráficos de dispersão, sugere que ambos detectaram efeitos de mesma magnitude para os mesmos marcadores. Dado que, no presente estudo, as estimativas de π pelo método Bayes Cπ foram 80%, 83% e 76% acredita-se que 20%, 17% e 24% dos marcadores apresentam efeitos colineares para GP, PE e EM, respectivamente, corroborando com os resultados descritos anteriormente para LASSO com $\lambda > 0$, que sugeriam a existência de colinearidade entre os efeitos dos marcadores.

A similaridade entre as estimativas dos métodos Bayes A e Bayes B sugerida pela proximidade das médias dos VGM não foi comprovada graficamente, tendo em vista a grande dispersão dos pontos entre eles, principalmente para as características PE e EM. As maiores divergências entre os VGM foram observadas para a metodologia Bayes B, dada a alta dispersão dos pontos para esse método nas três características analisadas, indicando baixa correlação linear entre essas estimativas e as fornecidas pelos demais métodos. Segundo Gianola et al. (2009) as estimativas fornecidas pelos métodos Bayes A e Bayes B são muito influenciadas pelos hiperparâmetros (ν, s^2) da distribuição a priori assumida para a variância

dos marcadores, limitando o aprendizado baseado nos dados, o que justificaria as divergências observadas em relação às estimativas dos outros métodos, que não apresentam essa limitação.

3.2. Variância aditiva, herdabilidade e correlação genética

As estimativas dos componentes de variância aditiva, dos coeficientes de herdabilidade e das correlações genéticas estimadas em análises bicaracterísticas entre o fenótipo observado e os valores genéticos moleculares, obtidos pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para as características ganho de peso pós-desmana (GPD), perímetro escrotal (PE) e escore de musculosidade (EM) estão apresentadas nas Tabelas 7 a 9.

Tabela 7. Estimativas das variâncias aditivas ($\hat{\sigma}_a^2$) e dos coeficientes de herdabilidade (h^2) para o ganho de peso pós-desmama e das correlações genéticas entre o fenótipo e o valor genético molecular ($r_{P,VGM}$), obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

Método	$\hat{\sigma}_a^2$	h^2	$r_{P,VGM}$
RMB	87,23	0,21	0,44
RCB	87,16	0,21	0,42
Bayes A	118,86	0,27	0,73
Bayes B	120,76	0,24	0,98
Bayes Cπ	87,18	0,21	0,42
LASSO	87,24	0,21	0,44

Tabela 8. Estimativas das variâncias aditivas ($\hat{\sigma}_a^2$) e dos coeficientes de herdabilidade (h^2) para o perímetro escrotal e das correlações genéticas entre o fenótipo e o valor genético molecular ($r_{P,VGM}$), obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

Método	$\hat{\sigma}_a^2$	h^2	$r_{P,VGM}$
RMB	3,57	0,51	0,40
RCB	3,56	0,51	0,35
Bayes A	3,52	0,39	0,41
Bayes B	5,14	0,63	0,84
Bayes Cπ	3,68	0,46	0,28
LASSO	3,57	0,51	0,39

Tabela 9. Estimativas das variâncias aditivas ($\hat{\sigma}_a^2$) e dos coeficientes de herdabilidade (h^2) para o escore de musculosidade e das correlações genéticas entre o fenótipo e o valor genético molecular ($r_{P,VGM}$), obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

Método	$\hat{\sigma}_a^2$	h^2	$r_{P,VGM}$
RMB	0,14	0,22	0,41
RCB	0,12	0,19	0,46
Bayes A	0,14	0,23	0,38
Bayes B	0,16	0,24	0,29
Bayes Cπ	0,12	0,19	0,46
LASSO	0,14	0,22	0,41

As estimativas dos coeficientes de herdabilidade obtidas para GPD, PE e EM apresentaram pequenas variações em virtude do método utilizado para a estimação dos VGM e têm magnitudes semelhantes àquelas obtidas nas análises unicaracterística realizadas por meio do *software* MTDFREML utilizando o banco de dados completo, que foram 0,21, 0,50 e 0,21, respectivamente. Além disso, as estimativas de herdabilidade apresentadas foram condizentes com as observadas na última avaliação genética realizada pelo GMAB-FZEA para esse rebanho, as quais foram 0,32, 0,45 e 0,25 para GPD, PE e EM, respectivamente (comunicação pessoal). Horimoto et al. (2007) e Van Melis et al. (2010) analisando outras amostras dessa população relataram estimativas de herdabilidade similares para GPD (0,32 e 0,25), PE (0,55 e 0,48) e EM (0,16 e 0,23).

3.3. Impacto na acurácia

As alterações percentuais médias na acurácia das estimativas dos valores genéticos para cada uma das cinco categorias de animais ao serem incorporadas as informações dos marcadores moleculares na avaliação genética das características GPD, PE e EM estão apresentadas nas Figuras 4 a 6.

