

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS DE JABOTICABAL

**SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS
LONGITUDINAIS DE BOVINOS DA RAÇA HOLANDESA**

Daiane Cristina Becker Scalez

Zootecnista

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS DE JABOTICABAL

**SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS
LONGITUDINAIS DE BOVINOS DA RAÇA HOLANDESA**

Daiane Cristina Becker Scalez

Orientadora: Profa. Dra. Sandra Aidar de Queiroz

Co-orientador: Dr. Haroldo Henrique de Rezende Neves

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – UNESP, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Zootecnia.

S281s Scalez, Daiane Cristina Becker
Seleção genômica para características longitudinais de bovinos da raça Holandesa. / Daiane Cristina Becker Scalez. -- Jaboticabal, 2016
xi, 52 p. ; 28 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2016
Orientador: Sandra Aidar de Queiroz
Coorientador: Haroldo Henrique de Rezende Neves
Banca examinadora: Henrique Nunes de Oliveira, Danísio Prado Munari, Lenira El Faro, Patrícia Tholon
Bibliografia

1. LASSO Bayesiano. 2. Modelos de regressão aleatória. 3. Modelos não lineares. 4. RR-BLUP. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.2

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

DADOS CURRICULARES DO AUTOR

Daiane Cristina Becker Scalez nasceu na cidade de Rondonópolis - MT, em 13 de Agosto de 1988, filha de Marcos Antonio Scalez e Elenice Becker. Em 2006, ingressou no curso de Zootecnia na Universidade Federal do Mato Grosso – UFMT, Câmpus de Rondonópolis. Graduou-se em fevereiro de 2010 e, em março de 2010, ingressou no programa de Pós-Graduação em Ciência Animal (área de concentração: Produção e Melhoramento animal) da Universidade Federal do Mato Grosso, orientada pelo Prof. Dr. Fábio Luiz Buranelo Toral, obteve o título de Mestre em fevereiro de 2012. Em março de 2012, iniciou como aluna de doutorado do programa de Pós-Graduação em Zootecnia (área de concentração: Genética e Melhoramento dos Animais Domésticos) da Universidade Estadual “Júlio de Mesquita Filho”, Câmpus de Jaboticabal (UNESP – FCAV) tendo como orientadora a Prof. Dra Sandra Aidar de Queiroz. Em Abril de 2014, iniciou o estágio sanduíche no exterior, permanecendo doze meses na Universität Kassel na Alemanha, orientada pelo Prof. Dr. Sven König.

“Não há acaso, sina, destino, que possa limitar, impedir ou controlar
a firme resolução de uma alma determinada”

(Ella Wheller Wilcox)

Ao meu avô, Jaci Reinaldo Becker, que sempre me apoiou, acreditou e fez tudo que era possível para que eu realizasse meus sonhos.

DEDICO

À minha mãe, à minha vó e ao meu irmão que são a família que cuida de mim.

OFEREÇO

AGRADECIMENTOS

Agradeço...

A Deus por me iluminar a cada amanhecer.

À minha orientadora, Prof^a Dr^a Sandra Aidar de Queiroz, pela oportunidade de fazer o doutorado sob sua orientação, por me aceitar e me ajudar tanto durante esse período em Jaboticabal.

Ao meu co-orientador, Dr. Haroldo Henrique de Rezende Neves, pelo incentivo e pela imensa contribuição para que o trabalho fosse realizado.

Aos membros da banca, Prof. Henrique, Prof. Danísio, Dr^a Lenira e Dr^a Patrícia, pelas valiosas sugestões.

Ao Diercles, que me ajudou de todas as maneiras possíveis, não só com o suporte emocional, mas com valiosas contribuições para o trabalho. Muito obrigada por me ensinar a cada dia ser uma pessoa melhor e mais confiante, você é mais que um noivo, é um amigo e um companheiro de jornada.

Aos meus amigos da Pós-graduação, Cruz Elena, Soliris, Guilherme Venturini, Rafael Medeiros, Rafael Tonussi, Fabrícia, Danielly, Ana Cristina, Giovana, Henrique, Grazyella, Laiza, William, Tiago, Camila, Daniele, Daniel, Naudin, vocês tornaram o fardo mais leve e a caminhada muito mais divertida.

À minha família, que entendeu minhas escolhas e me deu todo o suporte para viver longe de casa, mesmo em momentos de perdas e muita tristeza.

À família do Diercles que me ajudou mais do que eu poderia imaginar que qualquer pessoa pudesse, me proporcionando um novo lar fora de casa.

Aos amigos de Pitangueiras, que me receberam tão bem todas as vezes que estive lá e me proporcionaram muitos finais de semana divertidos.

Ao Prof. Dr. Sven König, que me recebeu na Alemanha e me proporcionou adquirir mais conhecimento.

A todos da Universidade de Kassel na Alemanha, pela convivência e pelo aprendizado, especialmente à Tong Yin, que me ajudou e continua me ajudando todos os dias.

A Capes pela concessão da bolsa no início do doutorado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão da bolsa de doutorado no país (Processo n^o: 2013/ 01068-8) e pela bolsa no exterior (Processo n^o: 2013/23530-5).

Ao financiamento do BMBF e do FBF para o projeto colaborativo intitulado “KMU-innovativ-10: Kuh-L – cow calibration groups for the implementation of selection strategies based on high-density genotyping in dairy cattle”.

SUMÁRIO

	Página
1. INTRODUÇÃO	1
2. REVISÃO DE LITERATURA	3
2.1. Modelos de regressão aleatória (MRA)	3
2.2. Modelos não lineares – curvas de lactação	4
2.3. Seleção genômica	6
2.3.1. Metodologias de predição genômica e distribuição dos efeitos dos QTL ...	9
2.3.2. Validação da seleção genômica	13
3. OBJETIVOS	14
3.1. Objetivo geral	14
3.2. Objetivos específicos	14
4. MATERIAL E MÉTODOS	14
4.1. Descrição dos dados	14
4.2. Análise quantitativa	15
4.2.1. Edição dos dados – Modelos não lineares	15
4.2.2. Modelos não lineares	15
4.2.3. Edição dos dados para análises genômicas – modelos não lineares	16
4.2.4. Edição dos dados – Modelos de regressão aleatória	19
4.2.5. Modelos de regressão aleatória	20
4.3. Controle de qualidade dos dados genômicos – Modelos não lineares e modelos de regressão aleatória	22
4.4. Métodos de seleção genômica ampla (segundo estágio)	23
4.4.1. Comparação dos métodos	24
5. RESULTADOS E DISCUSSÃO	25
5.1. Ajuste das curvas individuais	25
5.2. Predição genômica – modelos não lineares	28
5.3. Modelo de regressão aleatória	32
5.4. Predição genômica – modelo de regressão aleatória	33
6. CONCLUSÕES	37
7. REFERÊNCIAS	38

SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS LONGITUDINAIS DE BOVINOS DA RAÇA HOLANDESA

RESUMO - Características longitudinais são de grande importância para a produção animal, no entanto, quando se trata de seleção genômica vários métodos estão sendo avaliados para características pontuais, como produção de leite em idades específicas ou produção total. Dessa forma, a aplicação destes métodos a trajetórias completas do fenótipo de interesse ao longo do tempo pode ser uma ferramenta importante para a tomada de decisão do melhor momento para seleção dos animais. No presente estudo foram utilizados dados de produção de leite, porcentagem de gordura, porcentagem de proteína e escore de células somáticas de bovinos leiteiros da raça Holandesa em uma metodologia de dois estágios: i) os modelos não lineares de Wood (1967), Cobby & Le Du (1978) e Wilmink (1987) foram ajustados aos dados de cada característica; ii) as estimativas de parâmetros obtidas para as curvas individuais no modelo que proporcionou melhor ajuste aos dados (Wilmink) foram utilizadas como fenótipos nas análises genômicas. Além dos modelos não lineares, também foi utilizado um modelo de regressão aleatória (RRM), considerando polinômios de Legendre quadráticos para a modelagem dos efeitos genético aditivo e de ambiente permanente. No segundo estágio foram testados diferentes métodos para estimação dos valores genéticos genômicos (GEBV): RR-BLUP e LASSO, que consideraram como fenótipos as estimativas dos parâmetros obtidas anteriormente e os coeficientes de regressão estimados via RRM. A correspondência de fenótipos e GEBV foi verificada por meio de validação cruzada, utilizando o método k-fold. Para todas as características estudadas e todos os parâmetros estimados pelo modelo de Wilmink (1987), as acurácias da predição genômica foram baixas quando o RR-BLUP foi utilizado. Além disso, o parâmetro c apresentou acurácias próximas de zero para quase todas as características. As acurácias de predição obtidas para os coeficientes de regressão aleatória foram de magnitude moderada a alta ($>0,67$) demonstrando que o uso de coeficientes de regressão aleatória pode ser uma boa alternativa para predição genômica considerando dados longitudinais.