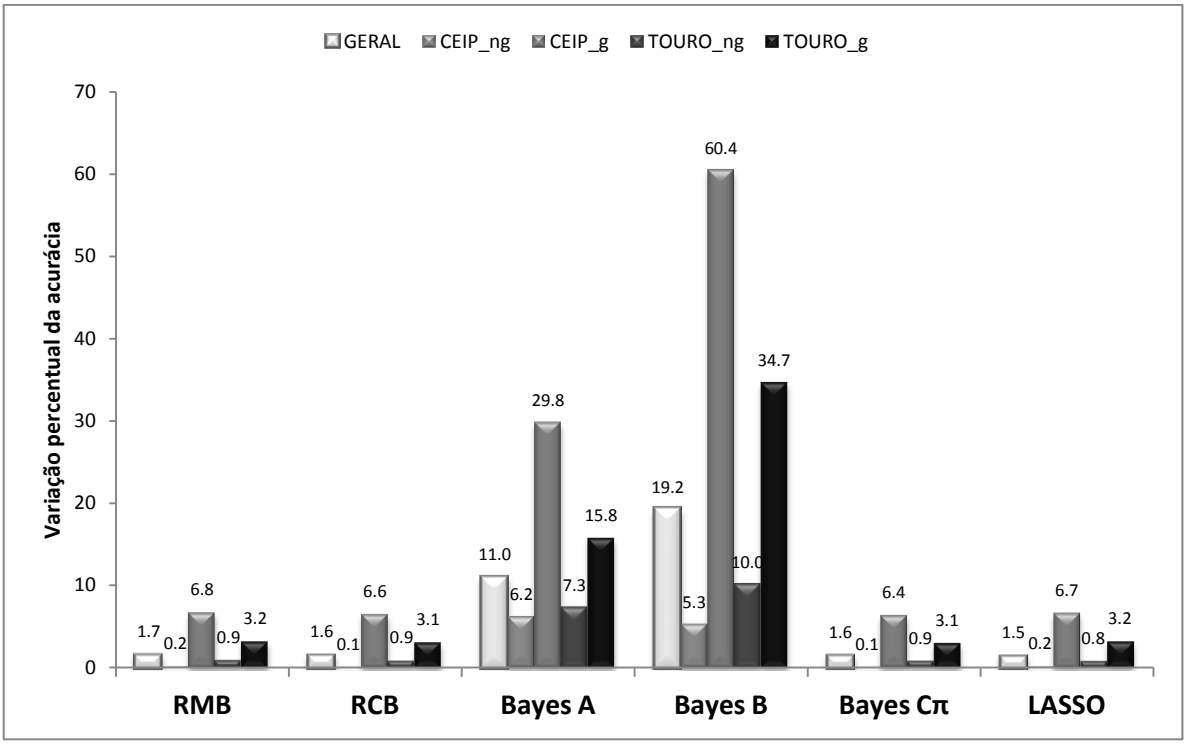


Figura 4. Variações percentuais médias na acurácia das estimativas de valores genéticos para o ganho de peso pós-desmama devido à inclusão dos efeitos dos marcadores estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

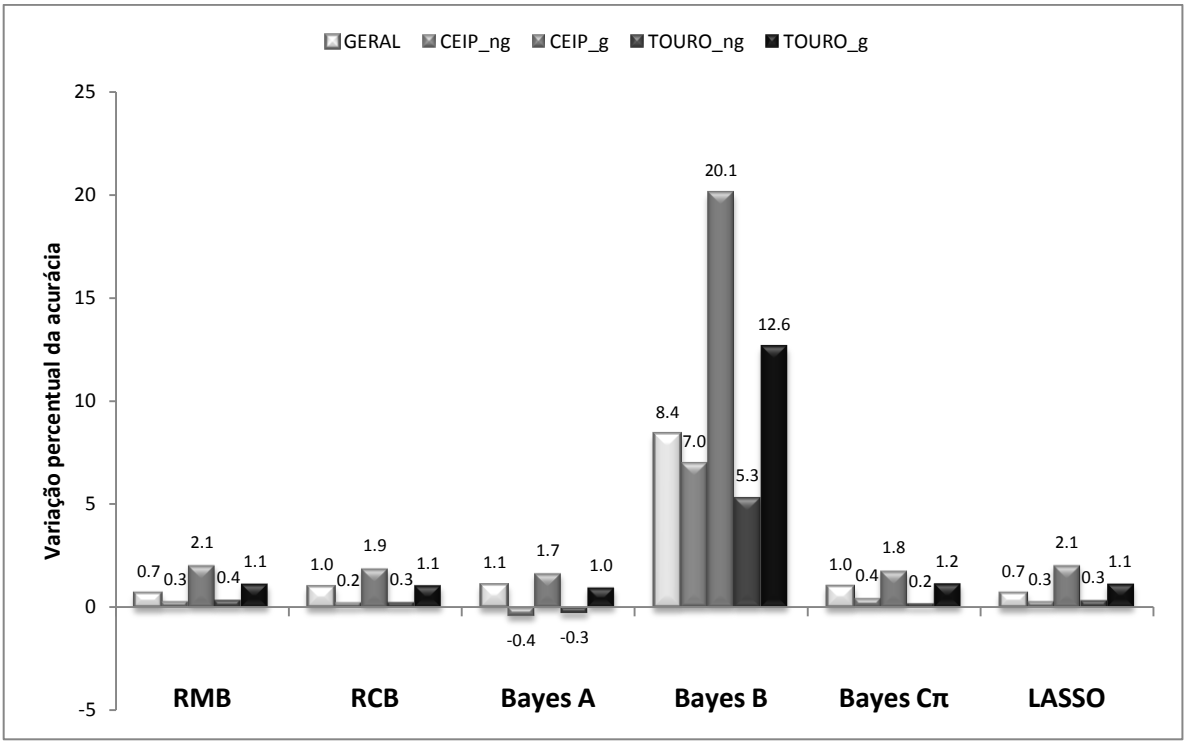


Figura 5. Variações percentuais médias na acurácia das estimativas de valores genéticos para o perímetro escrotal pela inclusão dos efeitos dos marcadores estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