Palavras chave: LASSO Bayesiano, modelos de regressão aleatória, modelos não lineares, RR-BLUP

GENOMIC SELECTION FOR LONGITUDINAL TRAITS OF HOLSTEIN CATTLE

ABSTRACT - Longitudinal traits are of great importance for animal production, however, when it comes to genomic selection several methods are being evaluated for specific traits, such as milk yield in specific ages or total yield. Thus, applying these methods to complete trajectories of the phenotype of interest over time can be an important tool for helping to decide the best time to select the animals. In this study, milk yield, fat and protein percentage, and somatic cell score data of dairy cattle were used, in a two-stage procedure: i) where the nonlinear Wood (1967), Cobby & Le Du (1978) and Wilmink (1987) were adjusted; ii) the model that provided the best fit was the Wilmink model (1987), whose parameter estimates were used as phenotypes later. In addition to the non-linear models, it was also used a random regression model, considering quadratic Legendre polynomials to model the additive genetic and permanent environment effects. In the second stage different methods for estimation of genomic breeding values (GEBV) were applied: RR-BLUP and LASSO, which considered as phenotypes the estimated parameters previously obtained and the random regression coefficients. The correspondence between phenotypes and GEBV was verified by through cross validation using the k-fold method. For all traits, the genomic prediction accuracies of Wilmink's model parameters were low when RR-BLUP was used. Furthermore, the accuracies for parameter c showed values close to zero for most traits. The prediction accuracies for random regression coefficients were moderate to high, above 0.67, and demonstrated that the use of random regression coefficients can be a good alternative for genomic prediction using longitudinal data.

Keywords: Bayesian LASSO, random regression models, nonlinear models, RR-BLUP

1. INTRODUÇÃO

A União Europeia (UE) apresentou a maior produção mundial de leite em 2014, com 146,5 bilhões de litros produzidos, seguida pela Índia, Estados Unidos, China e Brasil. O bloco europeu também liderou a lista de maior exportador mundial de leite (USDA, 2015). Apesar de ser o maior produtor e exportador de leite do mundo, a atividade profissionalizada na UE é restrita a alguns países como, por exemplo, Alemanha, França, Dinamarca, Reino Unido e Espanha. Outros países, principalmente os do leste Europeu, têm pouca representatividade no total. Dentre os países citados, a Alemanha destaca-se, correspondendo a aproximadamente 20% da produção de leite da União Europeia, sendo o país de maior produção deste bloco (EUROSTAT, 2015).

Quando se trata de produtividade, a Alemanha ficou em terceiro lugar na UE, sendo os primeiros, Dinamarca e Holanda (EUROSTAT, 2015). A produção média na Alemanha é de 9.206 kg de leite por lactação, correspondendo à produção diária de 30,18 kg de leite. A média de produção das vacas de primeira lactação é de 8.000 kg, com média diária de 26,22 kg (DHV, 2015).

A raça Holandesa se destaca com o maior número de animais na Alemanha. O rebanho bovino alemão conta com 12,686 milhões cabeças, sendo que 4,267 milhões são vacas, dentre estas 2,123 milhões são da raça Holandesa, demonstrando a importância da raça para a produção de leite no país e na União Europeia (DHV, 2015).

As avaliações genéticas realizadas pela organização oficial de melhoramento da raça Holandesa na Alemanha (DHV, 2015) já estão levando em consideração a informação genômica para algumas características produtivas e para a fertilidade da fêmea. Dessa forma, os valores genéticos genômicos (GEBVs), que permitem a tomada de decisão baseada na combinação de informações fenotípicas e genotípicas, já estão sendo utilizados para a seleção de animais.

Neste contexto, vários métodos estatísticos vêm sendo utilizados para avaliação de características pontuais, como produção de leite em idades específicas ou produção total. Porém, isso ainda não vem sendo praticado para as

características longitudinais (i.e. aquelas tomadas ao longo do tempo, como produção de leite, de gordura e de proteína), situação em que as predições considerariam a trajetória de produção como um todo.

Inicialmente, essas características longitudinais eram avaliadas por meio de modelos não lineares, as chamadas curvas de lactação. Atualmente, a avaliação genética de medidas repetidas vem sendo realizada mais frequentemente por meio de *test-day models*, como modelos de regressão aleatória (RRM). Apesar da maior complexidade computacional, a adoção de métodos estatísticos que consideram a natureza longitudinal dos dados foi motivada por diferentes estudos que indicaram que a mesma proporcionaria maior precisão do que a avaliação da produção de leite acumulada aos 305 dias, além de permitir contemplar objetivos de seleção como persistência de produção.

Quando se trata de dados longitudinais, metodologias de avaliação genômica para curvas de crescimento têm sido propostas em alguns estudos (IBÁÑEZ-ESCRICHE; BLASCO, 2011; SILVA et al., 2013). Ibáñez-Escriche e Blasco (2011), considerando os três parâmetros da curva de crescimento de Gompertz e, utilizando simulação de dados, indicaram que a seleção genômica pode ser usada para alterar a forma da curva de crescimento de animais. No entanto, a obtenção de resposta em termos de alteração da curva completa (peso ao nascer, peso adulto e a forma da curva) não é fácil, devido às fortes correlações existentes entre todos os pesos corporais ao longo do crescimento do animal.

Em gado leiteiro, Koivula et al. (2012) utilizaram a metodologia de um único passo (*single-step*) aplicada a dados longitudinais de produção de leite, proteína, gordura e contagem de células somáticas e reportaram maior habilidade de predição deste método quando comparado a avaliações tradicionais em que a informação genômica não foi considerada.

Santos et al. (2014) propuseram o uso de coeficientes de regressão aleatória como fenótipos para a avaliação genômica de dados de produção de leite e consideraram essa abordagem viável, pois a habilidade de predição destes coeficientes foi satisfatória, sugerindo sua utilização para prever valores genéticos genômicos de animais jovens.

Desta forma, a aplicação de métodos de seleção genômica a trajetórias completas do fenótipo de interesse ao longo do tempo pode ser uma ferramenta importante para a tomada de decisão do melhor momento para seleção dos animais e deve ser estudada antes da sua aplicação.

2. REVISÃO DE LITERATURA

2.1. Modelos de regressão aleatória (MRA)

Algumas características economicamente importantes, tais como produção de leite, peso corporal, produção de ovos, são mensuradas em vários estágios da vida do indivíduo. Os registros deste tipo de característica são chamados de dados longitudinais.

Estas características têm sido denominadas características de dimensão infinita, conceito definido por Kirkpatrick e Heckman (1989) para denominar características que podem ser descritas por funções contínuas em relação ao tempo (idade, dias em lactação, etc.). Desta forma, a trajetória de crescimento de um indivíduo pode ser descrita como a soma de duas funções contínuas, em que a primeira representa o componente genético aditivo, e a segunda, os efeitos ambientais. Os coeficientes da função que representa o efeito genético aditivo podem ser estimados no contexto da metodologia de modelos mistos por meio de modelos de regressão aleatória (MEYER; HILL, 1997).

Segundo Henderson (1984), diferentes coeficientes de regressão podem ser associados aos diferentes indivíduos da população, de forma que se os indivíduos são uma amostra aleatória da população, os coeficientes de regressão associados a estes indivíduos também devem ser tratados como efeitos aleatórios. A teoria associada à regressão aleatória foi apresentada inicialmente por Henderson Junior (1982).

Utilizando-se modelos de regressão aleatória, os efeitos genéticos aditivos associados a um indivíduo são obtidos como desvios de uma curva para cada animal em relação a uma curva comum a todos os animais. Nesses modelos, os coeficientes de regressão aleatória que representam, por exemplo, a trajetória de

crescimento de um animal, são utilizados na obtenção das funções de covariância, por meio de variâncias e covariâncias entre os coeficientes aleatórios de cada animal (MEYER, 1998).

As funções de covariâncias servem para descrever a estrutura de covariâncias do caráter de forma contínua para a amplitude de idades abrangida pelos dados (KIRKPATRICK et al., 1990). Segundo Meyer (2004), os modelos de regressão aleatória permitem obter as estimativas dos parâmetros genéticos e predições dos valores genéticos para as observações em qualquer ponto do intervalo analisado.

Os MRA apresentam várias vantagens em relação aos modelos tradicionais, como incremento da acurácia na avaliação dos indivíduos (MEYER, 2004); e a inclusão de lactações incompletas, sem a necessidade de projeção (SWALVE, 2000).

2.2. Modelos não lineares (curvas de lactação)

A curva de lactação é a representação gráfica da produção diária, semanal ou mensal e pode ser usada para estimar a produção de leite em qualquer período da lactação (DURÃES et al., 1991). A estimação da produção total de uma vaca enquanto uma lactação ainda está em progresso permite ao produtor identificar antecipadamente o potencial de vacas superiores, facilitando as decisões de descarte e de manejo, além de permitir considerar registros de um maior número de filhas na avaliação de reprodutores (CONGLETON; EVERET, 1980).

Várias funções matemáticas foram utilizadas para determinar a produção de leite de uma vaca e a forma de sua lactação. A primeira função matemática usada para descrever a curva de lactação de uma vaca foi proposta por Brody et al. (1923) e utilizada por Gaines (1927), cuja equação é:

$$y_t = a_0 \cdot e^{-k \cdot t},$$

em que, y_t é a produção de leite num momento t ; e é a base do logaritmo natural; a_0 e k são constantes do modelo. Esse modelo, que é uma função exponencial simples, é mais comumente conhecido como a "Função de Gaines" e segundo Wood (1967) não fornece um bom ajuste dos dados.

Wood (1967) propôs o uso de uma função gama incompleta (FGI), produto de uma potência e um exponencial:

$$y_t = a \cdot t^b \cdot e^{-c \cdot t},$$

em que, a , b e c são constantes relacionadas ao valor inicial de produção, aumento da produção antes do pico de lactação e decréscimo depois do pico de lactação, respectivamente. O pico de produção é dado por $PP = a \left(\frac{b}{c}\right)^b e^{-b}$, e o tempo de pico é dado por: $TP = \frac{b}{c}$. Desde que foi proposto, o modelo de Wood (1967) tem sido amplamente utilizado, principalmente porque com somente três parâmetros se ajusta apropriadamente a um grande número de dados.

Cobby e Le Du (1978) reportaram a dificuldade de interpretação biológica dos parâmetros do modelo de Wood (1967) e a tendência deste modelo subestimar a produção de leite no meio da lactação e superestimar na parte final. Assim, eles propuseram uma nova equação que combina função linear e exponencial:

$$y_t = a - b \cdot t - a \cdot e^{-c \cdot t},$$

onde a , b e c são constantes relacionadas ao valor inicial de produção, ao aumento de produção antes do pico de lactação e ao decréscimo após o pico, respectivamente. A produção no pico (PP) e o tempo do pico de lactação (TP) são dados como: $PP = c^{-1} \left(\ln\left(\frac{a \cdot c}{b}\right)\right)$ e $TP = \ln\left(\frac{a \cdot c}{b}\right)$.

Wilmink (1987), modificando o modelo proposto por Cobby e Le Du (1978), propuseram o uso do seguinte modelo:

$$y_t = a + b \cdot t + c \cdot e^{-0,05 \cdot t},$$

onde a é o parâmetro associado ao nível de produção, b é relacionado ao decréscimo de produção depois do pico, c é relacionado ao aumento de produção até o pico e o fator “-0.05” está relacionado ao tempo de pico de produção, ou seja, aproximadamente 50 dias após o parto. A produção no pico é estimada por $PP = a + b \left(-20 \left(\ln\left(\frac{20b}{c}\right)\right)\right) + c^{e^{-0,05 \cdot (-20 \left(\ln\left(\frac{20b}{c}\right)\right)}}$. O tempo de pico é estimado pela seguinte função: $TP = -20 \left(\ln\left(\frac{20b}{c}\right)\right)$.

2.3. Seleção genômica

Acredita-se que as características economicamente importantes são, em sua maioria, poligênicas, ou seja, determinadas por vários pares de genes com pequeno efeito individual e sofrem grande influência do ambiente (FALCONER; MACKAY, 1996). A seleção para estas características foi realizada tradicionalmente com base nos valores fenotípicos do indivíduo e/ou de seus parentes. Desde a década de 1980, a maior parte das avaliações genéticas vem sendo realizada utilizando o modelo animal, que permite avaliar os animais desde a geração a base até a última geração considerada.

A apresentação da inversa da matriz do numerador dos coeficientes de parentesco entre todos os animais (HENDERSON, 1976) teve papel de destaque nesta fase, além da evolução da capacidade computacional que permitiu trabalhar com um maior número de observações fenotípicas e de pedigree.

Em algumas situações, além da seleção tradicional, baseada em registros fenotípicos e pedigree, informações moleculares passaram a ser utilizadas visando aumentar a eficiência de seleção. Tais informações moleculares compreendem genótipos para marcadores moleculares, bem como genes e QTLs (*Quantitative Trait Loci*) identificados como associados a características de interesse, e podem ser utilizadas em conjunto aos métodos tradicionais de seleção, caracterizando um processo conhecido como seleção assistida por marcadores (MAS) (SILVA et al., 2009). Ganho genético adicional pode ser obtido utilizando-se MAS, em comparação aos métodos tradicionais de seleção (HENSHALL; GODDARD, 1997).

O uso de marcadores genéticos para fins de seleção fundamenta-se na ligação gênica entre os marcadores e os *loci* que determinam as características de interesse (QTL). Contudo, para que a ligação entre marcadores e QTL seja detectada e usada na seleção, é necessário que eles estejam em desequilíbrio de ligação. De modo geral, o desequilíbrio de ligação caracteriza-se pela associação não-aleatória entre diferentes loci. Por exemplo, se dois alelos de *loci* diferentes são encontrados juntos com frequência maior do que aquela esperada baseada na hipótese de segregação independente, infere-se que tais alelos estão em desequilíbrio de ligação (ARDLIE; KRUGLYAK; SEIELSTAD, 2002).

O desequilíbrio de ligação, em nível populacional, é importante para o mapeamento de QTLs na seleção auxiliada por marcadores (MAS) proposta por Lande e Thompson (1990) e, também, para a seleção genômica ampla (GS) proposta por Meuwissen, Hayes e Goddard (2001).

Dekkers (2004) citou três tipos de seleção que utilizam marcadores moleculares, a saber: (i) seleção auxiliada por genes conhecidos (GAS), onde os marcadores são os próprios genes; (ii) seleção auxiliada por marcadores em equilíbrio de ligação com QTLs na população, mas em desequilíbrio de ligação dentro de famílias (LE-MAS); (iii) seleção auxiliada por marcadores em desequilíbrio de ligação com QTLs em nível populacional (LD-MAS). Além destas, a seleção genômica (GS), proposta por Meuwissen, Hayes e Goddard (2001), que fundamenta-se em milhares de marcadores amplamente distribuídos ao longo do genoma e potencialmente em desequilíbrio de ligação populacional com os QTLs que influenciam uma dada característica, também faz uso de informações de marcadores moleculares.

A GS consiste em decisões de seleção considerando o valor genético genômico (GEBV), calculado como a soma dos efeitos de marcadores genéticos (ou haplótipos) distribuídos densamente ao longo do genoma, teoricamente capturando o efeito de todos os loci associados a uma dada característica (QTL) e permitindo a predição acurada do mérito dos animais tão logo fossem genotipados (MEUWISSEN; HAYES; GODDARD, 2001).

O grande impacto da seleção genômica deu-se, principalmente, pelo incremento na acurácia de predição dos valores genéticos genômicos (GEBVs) e possibilidade de redução do intervalo de gerações. Meuwissen, Hayes e Goddard (2001) demonstraram, em um estudo de simulação, que seria possível prever o mérito genético dos animais com alta acurácia (até 0,85), somente com base na informação de genótipos obtida com painéis densos de marcadores moleculares, antes mesmo que os candidatos a seleção tivessem seus próprios registros fenotípicos ou de progênie.

Por essa abordagem, Meuwissen, Hayes e Goddard (2001) compararam os métodos de quadrados mínimos (LS), BLUP e dois métodos de regressão Bayesiana. Exceto no caso de LS, os efeitos dos SNPs (polimorfismos de

nucleotídeo único) foram modelados como efeitos aleatórios. Estes autores constataram que a seleção genômica melhorou a acurácia de predição do mérito genético dos indivíduos, por possibilitar o uso mais eficiente de toda a informação disponível dos marcadores.

Apesar da proposição da GS por Meuwissen, Hayes e Goddard (2001), seu uso permaneceu discreto por cerca de cinco anos, devido ao custo relativamente elevado e restrição dos marcadores disponíveis (RESENDE, 2008). O sequenciamento do genoma bovino em 2009 e o advento da genotipagem por meio de painéis de marcadores SNP tornaram-se base para a aplicação subsequente da GS em melhoramento animal (HAYES et al., 2009).

Avaliando a acurácia das predições genômicas em bovinos leiteiros, considerando quatro países (Estados Unidos, Nova Zelândia, Austrália e Holanda), Hayes et al. (2009) reportaram que a habilidade preditiva dos GEBVs foi substancialmente maior do que a dos valores genéticos provenientes das avaliações tradicionais. Tais resultados permitiram que as companhias de melhoramento genético de bovinos leiteiros aumentassem as taxas de ganho genético e/ou reduzissem custos, principalmente, com testes de progênie.

Pryce e Daetwyler (2012) também relataram as vantagens que podem ser obtidas empregando-se os GEBVs, dentre as quais selecionar animais jovens com alta confiabilidade, reduzir o intervalo de gerações e, conseqüentemente, aumentar o ganho genético. A seleção de animais jovens tem substanciais implicações no delineamento de esquemas de melhoramento de bovinos leiteiros. Por exemplo, para tomar decisões de seleção, não é preciso esperar que um touro tenha filhas com registros fenotípicos, um processo que normalmente leva de cinco a seis anos. Touros jovens sem progênie já poderiam ser selecionados e utilizados como reprodutores em esquemas de seleção genômica.

O potencial descrito por Goddard e Hayes (2007), de grande modificação na estrutura de programas de melhoramento de bovinos, principalmente quanto à seleção de animais para teste de progênie vem se confirmando. Hutchison, Cole e Bickhart (2014) apresentaram resultados referentes ao intenso uso de touros jovens da raça Holandesa em avaliações genômicas nos Estados Unidos que corroboram as previsões relatadas anteriormente. Estes autores reportaram que houve uma

grande redução do intervalo de gerações e aumento da taxa de ganho genético. A idade média dos reprodutores foi reduzida em 2,7 anos para os machos nascidos entre 2006 e 2012 e, para as fêmeas, em 1,3 anos. Além disso, a média das DEPs dos machos nascidos neste período teve aumento de 43%.

Resultados positivos também vêm sendo obtidos no Canadá. Um relatório da rede de pesquisa de bovinos de leite do Canadá (“Canadian Dairy Network – CDN”) evidenciou a redução do intervalo de gerações em animais da raça Holandesa, decorrente da aplicação da GS naquele país. Em 2013, para os touros jovens, o intervalo de gerações foi igual a três anos, representando redução de 55% em relação ao período anterior à aplicação da GS. Para as fêmeas também houve redução de 25%, também atingindo os três anos (VAN DOORMAAL, 2014).