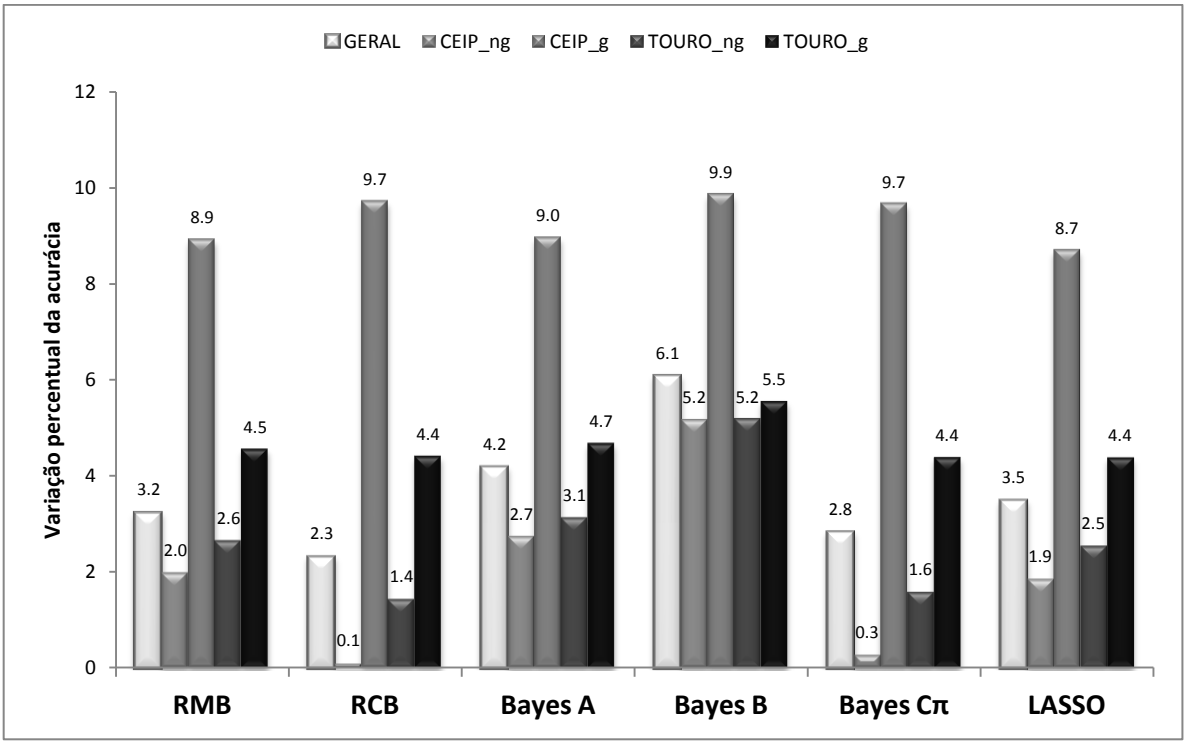


Figura 6. Variações percentuais médias na acurácia das estimativas de valores genéticos para o escore de musculosidade pela inclusão dos efeitos dos marcadores estimados pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO).

Independente da metodologia utilizada para a estimação dos efeitos dos marcadores, incrementos nas acurácias dos valores genéticos foram observados para as cinco categorias de animais analisadas. Os maiores impactos foram observados nos animais genotipados, sendo os incrementos mais elevados para os animais jovens, no caso, os tourinhos de reposição. As variações do impacto da inclusão dos VGM, estimados pelas diferentes metodologias dentro de cada característica, podem ser explicadas pela correlação genética estimada entre o fenótipo observado e o VGM (Tabelas 7 a 9), nas quais as maiores alterações se devem às maiores correlações genética entre eles. Embora as alterações sejam muito variáveis entre as diferentes categorias de animais, esses resultados demonstram a potencialidade da incorporação da informação dos marcadores nas avaliações genéticas. Em termos práticos, o aumento da acurácia leva ao aumento do ganho genético por promover o encurtamento do intervalo de gerações, já que os animais jovens passam a ser utilizados como reprodutor precocemente, e pelo aumento acurácia em si, que apresenta relação proporcional com ganho genético.

3.4. Conflitos de seleção

Nas Figuras 7 a 9 estão apresentadas as dispersões entre os valores genéticos clássicos (VG) e os valores genéticos assistidos por marcadores (VG_{AM}) estimados considerando os valores genéticos moleculares obtidos com base na metodologia LASSO bayesiano, para as características GPD, PE e EM. Nessas figuras, estão indicados os pontos de truncamento para a seleção dos 20% melhores animais em cada critério de classificação com base em todo rebanho analisado, a exemplo do que é permitido para os programas autorizados pelo MAPA a emitir CEIP.

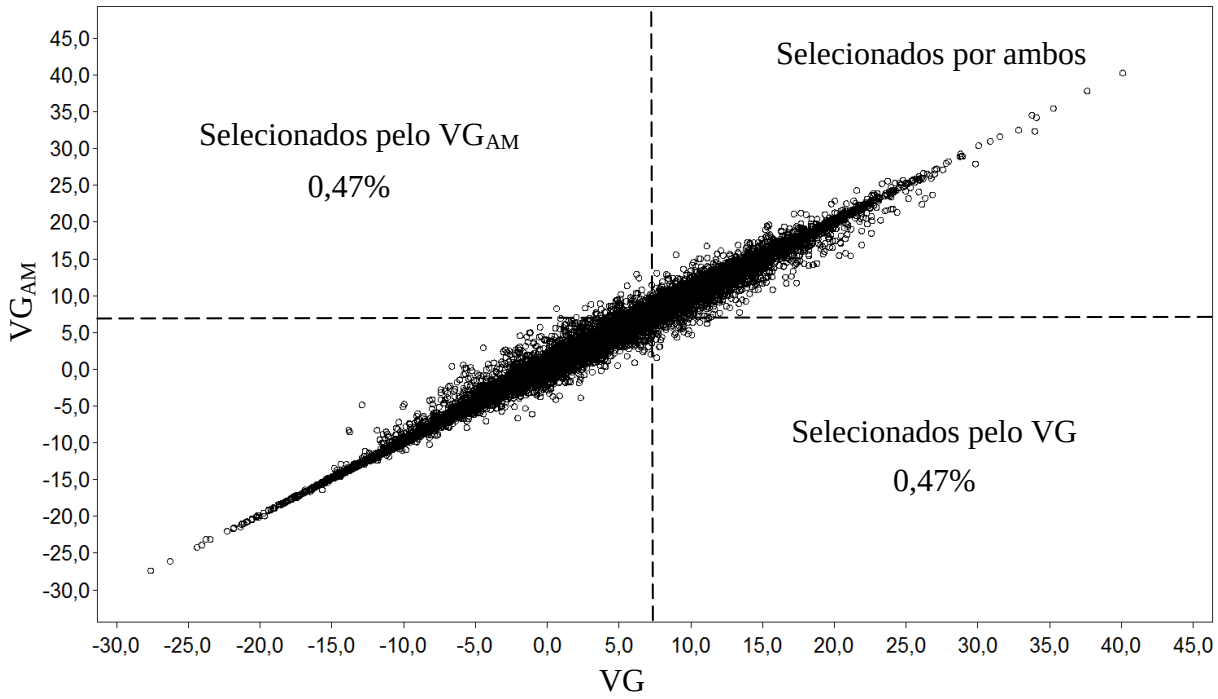


Figura 7. Dispersões dos valores genéticos clássicos (VG) e valores genéticos assistidos por marcadores (VG_{AM}) estimados com base na metodologia LASSO bayesiano para o ganho de peso pós-desmama (linha pontilhada indica ponto de truncamento dos 20% melhores animais).

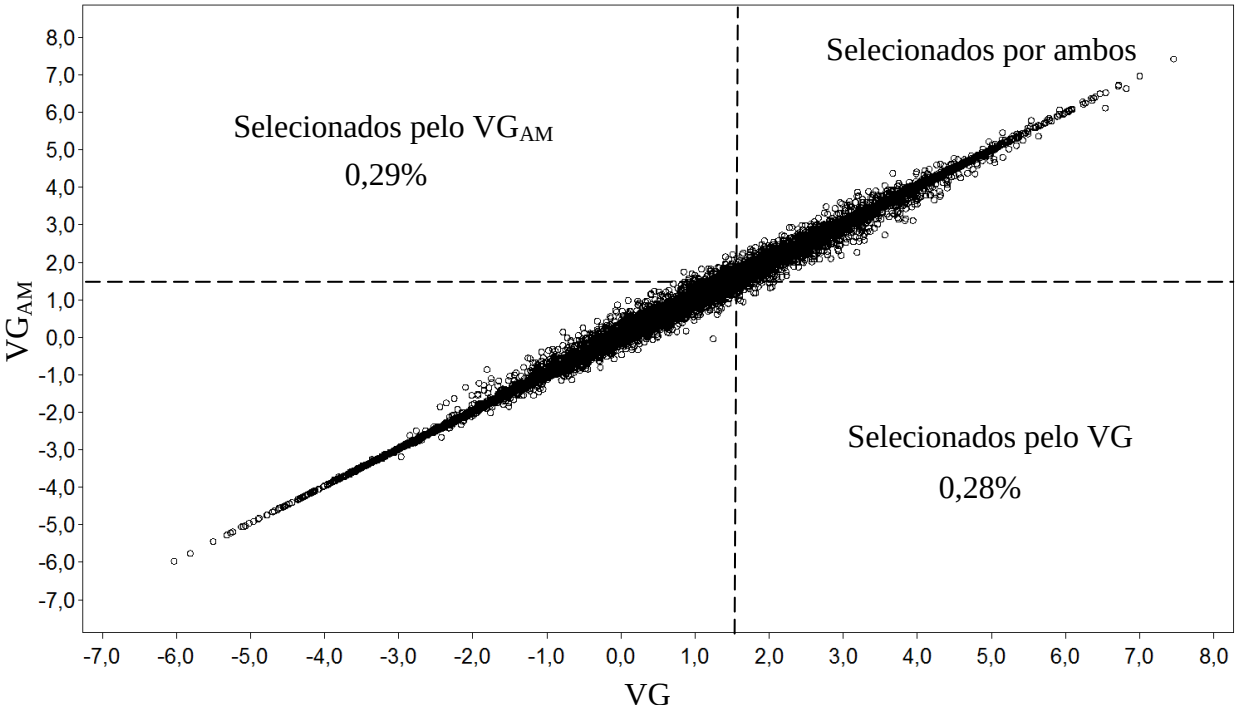


Figura 8. Dispersões dos valores genéticos clássicos (VG) e valores genéticos assistidos por marcadores (VG_{AM}) estimados com base na metodologia LASSO bayesiano para o perímetro escrotal (linha pontilhada indica ponto de truncamento dos 20% melhores animais).

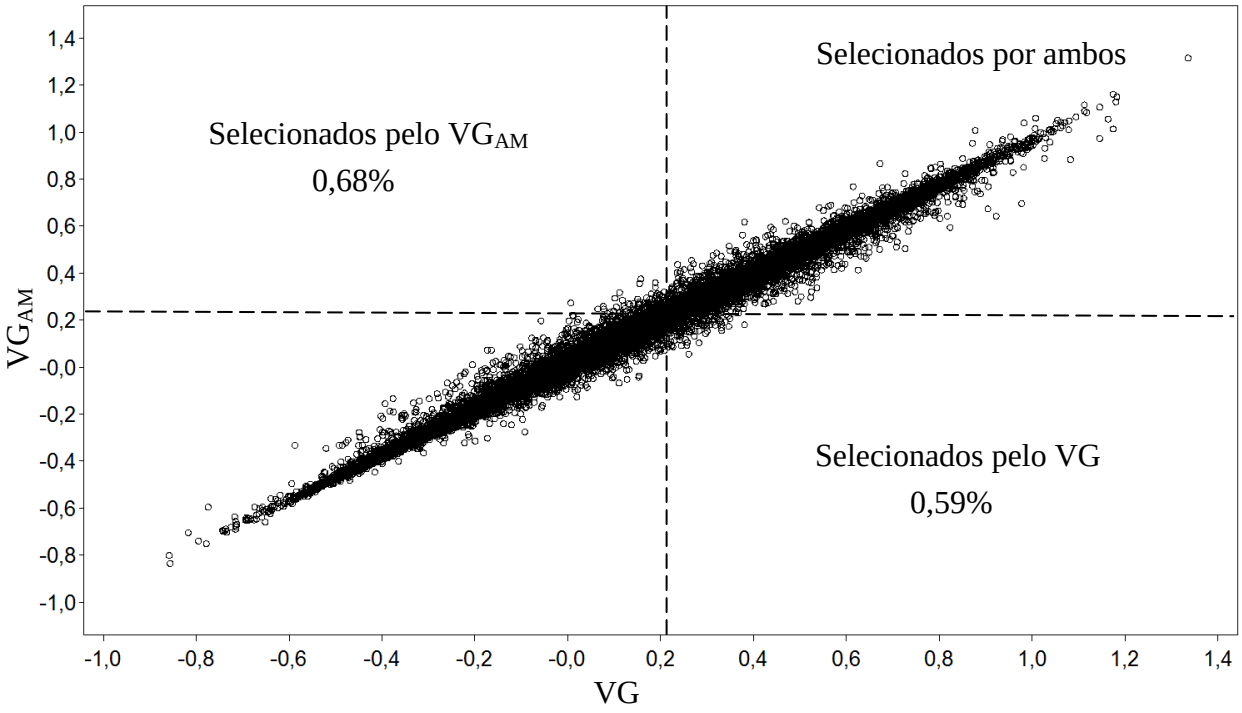


Figura 9. Dispersões dos valores genéticos clássicos (VG) e valores genéticos assistidos por marcadores (VG_{AM}) estimados com base na metodologia LASSO bayesiano para o escore de musculosidade (linha pontilhada indica ponto de truncamento dos 20% melhores animais).