2.3.1. Metodologias de predição genômica e distribuição dos efeitos dos QTL

Métodos estatísticos utilizados em predições genômicas baseiam-se em associações entre um grande número de marcadores SNPs, distribuídos por todo o genoma, e fenótipos, condicionadas ao desequilíbrio de ligação entre os marcadores e os QTLs proximamente ligados. A predição dos efeitos dos marcadores é realizada com base em dados genotípicos e fenotípicos de indivíduos pertencentes a uma amostra da população de seleção. Esses efeitos dos marcadores sobre fenótipos de características quantitativas são somados e usados na predição dos valores genéticos de indivíduos que foram genotipados para os marcadores (RESENDE, 2010), permitindo a predição dos valores genéticos tão logo os animais sejam genotipados.

Entretanto, tal cenário representa um problema do ponto de vista estatístico, pois o número de preditores (marcadores) é geralmente muito superior ao número de observações fenotípicas, não havendo graus de liberdade suficientes para estimar os efeitos de todos os marcadores simultaneamente por regressão (SOLBERG et al., 2008). A regressão por quadrados mínimos poderia ser utilizada para pré-seleção de marcadores com efeitos significativos seguida de regressão múltipla com os marcadores selecionados. No entanto, tal abordagem implicaria em

superestimação dos efeitos dos marcadores e não seria capaz de detectar todos os QTLs (MOSER et al., 2009).

Para contornar tais problemas, três metodologias que tratam os marcadores como efeitos aleatórios foram utilizadas no trabalho de Meuwissen, Hayes e Goddard (2001) e têm sido amplamente empregadas em estudos de seleção genômica, são elas: RR-BLUP, *Bayes A* e *Bayes B*. Há várias diferenças quanto às pressuposições destes métodos e suas principais implicações foram discutidas por Gianola et al. (2009).

O modelo básico considerado por Meuwissen, Hayes e Goddard (2001) pode ser descrito por:

$$y = 1_n\mu + \sum_{j=1}^p X_j g_j + e,$$

em que, y é o vetor de dados (fenótipos) de ordem n , μ é a média geral, 1_n é um vetor de uns de ordem n , g_j e X_j representam o efeito do j -ésimo marcador e uma matriz de incidência relacionando os dados ao efeito do marcador j , respectivamente, e e é um termo residual aleatório, seguindo uma distribuição normal, $N(0, \sigma^2_e)$.

A principal diferença entre os métodos propostos por estes autores reside na definição da variância dos efeitos dos marcadores, como descrito a seguir.

O método referido posteriormente por RR-BLUP permite a obtenção da melhor predição linear não-viesada (BLUP) dos efeitos dos marcadores. Nesta abordagem, assume-se um modelo infinitesimal, sob a pressuposição de que os efeitos de marcadores são considerados aleatórios, com distribuição normal e variância constante (σ^2_a) para todos os segmentos cromossômicos. Uma vantagem deste método é que a única informação requerida *a priori* é uma razão de variâncias utilizada para penalização dos coeficientes de regressão. Todavia a pressuposição de mesma variância para todos os marcadores pode ser uma desvantagem deste método, implicando em predições muito regredidas mesmo para *loci* próximos de QTL, o que pode limitar as acurácias dos GEBVs. Estudos de simulação reportaram menor acurácia deste método comparado a *Bayes B* (MEUWISSEN; HAYES; GODDARD, 2001; HABIER; FERNANDO; DEKKERS, 2007).

O método *Bayes A* pode ser descrito por modelo similar ao apresentado no caso de RR-BLUP, porém considerando variâncias heterogêneas, de modo que assumem-se variâncias específicas (σ^2_j) para cada segmento cromossômico, as quais são estimadas levando em consideração a informação combinada dos dados e da distribuição *a priori* para essas variâncias. Neste método, implementa-se uma regressão bayesiana em que se assume uma distribuição $\chi^2(v, S)$ como priori para σ^2_j , definindo-se hiperparâmetros (v e S) que resultam numa distribuição *t* de Student para os efeitos *a posteriori* dos marcadores.

O método *Bayes B* é uma variação do método *Bayes A*, em que é incluído um passo adicional, implementado via algoritmo de Metropolis-Hastings, em que avalia-se se cada marcador é associado ou não ao QTL e, em caso negativo, atribui-se efeito nulo ao marcador em questão. Na prática, isto implica numa mistura de distribuições para os efeitos dos marcadores que possui maior densidade em zero (DE LOS CAMPOS et al., 2009).

Outro método amplamente estudado é o LASSO (*Least Absolute Shrinkage and Selection Operator*), que associa seleção de variáveis e regularização, via penalização dos coeficientes de regressão. Este método foi proposto por Tibshirani (1996), e modificado por Legarra et al. (2011). Na parametrização proposta por Legarra et al. (2011), assume-se distribuição exponencial dupla para os efeitos dos SNPs, num procedimento equivalente a estimar variâncias individuais para cada marcador.

Para grande parte das características de produção, assumindo-se que apresentam herança poligênica, a distribuição de efeitos de QTL próxima do modelo infinitesimal parece um pressuposto razoável, porém há características em que um único gene é responsável por boa parte da variabilidade. Por exemplo, para porcentagem de gordura no leite de bovinos, uma mutação no gene DGAT1 responde por cerca de 40% da variância fenotípica (GRISART et al., 2004). Neste caso, as pressuposições de métodos como *Bayes A*, *Bayes B* e LASSO (i.e. distribuições específicas por marcador) seriam mais adequadas, uma vez que espera-se que os efeitos de boa parte dos marcadores sejam nulos.

Outro método de regressão Bayesiana, proposto por Habier et al. (2011), trata a probabilidade (π) do SNP ter efeito não-nulo como parâmetro desconhecido, e

essa probabilidade pode ser estimada a partir dos dados. Este método é denominado por BayesCπ. Uma variância comum é especificada para todos os loci, a qual é estimada a partir dos dados, assumindo-se como priori uma distribuição qui-quadrado escalonada invertida.

A habilidade de predição de diversos métodos de estimação dos efeitos dos marcadores foi investigada por diferentes pesquisadores. Hayes et al. (2009), em um estudo comparando oito diferentes metodologias de predição de efeitos de marcadores SNP para seis características de bovinos de leite, verificaram acurácias bastante similares, da ordem de 0,60. Habier et al. (2011) e Hayes et al. (2009) trabalharam com dados reais e também não encontraram diferenças significativas entre os métodos. Lund et al. (2009) e Habier, Fernando e Dekkers (2007) apresentaram resultados com melhor habilidade preditiva para metodologias bayesianas.

Segundo Calus (2010), aplicações em gado de leite com dados reais, assumindo contribuições desiguais dos SNPs para a variância genética, permitiram prever GEBVs com confiabilidade ligeiramente superior aos obtidos a partir de métodos que consideram contribuições iguais dos SNPs.

Esses resultados evidenciam que não há consenso sobre o melhor método a ser utilizado. Hayes et al. (2009) observaram que a metodologia e distribuição mais adequadas para a estimação dos efeitos dos SNPs dependem da característica que está sendo considerada, uma vez que o número de QTLs, assim como a proporção da variância genética explicada pelos SNPs podem variar em função da característica. Dessa forma, torna-se fundamental estudar a metodologia mais apropriada para cada conjunto de dados.

Outro fator que pode influenciar a acurácia de predição é a inclusão de um efeito poligênico adicional no modelo original proposto por Meuwissen, Hayes e Goddard (2001), e esta tem sido apontada como vantajosa. Hayes et al. (2009) reportaram a consideração do efeito poligênico, além dos efeitos dos marcadores, em todos os modelos utilizados em quatro países para predição genômica em bovinos de leite. Tal estratégia permitiria capturar qualquer variação aditiva não associada aos marcadores, além de impor alguma pressão de seleção sobre QTLs de baixa frequência que poderiam não ser capturados por marcadores.

2.3.2. Validação da seleção genômica

Uma estratégia comumente utilizada antes da aplicação prática da seleção genômica, consiste em avaliar o potencial desta tecnologia por meio de estudos de validação. Tal processo pode ser resumido pelas seguintes etapas:

A) Treinamento: modelos estatísticos são ajustados, associando-se os fenótipos à informação genotípica numa população de referência, visando estimar os efeitos de cada marcador sobre uma dada característica. Os fenótipos utilizados podem ser dados de desempenho próprio ou de parentes, valor genético predito usando dados de desempenho e parentesco (EBV) ou EBVs desregredidos.

B) Validação: a confiabilidade dos efeitos estimados é verificada num conjunto independente de dados. Neste caso, têm sido usados procedimentos para verificar a correspondência entre o valor genético predito com base nos efeitos dos marcadores estimados previamente (GEBV) e o EBV tradicional (Hayes et al., 2009).

Uma forma de verificar a correspondência de fenótipos e GEBV é a validação cruzada (Legarra et al., 2008). Erbe et al. (2010) estudaram as mudanças nos resultados usando diferentes números de animais para o grupo de referência e validação e concluíram que uma subdivisão em 5 grupos (5-fold), usando 20% dos dados como conjunto de validação, mostrou-se uma boa estratégia. Nesse método, os dados são divididos em um conjunto de referência e um conjunto de validação. Os parâmetros do modelo são estimados no conjunto de referência e utilizados para prever as observações no conjunto de validação. A correlação de Pearson entre o valor predito e o valor observado indica a adequação ou habilidade preditiva do modelo (LEGARRA et al., 2008).

C) Seleção: uma vez confirmada a confiabilidade das estimativas do GEBV com base em informação de marcadores, é possível a obtenção de predições do mérito genético dos animais candidatos à seleção tão logo forem genotipados e, assim, tomar decisões de seleção com base no GEBV.

3. OBJETIVOS

3.1. Objetivo geral

Objetivou-se avaliar diferentes metodologias de seleção genômica para características longitudinais, como produção de leite, porcentagem de gordura, porcentagem de proteína e escore de células somáticas, além de comparar estas metodologias, indicando eventuais vantagens e desvantagens.

3.2. Objetivos específicos

- Ajustar diferentes modelos longitudinais aos dados de produção de leite, porcentagem de gordura, porcentagem de proteína e escore de células somáticas;
- Comparar a habilidade preditiva dos métodos RR-BLUP e LASSO bayesiano para as características estudadas, sob diferentes modelos longitudinais.

4. MATERIAL E MÉTODOS

4.1. Descrição dos dados

Os dados foram coletados em animais da raça Holandesa provenientes de 507 rebanhos situados na região nordeste da Alemanha, compreendendo 520.779 registros de lactações com medidas de produção de leite (PL), porcentagem de gordura (PG), porcentagem de proteína (PP) e escore de células somáticas (SCS). A contagem de células somáticas (CCS) foi transformada em SCS, conforme descrito por Shook (1982): $SCS = \log_2 (CCS/100) + 3$.

Os animais utilizados na abordagem com modelos não lineares foram genotipados utilizando o painel Illumina BovineSNP50 (Illumina Inc., San Diego, CA), contendo 54.609 marcadores SNP (4.960 animais). Para a metodologia que considerou MRA foram utilizados 5.143 animais genotipados utilizando o painel Illumina BovineSNP50 (Illumina Inc., San Diego, CA) e 2.204 animais genotipados utilizando o painel EuroG10k BeadChip, contendo 11.490 SNP. A imputação dos

genótipos foi feita de acordo com Segelke et al. (2012). Estes dados genotípicos fazem parte de um trabalho colaborativo financiado pelo BMBF e FBF ("*KMU-innovativ-10: Kuh-L – cow calibration groups for the implementation of selection strategies based on high-density genotyping in dairy cattle*").

4.2. Análise quantitativa

4.2.1. Edição dos dados – Modelos não lineares

Para as diferentes características, foram considerados os dados entre o 5º e o 305º dia de lactação; sendo que somente registros de primeira lactação com, pelo menos, quatro controles foram mantidos. Foram desconsideradas do conjunto de dados medidas cujo desvio absoluto em relação à média por controle foi maior que três desvios-padrão. Vacas sem informação de data de nascimento foram excluídas do conjunto de dados. Lactações com o primeiro controle depois dos 60 dias de lactação também foram excluídas, porque poderiam afetar a estimação do pico de lactação, que é o período crítico para a estimação dos parâmetros de qualquer função que represente uma lactação. Para a escolha do modelo mais adequado para cada característica foram mantidas somente vacas genotipadas no arquivo de dados.

Após efetuada a edição dos dados, o arquivo geral continha 35.735 dados mensais para produção de leite (PL), 35.694 dados para porcentagem de gordura (PG), 35.823 dados para porcentagem de proteína (PP) e, para SCS, 35.347 dados.

4.2.2. Modelos não lineares

Três modelos foram usados para descrever as curvas de lactação: Wood (1967), Cobby e Le Du (1978) e Wilmink (1987). O método modificado de Gauss-Newton foi utilizado para estimação dos parâmetros, por meio do PROC NLIN (SAS, 2013).

O modelo não linear mais adequado foi escolhido considerando-se as seguintes estatísticas: R^2 , R^2 ajustado, coeficiente de variação (CV) e desvio-padrão.

O coeficiente de determinação ajustado foi proposto por Draper e Smith (1981) e é representado por:

$$R_A^2 = \frac{(n-1)R^2 - k}{n-k-1},$$

em que, k representa o número de parâmetros do modelo de regressão; n o número de observações, R^2 é o coeficiente de determinação.

4.2.3. Edição dos dados para análises genômicas – modelos não lineares

Após a escolha do modelo não linear mais adequado, foram mantidas todas as vacas do conjunto de dados inicial para estimação dos parâmetros da função de regressão e posterior formação de grupos de contemporâneos.

Os grupos de contemporâneos (GC) foram formados considerando rebanho, ano e estação de parto. Foi utilizado o software AMC (ROSO; SCHENKEL, 2006) para avaliar a conectabilidade entre os grupos. Os GC não conectados ao conjunto de dados principal foram descartados.

Após efetuadas estas análises preliminares, o arquivo geral para PL contou com 1.589.121 dados mensais de 190.873 vacas da raça Holandesa (Figura 1), com média de produção de 28,49kg; o arquivo para PG contou com 1.582.687 informações de 190.925 vacas (Figura 2), com média de 4%; para PP foram considerados 1.589.669 dados de 191.310 vacas (Figura 3), com média de 3,41%; e o arquivo para SCS possuiu 1.570.618 dados de 190.297 vacas (Figura 4), nascidas entre 1996 e 2013, com média de 2,42. Os animais considerados nesta etapa foram nascidos entre 1996 e 2013.

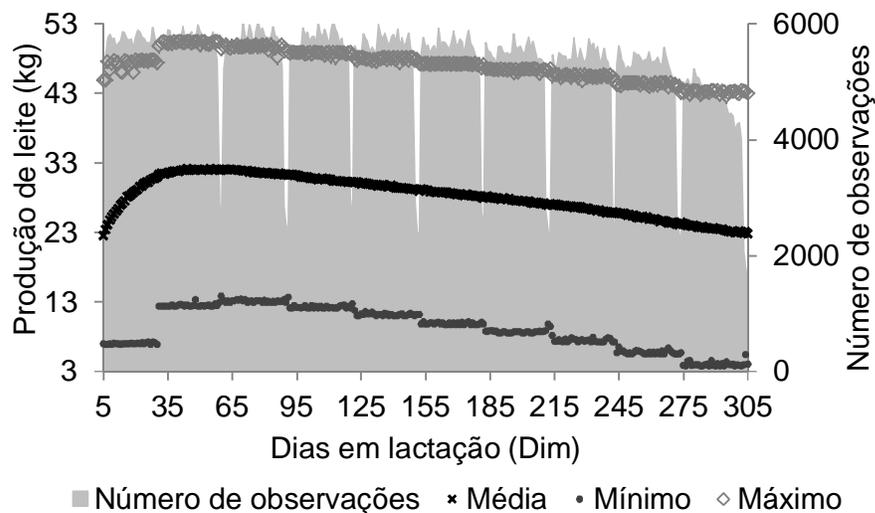


Figura 1. Distribuição do número de observações, média (kg), mínimo (kg), e máximo (kg) de produção de leite de vacas da raça Holandesa de acordo com os dias em lactação

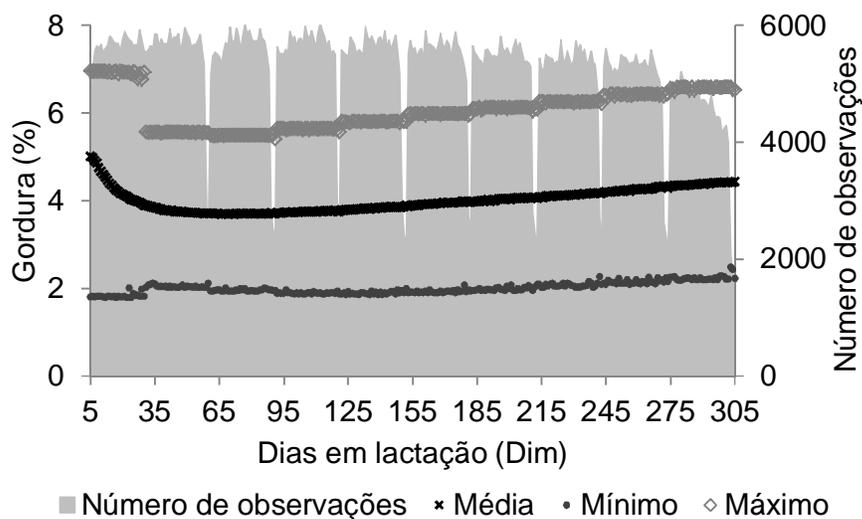


Figura 2. Distribuição do número de observações, média (kg), mínimo (kg), e máximo (kg) de porcentagem de gordura no leite de vacas da raça Holandesa de acordo com os dias em lactação

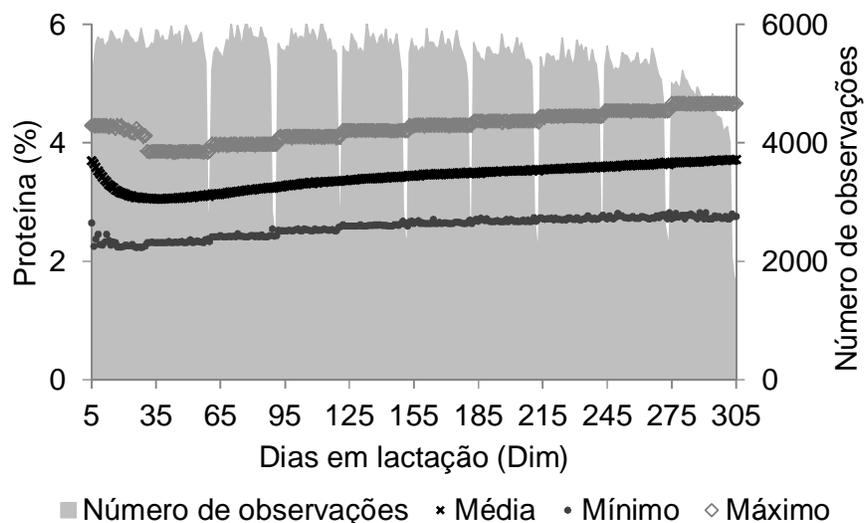


Figura 3. Distribuição do número de observações, média (kg), mínimo (kg), e máximo (kg) de porcentagem de proteína no leite de vacas da raça Holandesa de acordo com os dias em lactação