Os conflitos de seleção são indicados pelas divergências nas classificações dos animais, ou seja, quando um animal é selecionado por um critério e descartado por outro. Os maiores conflitos foram observados para o escore de musculosidade, com 1,27% dos animais sendo classificados por apenas um dos critérios como pertencentes ao grupo dos 20% melhores avaliados.

Esse mesmo enfoque pode ser aplicado para avaliar os conflitos de seleção dentro das categorias de animais, conforme disposto na Tabela 10 para o escore de musculosidade.

Tabela 10. Conflitos de seleção nas categorias de animais classificados pelo valor genético clássico e pelo valor genético assistido por marcadores estimados para o escore de musculosidade.

Categoria	N	VG (%)	VG_{AM} (%)	CONFLITO (%)
GERAL	113.301	0,59	0,68	1,27
CEIP_ng	2.744	0,73	0,76	1,49
CEIP_g	1.374	2,76	2,69	5,45
TOURO_ng	1.103	0,82	0,91	1,73
TOURO_g	296	2,36	2,36	4,72

N: número de animais avaliados; VG: proporção de animais selecionados pelo valor genético clássico e não pelo valor genético assistido por marcadores; VG_{AM}: proporção de animais selecionados pelo valor genético assistido por marcadores e não pelo valor genético clássico; CONFLITO: proporção total de animais que apresentaram conflitos de seleção.

A análise dos conflitos de seleção dentro das categorias evidenciou que as maiores discrepâncias entre as classificações dos animais com base no VG e no VG_{AM} ocorrem para as categorias de animais genotipados, sendo superior nos animais jovens. Considerando que nos programas de melhoramento genético autorizados a emitir CEIP, pouco mais de 20% de todos os animais avaliados são considerados geneticamente superiores e aptos à certificação, conflitos de seleção ao redor de 5% são muito importantes, principalmente por ocorrerem em categorias de animais responsáveis pela disseminação de material genético em ampla escala, como são os touros e tourinhos de reposição. Dados os conflitos de seleção observados para todas as categorias, indica-se que a seleção dos animais seja realizada com base nos valores genéticos assistidos por marcadores, uma vez que esses apresentam estimativas mais acuradas do que os valores genéticos estimados pelas avaliações genéticas clássicas, conforme demonstrado anteriormente nos gráficos de impacto na acurácia.

4. Conclusão

As diferenças nas estimativas dos efeitos dos marcadores e, por consequência, nos valores genéticos moleculares, em função da metodologia utilizada (regressão múltipla bayesiana, regressão de cumeeira bayesiana, Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano), ocorreram devido à maneira como cada método regulariza as estimativas dos efeitos dos marcadores.

A incorporação das informações dos marcadores nas avaliações genéticas possibilitou a obtenção de estimativas de valores genéticos, no geral, mais acuradas para todas as categorias de animais, com os maiores impactos sendo observados para os tourinhos jovens de reposição, que foram genotipados.

As maiores divergências na seleção dos 20% melhores animais, ao serem classificados com base no valor genético clássico e no valor genético assistido por marcadores, foram observadas para os touros e tourinhos genotipados. Embora esses resultados tenham sido obtidos para um pequeno conjunto de marcadores, eles demonstram a potencialidade do uso da informação molecular nos programas de melhoramento genético, possibilitando a tomada de decisões mais seguras em relação à escolha de animais de maior mérito genético.

5. Referências bibliográficas

AGENCIA BRASIL. 2012. **Com US$ 94,6 bilhões em exportações, agronegócio tem melhor resultado desde 1997**. Disponível em: http://agenciabrasil.ebc.com.br/noticia/2012-01-10/com-us-946-bilhoes-em-exportacoes-agronegocio-tem-melhor-resultado-desde-1997. Acesso em 15 abr. 2012.

AYRES, D.R. et al. 2010. Evaluation of TFAM and FABP4 gene polymorphisms in three lines of Nellore cattle selected for growth. **Genetics and Molecular Research**, 9(4):2050-2059.

BLASCO, A.; PILES, M.; VARONA, L. 2003. A Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits. **Genetics Selection Evolution**, 35:21-41.

BOLDMAN, K.G. et al. 1995. **A manual for use of MTDFREML:** a set of programs to obtain estimates of variances and covariances. Nebraska: United States Department of Agriculture-Agricultural Research Service. 115p.

CEPEA-USP. 2012. **PIB Agro CEPEA-USP/CNA**. Disponível em: http://www.cepea.esalq.usp.br/pib/. Acesso em 15 abr. 2012.

CNA/CEPEA. 2012. **Análise do PIB das cadeias produtivas de algodão, cana-de-açúcar, soja, pecuária de corte e de leite no Brasil**. Disponível em: http://www.canaldoprodutor.com.br/sites/default/files/pib-cadeias-produtivas-web.pdf. Acesso em 15 abr. 2012.

DE LOS CAMPOS, G. et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, 182:375-385.

DEKKERS, J.C.M.; HOSPITAL, F. 2002. The use of molecular genetics in the improvement of agricultural populations. **Nature Reviews: Genetics**, 3:22-32.

FERNANDO, R.L.; GROSSMAN, M. 1989. Marker-assisted selection using best linear unbiased prediction. **Genetics Selection Evolution**, 21:467-477.

FERRAZ, J.B.S. et al. 2009. Association of single nucleotide polymorphisms with carcass traits in Nellore cattle. **Genetics and Molecular Research**, 8(4):1360-1366.

GARRICK, D.J.; SAATCHI, M. 2011. Opportunities and challenges for genomic selection of beef cattle. **Revista Brasileira de Zootecnia**, 40:310-316.

GELDERMANN, H. 1975. Investigation on inheritance of quantitative characters in animals by gene markers. 1. Methods. **Theoretical and Applied Genetics**, 46:319-33.

GEORGE, A.; LIU, J.; NG, E. 1980. **User guide for SPARSPACK:** Waterloo sparse linear equations package. CS-78-30, Dept. Computer Sci., Univ. Wateerloo, ON, Canada.

GEWEKE, J. 1992. **Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments** (with discussion). Pages 169-193 in Bayesian Statistics , Bernardo, J.M.; Berger, J.O.; Dawid, A.P.; Smith, A.F.M., ed. Oxford Univ. Press, Oxford, GB.

GEYER, C.J. 1992. Practical Markov chain Monte Carlo. **Statistical Science**. 7:473-511.

GIANOLA, D. et al. 2009. Additive genetic variability and the bayesian alphabet. **Genetics**, 183:347-363.

GIANOLA, D.; FERNANDO, R.L. 1986. Bayesian methods in animal breeding theory. **Journal of animal Science**, 63:217-244.

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

GIANOLA, D.; PEREZ-ENCISO, M.; TORO, M.A. 2003. On Marker-Assisted Prediction of Genetic Value: Beyond the Ridge. **Genetics**, 163:347-365

GIANOLA, D.; VAN KAAM, J.B.C.H.M. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, 178:2289-2303.

GODDARD, M.E. 1992. A mixed model for analyses of data on multiple genetic markers. **Theoretical and Applied Genetics**, 83:878-886.

GODDARD, M.E.; HAYES, B.J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews: Genetics**, 10:381-391.

GONZÁLEZ-RECIO, O. et al. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, 178: 2305-2313.

GUILLAUME, F. et al. 2008. Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. **Genetics Selection Evolution**, 40:91-102.

HABIER, D. et al. 2011. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, 12:186.

HABIER, D.; FERNANDO, R.L.; DEKKERS, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, 177:2389-2397.

HALEY, C.S.; VISSCHER, P.M. 1998. Strategies to utilize marke-quantitative trait loci associations. **Jounal of Dairy Science**, 81: 85-97 (Suppl. 2).

HENDERSON, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, 31:423.

HORIMOTO, A.R.V.R. et al. 2007. Phenotypic and genetic correlations for body structure scores (frame) with productive traits and index for CEIP classification in Nellore beef cattle. **Genetics and Molecular Research**, 6(1):188-196.

IBGE 2011. **PPM 2010:** Rebanho bovino nacional cresce 2,1% e chega a 209,5 milhões de cabeças. Disponível em: http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=2002&id_pagina=1. Acesso em 15 abr. 2012.

IBGE 2012a. **Em 2011, PIB cresce 2,7% e totaliza R$ 4,143 trilhões**. Disponível em: http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=2093. Acesso em 15 abr. 2012.

IBGE 2012b. **Abate de animais, produção de leite, couro e ovos**. Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/producaoagropecuaria/abate-leite-couro-ovos_201104_1.shtm. Acesso em 15 abr. 2012.

KENDALL, M.G. 1947. **The advanced theory of statistics**. London: Charles Griffin & Company Limited. 457p.

LANDE, R.; THOMPSON, R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, 124:743-756.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. 2009. A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, 92:4656-4663.

LEGARRA, A.; VARONA, L.; LÓPEZ DE MATURANA, E. 2008. **TM:** Threshold Model. Disponível em: http://snp.toulouse.inra.fr/~alegarra.

LONG, N. et al. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, 124:377-389.

MEUWISSEN, T.H.E. et al. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. **Genetics Selection Evolution**, 41:2.

MEUWISSEN, T.H.E.; GODDARD, M.E. 1996. The use of marker haplotypes in animal breeding schemes. **Genetics Selection Evolution**, 28:161-176.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157:1819-1829.

NEIMANN-SORENSEN, A.; ROBERSON, A. 1961.The association between blood groups and several production characters in three Danish cattle breeds. **Acta Agriculturae Scandinavica**, 11:163-196.

NOGUEIRA, M.P. **Perspectivas para o confinamento em 2012**. Disponível em: http://www.beefpoint.com.br/cadeia-produtiva/espaco-aberto/perspectivas-para-o-confinamento-em-2012/. Acesso em 15 abr. 2012.

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

PEDROSA, V.B. **Utilização de diferentes metodologias para a avaliação genética de bovinos de corte**. 2011. 87 f. Tese (Doutorado) - Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2011.

PEREZ-ENCISO, M.; MISZTAL, I.; ELZO, M.A. 1994. FSPAK: An interface for public domain sparse matrix subroutines. **Proceedings of 5th World Congress of Genetics Applied to Livestock Production**, 22:87-88.

PINTO, L.F.B. et al. 2010. Association of SNPs on CAPN1 and CAST genes with tenderness in Nellore cattle. **Genetics and Molecular Research**, 9(3):1431-1442.

QUAAS, R.L. 1976. Computing the diagonal elements and inverse os a large numerator relationship matrix. **Biometrics**, 32:949-953.

SMITH, S. P.; GRASER, H.U. 1986. Estimating variance components in a class of models by restricted maximum likelihood. **Journal of Dairy Science,** 69:1156-1165.

SOLBERG, T.R. et al. 2009. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, 41:29.

SOLLER, M. 1978. The use of loci associated with quantitative effects in dairy cattle improvement. **Animal Production**, 27: 133-139.

SOLLER, M.; BECKMANN, J.S. 1983. Genetic polymorphisms in varietal identification and genetic improvement. **Theoretical and Applied Genetics**, 67: 25-33.

SPIEGELHALTER, D.J. et al.2002. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistic Society**, Series B. 64:583-639.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistic Society**, 58(1):167-288.

VAN ARENDONK, J.A.M.; TIER, B.; KINGHORN, B.P. 1994. Use of multiple genetic markers in prediction of breeding values. **Genetics**, 137:319-329.

VAN MELIS, M.H. et al. 2010. Additive genetic relationship of longevity with fertility and production traits in Nellore cattle based on bivariate models. **Genetics and Molecular Research**, 9(1):176-187.

VAN RADEN, P.M. 2008. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91:4414-4423.

VAN RADEN, P.M. et al. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, 92:16-24.

WHITTAKER, J.C.; THOMPSON, R.; DENHAM, M.C. 2000. Marker-assisted selection using ridge regression. **Genetics Research**, 75: 249-252

WOLC, A. et al. 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. **Genetics Selection Evolution**, 43:5.