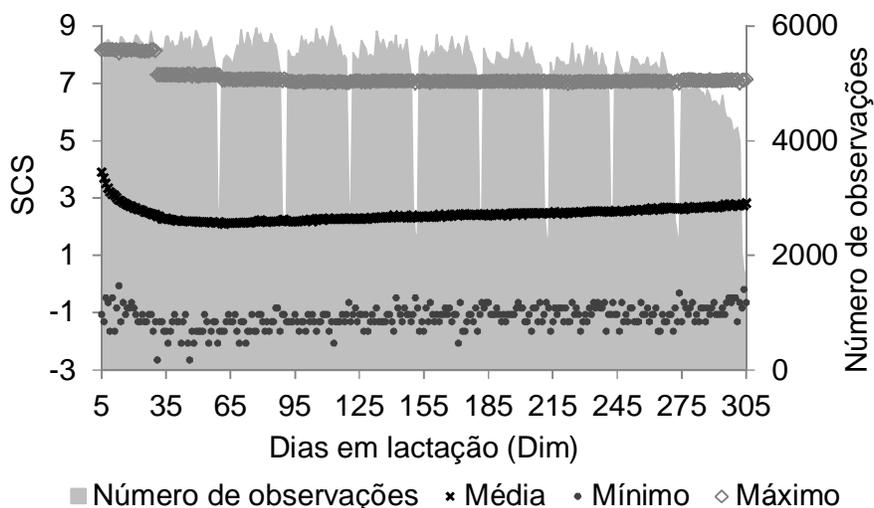


Figura 4. Distribuição do número de observações, média, mínimo, e máximo de escore de células somáticas (SCS) de vacas da raça Holandesa de acordo com os dias em lactação

Após a obtenção de parâmetros por meio do modelo escolhido e da formação de grupos de contemporâneos, somente as vacas cujas curvas de lactação

apresentaram coeficiente de regressão ajustado (R_A^2) maior que 0,50 foram mantidas para as posteriores análises genômicas (Tabela 1).

Um modelo animal uni-característica foi utilizado para estimar componentes de variância e prever os EBVs associados a cada parâmetro da curva de lactação. Para todos os parâmetros estimados da curva individual (a , b e c), o modelo incluiu os efeitos fixos de grupo de contemporâneos, e as covariáveis idade da vaca (efeito linear e quadrático), além dos efeitos aleatórios de animal e resíduo. Os componentes de variância foram estimados pelo método de máxima verossimilhança restrita (REML), utilizando o software AIREMLF90 (Misztal et al., 2002).

Tabela 1. Estatísticas descritivas e estimativas de herdabilidade das estimativas dos parâmetros a , b e c do modelo de Wilmink para as características relacionadas à produção de leite

Característica	N	Média	DP	Mínimo	Máximo	Mediana	h^2
PL – par. a	3283	4,55	2,20	-3,73	12,78	4,5173	0,38
PL – par. b	3283	0,001	0,006	-0,04	0,03	0,001	0,15
PL – par. c	3283	-2,23	2,78	-18,41	24,66	-2,24	0,09
PG – par. a	2823	-0,16	0,21	-1,31	0,77	-0,16	0,39
PG – par. b	2823	-0,0006	0,002	-0,01	0,02	-0,0006	0,63
PG – par. c	2823	0,29	0,44	-2,14	2,44	0,29	0,10
PP – par. a	4685	0,13	0,11	-0,31	0,57	0,13	0,45
PP – par. b	4685	-0,0002	0,0007	-0,004	0,005	-0,0002	0,64
PP – par. c	4685	0,07	0,09	-1,02	0,66	0,08	0,09
SCS – par. a	1301	-0,06	0,18	-0,68	0,57	-0,06	0,06
SCS – par. b	1301	-0,0004	0,01	-0,06	0,07	-0,0004	0,83
SCS – par. c	1301	0,37	0,39	-1,02	1,82	0,32	0,02

PL: produção de leite; PG: porcentagem de gordura; PP: porcentagem de proteína; SCS: escore de células somáticas; N: número de observações; DP: Desvio-padrão; h^2 : herdabilidade.

4.2.4. Edição dos dados – Modelos de regressão aleatória

Foram considerados os dados entre o 5º e o 305º dia de lactação; somente a primeira lactação de cada vaca foi mantida, considerando somente lactações com,

pelo menos, três controles. Foram desconsideradas nas análises as medidas cujo desvio absoluto em relação à média foi maior que três desvios-padrão. Os grupos de contemporâneos foram criados considerando rebanho, ano, mês e dia do controle (herd-test-day). Foram excluídos grupos de contemporâneos com menos de quatro observações.

As estatísticas descritivas dos dados de produção de leite, porcentagem de gordura e porcentagem de proteína são apresentadas na Tabela 2.

Tabela 2. Estatísticas descritivas dos dados utilizados para obtenção dos coeficientes de regressão aleatória para as características relacionadas à produção de leite

Característica	Número de observações	Média	Desvio Padrão	Mínimo	Máximo
Produção de leite (kg)	1.741.932	28,44	6,67	8,00	48,80
RRC, intercepto	336.431	1,98	3,07	-9,15	14,43
RRC, linear	336.431	0,22	1,06	-6,37	6,34
RRC, quadrático	336.431	-0,39	0,53	-3,34	1,87
Gordura (%)	1.735.645	4,00	0,69	1,85	6,18
RRC, intercepto	336.430	-0,07	0,35	-1,63	1,49
RRC, linear	336.430	-0,02	0,13	-0,65	0,55
RRC, quadrático	336.430	0,04	0,07	-0,30	0,37
Proteína (%)	1.742.141	3,41	0,37	2,38	4,44
RRC, intercepto	336.431	-0,02	0,14	-0,68	0,76
RRC, linear	336.431	0	0,6	-0,29	0,30
RRC, quadrático	336.431	0,01	0,03	-0,13	0,16

RRC: coeficiente de regressão aleatória

4.2.5. Modelos de regressão aleatória

O modelo geral utilizado para as análises de MRA das diferentes características apresentou como efeitos fixos a regressão fixa, considerando polinômio de Legendre quadrático para sua modelagem, o grupo de

contemporâneos e a idade vaca como covariável linear e quadrática. Como efeitos aleatórios foram considerados o efeito genético aditivo e o efeito de ambiente permanente, também modelados por polinômios de Legendre quadráticos, considerou-se a variância residual homogênea.

O modelo utilizado pode ser representado, sob a forma matricial, por:

$$\tilde{y} = X\tilde{\beta} + Z_1\tilde{\alpha} + Z_2\tilde{\rho} + \tilde{e},$$

em que: \tilde{y} , representa o vetor de cada uma das características (PL, PG e PP), X , uma matriz de incidência de efeitos fixos; $\tilde{\beta}$, um vetor com as soluções para os efeitos fixos; Z_1 , uma matriz com os coeficientes dos polinômios para modelar os efeitos genéticos aditivos diretos; $\tilde{\alpha}$, um vetor com as soluções para os coeficientes de regressão aleatória dos efeitos genéticos aditivos diretos; Z_2 , uma matriz com os coeficientes dos polinômios para modelar os efeitos de ambiente permanente diretos; $\tilde{\rho}$, um vetor com as soluções para os coeficientes de regressão aleatória dos efeitos de ambiente permanente diretos; e \tilde{e} , um vetor com os resíduos. As seguintes pressuposições podem ser assumidas com relação aos efeitos aleatórios:

$$V \begin{bmatrix} \tilde{\alpha} \\ \tilde{\rho} \\ \tilde{e} \end{bmatrix} = \begin{bmatrix} K_a \otimes A & 0 & 0 \\ 0 & K_p \otimes I & 0 \\ 0 & 0 & R \end{bmatrix},$$

em que, K_a e K_p , representam as matrizes dos coeficientes das funções de covariâncias dos efeitos genéticos aditivos diretos e de ambiente permanente direto, respectivamente; A , a matriz de parentesco; I , uma matriz identidade; e R , uma matriz de variâncias residuais.

No caso dos modelos de regressão aleatória, as soluções para os efeitos fixos e aleatórios podem ser obtidas resolvendo-se o seguinte sistema de equações de Modelos Mistos (MRODE, 2005):

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z_1 & X'R^{-1}Z_2 \\ Z_1'R^{-1}X & Z_1'R^{-1}Z_1 + K_a^{-1} \otimes A^{-1} & Z_1'R^{-1}Z_2 \\ Z_2'R^{-1}X & Z_2'R^{-1}Z_1 & Z_2'R^{-1}Z_2 + K_c^{-1} \otimes I \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \\ \hat{\rho} \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z_1'R^{-1}y \\ Z_2'R^{-1}y \end{pmatrix}$$

Então, o vetor $\hat{\beta}$ contém as estimativas dos efeitos fixos, inclusive os coeficientes que definem a trajetória fixa, $\hat{\alpha}$ contém os coeficientes de regressão aleatória genéticos aditivos diretos preditos de cada animal presente em A , e $\hat{\rho}$ contém os coeficientes de regressão aleatória de ambiente permanente preditos.