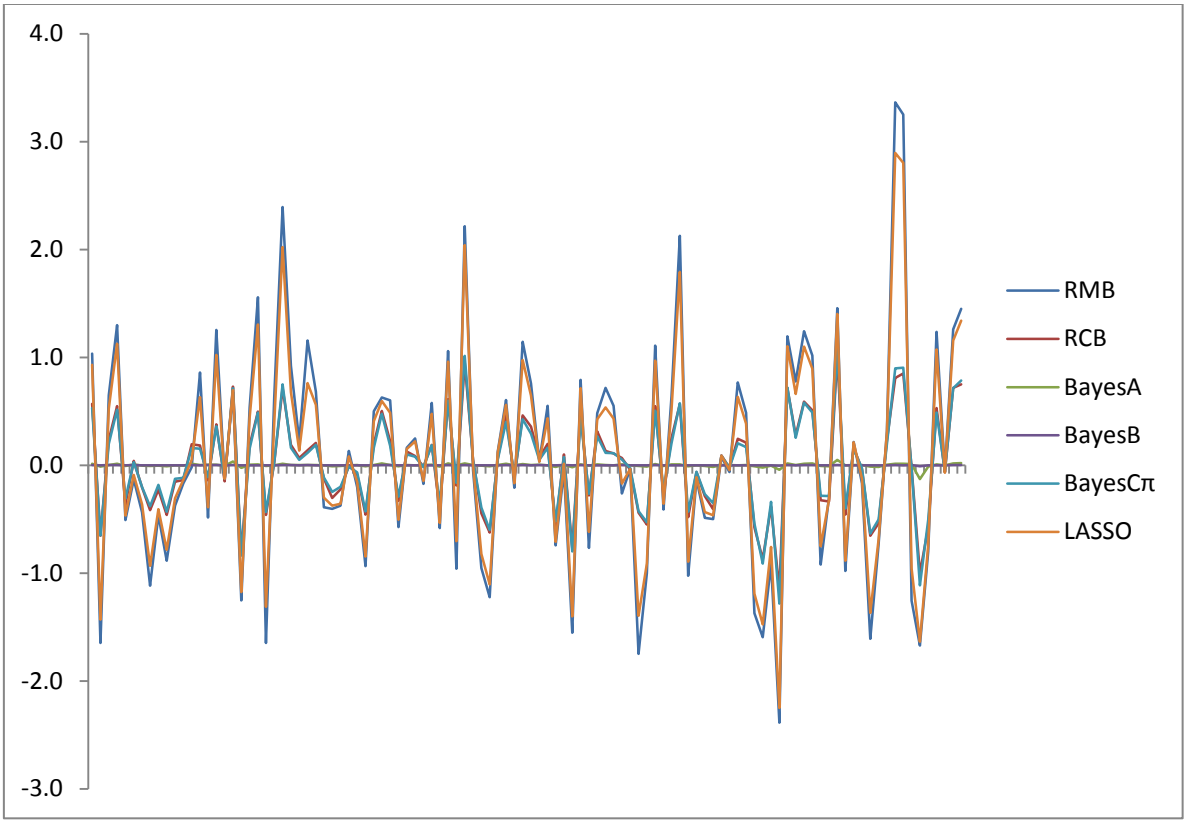
XU, S. 2007. An empirical bayes method for estimating epistatic effects of quantitative trait loci. **Biometrics**, 63:513-521.

6. Conclusões e perspectivas da tese

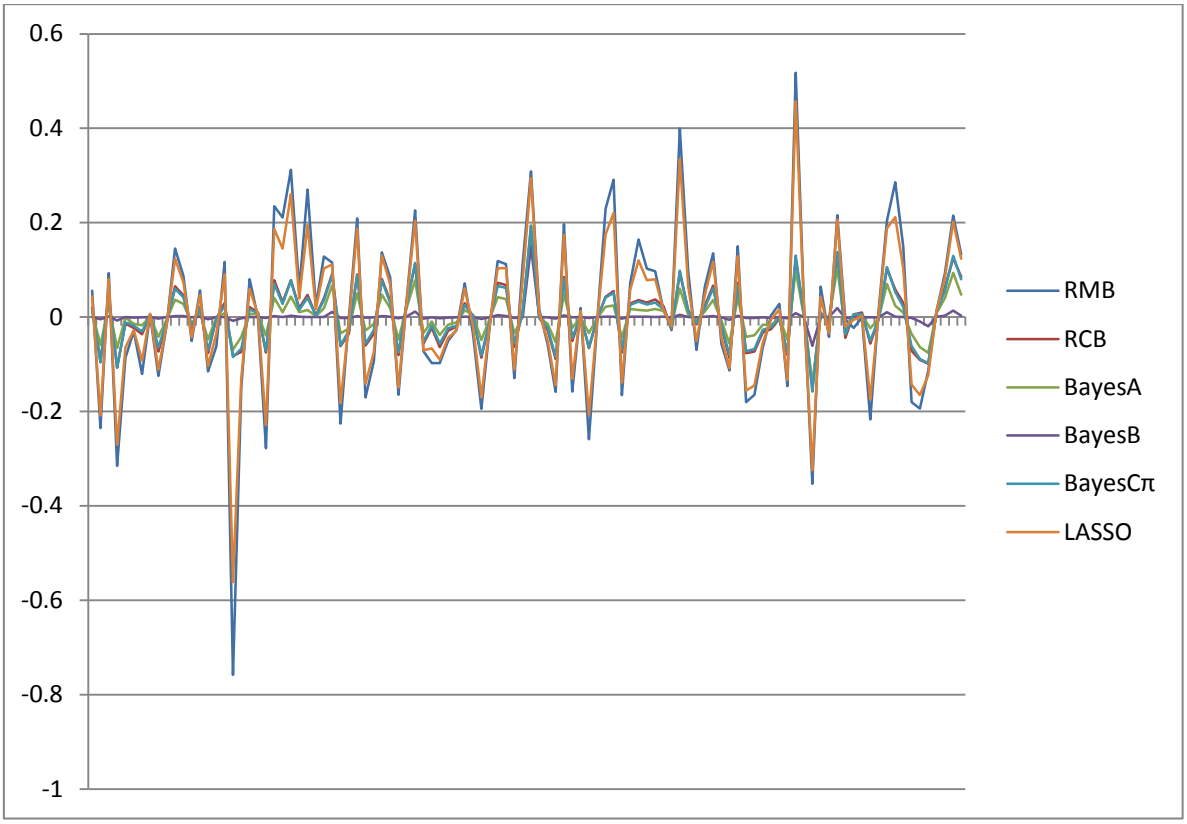
Como conclusão geral tem-se que mesmo para painéis de marcadores do tipo SNP de muito baixa densidade, a inclusão da informação molecular nas avaliações genéticas para a estimação dos valores genéticos dos animais contribui para o aumento da acurácia dessas estimativas, sendo, portanto, uma tecnologia muito promissora para aumentar o ganho genético nos programas de melhoramento animal, merecendo por isso contínuos estudos para seu aperfeiçoamento.