No caso da utilização de polinômios para obtenção dos coeficientes de regressão aleatória genéticos aditivos diretos, os coeficientes preditos não são utilizados diretamente para classificação dos animais. A classificação só é possível após a conversão dos coeficientes nos valores genéticos para o dia em lactação de interesse (MRODE, 2005). Para os candidatos à seleção (k), o valor genético aditivo para a característica no dia em lactação j (\hat{VG}_{jk}), pertencente ao intervalo considerado, pode ser obtido por:

$$\hat{VG}_{jk} = t_{\sim j} \hat{\alpha}_{\sim k}$$

em que, $t_{\sim j} = \sum_{l=0}^{L_a-1} \phi_l(id_j^*)$, representa um vetor com os coeficientes do polinômio para o dia em lactação j de interesse.

4.3. Controle de qualidade dos dados genômicos – Modelos não lineares e modelos de regressão aleatória

Para os procedimentos de controle de qualidade foi utilizado o software Plink, versão 1.07 (PURCELL et al., 2007). Foram mantidos para as análises seguintes, apenas os SNPs: com coordenadas genômicas não duplicadas; de posição conhecida em cromossomos autossômicos; que atendessem a *call rate* mínima de 0,98; sem desvio significativo no teste de equilíbrio de Hardy-Weinberg (i.e. p-value

$> 10^{-5}$); e com frequência alélica mínima (MAF) superior a 0,05. Em relação às amostras, foram mantidas apenas aquelas com *call rate* maior que 0,92. A adoção de tais critérios resultou em uma base de dados genotípicos composta por 4.960 animais e 40.884 SNPs para as análises com modelos não lineares e 7.347 animais e 40.597 SNPs para os modelos de regressão aleatória.

4.4. Métodos de seleção genômica ampla (segundo estágio)

Após a escolha do modelo não linear mais adequado para modelagem da PL, PP, PG e SCS, as estimativas dos parâmetros dos modelos escolhidos foram consideradas como fenótipos nos modelos de seleção genômica. Além dos parâmetros sem correção para qualquer efeito fixo, também foram utilizados os parâmetros que foram submetidos a uma análise uni-característica.

Dessa forma, foram testados dois tipos de pseudo-fenótipos nas análises genômicas: as estimativas de cada parâmetro obtidas para as curvas individuais (sem correção para qualquer efeito ambiental) e os EBVs preditos nas análises uni-características.

Os mesmos procedimentos foram realizados para os modelos de regressão aleatória, onde os coeficientes de regressão aleatória foram utilizados como pseudo-fenótipos. Todas as análises consideraram o modelo proposto por Meuwissen, Hayes e Goddard (2001):

$$y = 1_n \mu + \sum_{i=1}^p X_j g_i + e,$$

em que, y é o vetor de dados (fenótipos) de ordem n , μ é a média geral, 1_n é um vetor de uns de ordem n , g_i e X_j representam o efeito do j -ésimo marcador e uma matriz de incidência relacionando os dados ao efeito do marcador j , respectivamente, e e é o vetor de resíduos do modelo, seguindo uma distribuição $N(0, \sigma_e^2)$. Os genótipos foram codificados como 0, 1, 2 e 5 para genótipos AA, AB (BA), BB e genótipos faltantes, respectivamente.

Foram considerados diferentes métodos estatísticos para estimação dos efeitos dos marcadores SNP, implementados no software GS3 (LEGARRA;

RICARD; FILANGI, 2015). Para o método LASSO, a amostragem de Gibbs foi utilizada. Para as análises de validação cruzada foram utilizadas cadeias de 200.000 iterações, com *burn-in* de 40.000 e foram armazenadas a cada 50 amostras.

Para os efeitos dos SNPs foram consideradas diferentes distribuições *a priori*, conforme descrito abaixo:

RR-BLUP: O primeiro método usado para ajustar o modelo de seleção genômica foi o RR-BLUP. Nesta abordagem, assume-se um modelo infinitesimal, sob a pressuposição de que os efeitos de marcadores são considerados aleatórios, com distribuição normal e variância constante (σ_a^2) para todos os segmentos cromossômicos.

LASSO Bayesiano: Esse método foi proposto por Tibshirani (1996) e modificado por Legarra et al. (2011). A distribuição *a priori* para os efeitos dos SNPs (g_i) pode ser representada por:

$$p(g|\lambda) = \prod_{i=1}^p \frac{\lambda}{2} \exp(-\lambda|g_i|).$$

A distribuição *a priori* do efeito do marcador é uma exponencial. λ é o parâmetro "sharpness", ou seja, esse parâmetro determina a forma da distribuição do efeito do SNP. A *priori* de λ é uma distribuição gama, delimitada entre 0 e 10^7 . A parametrização consiste na estimativa de variâncias individuais para cada marcador (τ_i^2), condicional ao parâmetro de regularização λ .

Os valores genômicos (GEBV) foram preditos a partir dos efeitos dos SNPs estimados utilizando a seguinte fórmula:

$$GEBV = \sum_i^n X_i g_i,$$

em que, n é o número de SNPs, X_i é o genótipos do marcadores SNP i , e g_i é o efeito de substituição alélica estimado para o i -ésimo marcador.

4.4.1. Comparação dos métodos

A metodologia de validação cruzada (*cross-validation*) foi utilizada para verificar a adequação das predições obtidas com base nos efeitos dos marcadores. O conjunto de dados foi dividido aleatoriamente em cinco grupos de tamanhos similares (5-fold cross-validation, ERBE et al., 2010), com quatro grupos

representando o grupo de treinamento (80% da população total) e um o grupo de validação (20% da população total). Esse processo foi repetido cinco vezes, de forma que cada grupo fosse uma vez o grupo de validação. Como o processo de divisão tem um componente aleatório que pode afetar os resultados, o processo foi repetido 5 vezes, ou seja, 25 grupos foram criados para validação e 25 grupos para referência para cada parâmetro ou coeficiente, resultando em 75 análises para cada característica e método (5 repetições x 5 grupos x 3 parâmetros/coeficiente).

Os critérios utilizados para comparar os métodos, considerando os animais no grupo de validação, foram:

- Correlação de Pearson entre o GEBV e as respectivas estimativas dos parâmetros (a , b e c) das curvas de Wilmink e, entre os respectivos GEBVs e os coeficientes de regressão aleatória (intercepto, linear e quadrático);
- Correlação de Pearson entre o GEBV e o EBV ($r(\text{GEBV}, \text{EBV})$) para os parâmetros do modelo de Wilmink e para os coeficientes do RRM;
- A escala das predições genômicas (GEBV) foi avaliada com base no coeficiente angular de regressão dos parâmetros sobre o GEBVs preditos ($b(\text{Par}, \text{GEBV})$) e do EBV sobre os GEBVs ($b(\text{EBV}, \text{GEBV})$). Este critério mede o grau de inflação ou deflação da predição genômica, se o valor da regressão for maior que 1 o modelo está subestimando os GEBVs, e se esse valor for abaixo de 1 indica superestimação dos GEBVs.
- Foram calculadas a média e o desvio-padrão considerando as cinco repetições de cada combinação (característica/método e parâmetro) para as estatísticas: $r(\text{GEBV}, \text{parâmetro/coeficiente MRA})$, $r(\text{GEBV}, \text{EBV})$, $b(\text{GEBV}, \text{parâmetro/coeficiente MRA})$ e $b(\text{EBV}, \text{GEBV})$.

5. RESULTADOS E DISCUSSÃO

5.1. Ajuste das curvas individuais

Todas as lactações convergiram quando os modelos de Wood e Wilmink foram utilizados para a modelagem de todas as características estudadas. O modelo

de Cobby e Le Du também apresentou alto índice de convergência, 4.672 lactações para PL, 3.631 para PG, 4.719 lactações para PP e 3.625 para SCS.

Os coeficientes de determinação ajustados ao número de parâmetros (R^2_A) foram calculados para as lactações individuais, mas os valores foram baixos para as quatro características. De acordo com Bianchini Sobrinho (1984), são considerados modelos de bom ajuste aqueles que apresentam R^2_A igual ou maior que 0,80. Segundo esta perspectiva, os resultados demonstram que todos os modelos não ajustaram-se adequadamente aos dados (veja o apêndice para detalhes, Tabelas 1a, 4a, 7a e 10a). Os coeficientes de variação (CV) das estimativas dos parâmetros das curvas foram muito altos para todos os modelos estudados (veja o apêndice para detalhes, Tabelas 2a, 5a, 8a e 11a). Entretanto, menores CV e desvios-padrão (DP) foram observados para os modelos de Wood e Wilmink, comparados ao modelo de Cobby e Le Du para a produção de leite. Para PG, PP e SCS, os valores de CV observados foram altos, mas o modelo de Wilmink apresentou os menores valores.

Para PL, as estimativas dos parâmetros para as lactações individuais, quando ajustadas pelos modelos de Wood e Wilmink foram relativamente próximas do esperado, em que o parâmetro a indica o valor inicial de produção e o b indica o aumento de produção antes do pico para o modelo de Wood. Entretanto, apresentaram valores de CV muito altos. O parâmetro c da função de Cobby e Le Du apresentou um valor extremo de CV, provavelmente, porque a média é muito próxima de zero.

As estimativas dos parâmetros a e b foram relativamente próximas do esperado para PG somente para o modelo de Wilmink. Os parâmetros estimados por meio da função de Cobby e Le Du apresentaram valores de média inconsistentes.

Quando a característica PP foi considerada, as estimativas dos parâmetros para as lactações individuais proporcionadas pelo ajuste de quaisquer dos modelos testados apresentaram valores consistentes. Entretanto, todos os modelos apresentaram valores de CV muito altos para o parâmetro c .