APÊNDICES

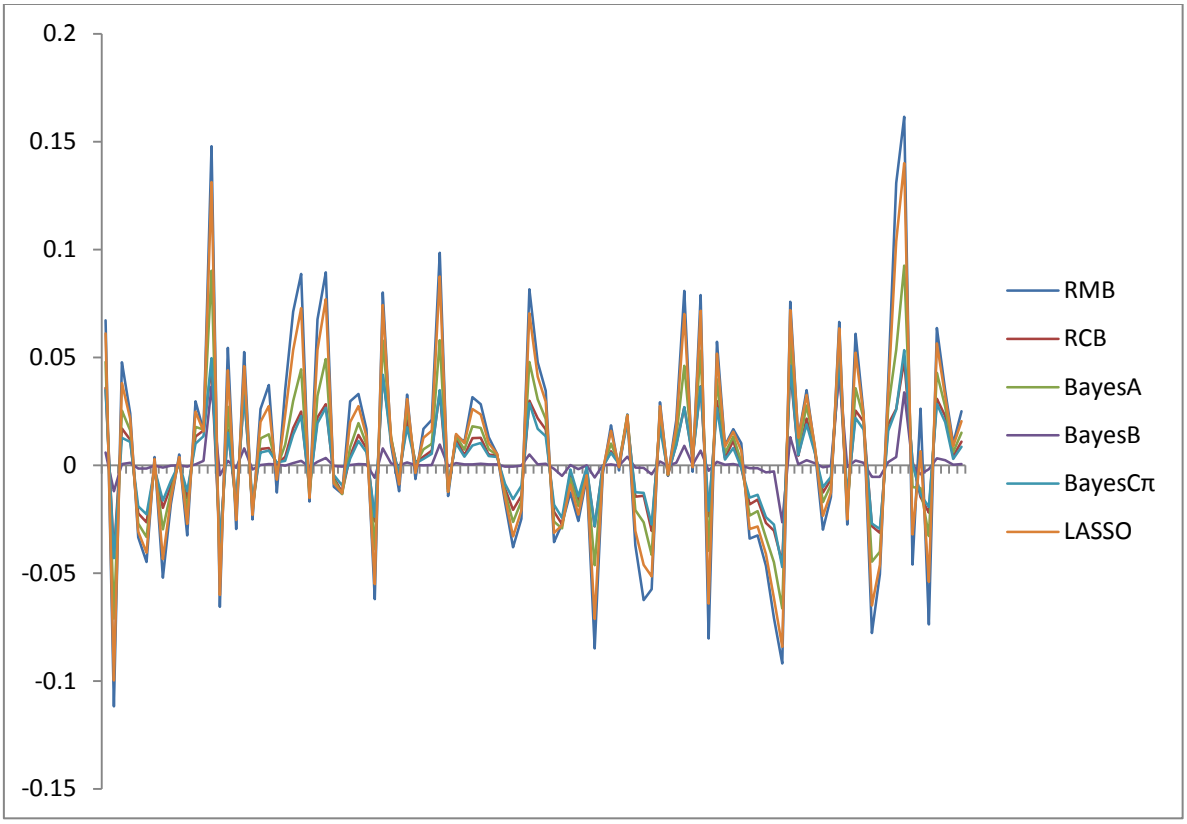
APÊNDICE A – Gráfico do tipo linha das estimativas dos efeitos dos marcadores obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o ganho de peso pós-desmama.



APÊNDICE B – Gráfico do tipo linha das estimativas dos efeitos dos marcadores obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o perímetro escrotal.



APÊNDICE C – Gráfico do tipo linha das estimativas dos efeitos dos marcadores obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o escore de musculosidade.



APÊNDICE D – Estimativas dos coeficientes de correlação de Pearson (acima da diagonal) e de Spearman (abaixo da diagonal) entre as estimativas dos valores genéticos moleculares, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o ganho de peso pós-desmama.

Método	RMB	RCB	Bayes A	Bayes B	Bayes Cπ	LASSO
RMB		0,95	0,75	0,68	0,95	1,00
RCB	0,95		0,87	0,79	1,00	0,97
Bayes A	0,75	0,87		0,97	0,88	0,77
Bayes B	0,69	0,80	0,96		0,81	0,71
Bayes Cπ	0,95	1,00	0,88	0,81		0,97
LASSO	1,00	0,97	0,77	0,72	0,96	

APÊNDICE E – Estimativas dos coeficientes de correlação de Pearson (acima da diagonal) e de Spearman (abaixo da diagonal) entre as estimativas dos valores genéticos moleculares, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o perímetro escrotal.

Método	RMB	RCB	Bayes A	Bayes B	Bayes Cπ	LASSO
RMB		0,94	0,88	0,50	0,93	1,00
RCB	0,93		0,97	0,59	1,00	0,96
Bayes A	0,88	0,97		0,75	0,98	0,91
Bayes B	0,51	0,61	0,76		0,63	0,53
Bayes Cπ	0,93	1,00	0,98	0,64		0,96
LASSO	1,00	0,95	0,91	0,54	0,95	

APÊNDICE F – Estimativas dos coeficientes de correlação de Pearson (acima da diagonal) e de Spearman (abaixo da diagonal) entre as estimativas dos valores genéticos moleculares, obtidas pelas metodologias regressão múltipla bayesiana (RMB), regressão de cumeeira bayesiana (RCB), Bayes A, Bayes B, Bayes Cπ e LASSO bayesiano (LASSO) para o escore de musculosidade.

Método	RMB	RCB	Bayes A	Bayes B	Bayes Cπ	LASSO
RMB		0,95	0,98	0,69	0,95	1,00
RCB	0,95		0,99	0,73	1,00	0,97
Bayes A	0,98	0,99		0,74	0,99	0,99
Bayes B	0,69	0,72	0,74		0,77	0,71
Bayes Cπ	0,95	1,00	0,99	0,77		0,97
LASSO	1,00	0,97	0,99	0,71	0,97	