Para SCS, as estimativas dos parâmetros para as lactações individuais, quando ajustadas pelos modelos de Wood e Cobby e Le Du ficaram totalmente fora do esperado, com valores muito altos para os parâmetros a e c .

As funções dos parâmetros, tempo até o pico (TP) e produção no pico, apresentaram os menores valores de CV para o modelo de Wilmink para PL e PG (veja o apêndice para detalhes, Tabelas 3a e 6a). As médias estimadas para pico de produção foram próximas do esperado para os modelos de Wood e Wilmink, com média de produção no pico por volta dos 33 kg para PL e de 3,5% para PG. O TP e o pico de produção apresentaram os menores valores de CV para o modelo de Wilmink para PP e para SCS (veja o apêndice para detalhes, Tabelas 9a e 12a). As médias estimadas para pico de produção ficaram próximas de 3,1% para PP. As médias estimadas para pico de produção e TP pelo modelo Cobby e Le Du apresentaram os maiores valores de CV e médias inconsistentes com os valores esperados.

Com base nos resultados apresentados e diante dos menores valores de R^2_A e menor variação das estimativas dos parâmetros, pode-se considerar que o modelo proposto por Wilmink foi o mais adequado dentre os modelos testados. Graficamente, as médias estimadas para as características, quando este modelo foi utilizado, foram próximas das médias observadas para todas as características (Figura 5).

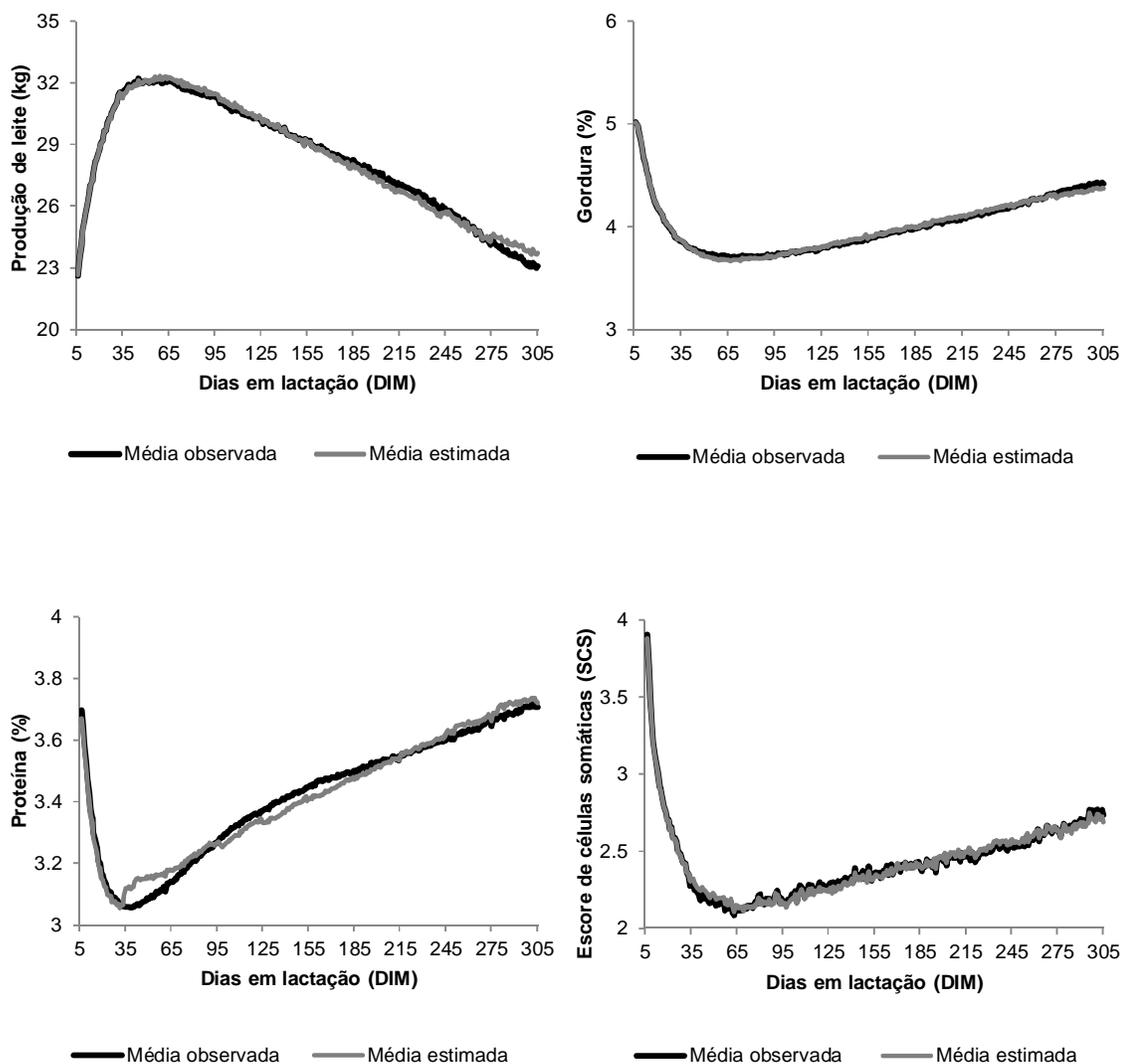


Figura 5. Curvas de lactação médias observadas e estimadas pelo modelo de Wilmink (1987), bem como para porcentagem de gordura e proteína e para escore de células somáticas, para vacas da raça Holandesa

5.2. Predição genômica – modelos não lineares

Para as quatro características estudadas e todos os parâmetros estimados por meio do modelo de Wilmink sem ajuste para efeitos ambientais, os valores para habilidade de predição foram baixos quando o método RR-BLUP foi utilizado (Tabela 3). Além disso, o parâmetro c apresentou habilidade de predição negativa para quase todas as características, dado o desvio-padrão, pode-se considerar que a

média foi próxima de zero. Nesta situação, o sinal tem pouco significado, uma vez que a média não seria significativamente diferente de zero. Isso pode ser explicado pela distribuição deste parâmetro que apresentou curtose positiva (caudas mais alongadas) e pelo fato da correlação de Pearson ser sensível a *outliers*. Mesmo tendo sido feita uma consistência antes da obtenção dos parâmetros, não foi feito um controle após a estimação dos parâmetros, permitindo que todos os valores estimados para o parâmetro *c* fossem mantidos na análise.

Quando o método LASSO foi utilizado, a habilidade de predição apresentou valores superiores aos obtidos pelo RR-BLUP e, somente o parâmetro *c* referente a produção de gordura foi negativo, indicando que este método pode proporcionar uma predição mais acurada. No entanto, os valores de desvio-padrão foram mais altos do que os obtidos por RR-BLUP.

Tabela 3. Habilidade de predição (acurácia média das 5 repetições – 5-fold) medida pela correlação de Pearson entre o pseudo-fenótipo (coeficientes estimados para os parâmetros da curva de Wilmink) e o valor genético genômico $r(y_i, GEBV)$ do pseudo fenótipo para as características produção de leite (PL), porcentagem de gordura (PG), porcentagem de proteína (PP) e escore de células somáticas (SCS) obtidas pelos métodos RR-BLUP e LASSO

Característica	RR-BLUP		LASSO	
	$r(y_i, GEBV)$	Desvio padrão	$r(y_i, GEBV)$	Desvio padrão
PL – parâmetro a	0,2475	0,0237	0,4723	0,3085
PL – parâmetro b	0,1608	0,0271	0,1831	0,1300
PL – parâmetro c	-0,0136	0,0477	0,0249	0,1582
PG – parâmetro a	0,2138	0,0262	0,2307	0,0258
PG – parâmetro b	0,1757	0,0243	0,1856	0,0226
PG – parâmetro c	-0,1336	0,1535	-0,1631	0,1795
PP – parâmetro a	0,3195	0,0223	0,3858	0,1891
PP – parâmetro b	0,1338	0,0329	0,2578	0,1061
PP – parâmetro c	0,0599	0,1284	0,0550	0,1239
SCS – parâmetro a	0,1360	0,0315	0,1603	0,1387
SCS – parâmetro b	0,0473	0,0301	0,1282	0,1898
SCS – parâmetro c	-0,0312	0,1356	0,0312	0,2723

Silva et al. (2013) estudaram a aplicação de seleção genômica a curvas de crescimento e obtiveram bons resultados, com acurácias acima de 0,60 para peso corporal e taxa de crescimento. No entanto, eles utilizaram o fenótipo pré-corrigido para os efeitos de sexo e grupo. No que se refere à comparação de métodos para a seleção genômica, os resultados do presente trabalho concordam com os de Silva et al. (2013), onde o LASSO foi superior ao RR-BLUP sugerindo a segregação de genes de grande efeito para as características estudadas como a razão para a superioridade do método LASSO.

De acordo com Hayes et al. (2009), a proporção da variância explicada pelos marcadores pode variar em função da característica. No entanto, uma das principais razões para a diferença entre os modelos estudados, está na distribuição *a priori* do efeito dos marcadores, sendo plausível supor que aquelas características em que maiores acurácias foram observadas utilizando LASSO, poderiam ser afetadas por genes de efeito moderado a grande (NEVES et al., 2014).

Utilizando o método RR-BLUP, as estimativas dos coeficientes de regressão ($b(y_i, \text{GEBV})$) foram próximas de 1 somente para os parâmetros a de todas as características estudadas e para o parâmetro b de produção de leite (Tabela 4).

Quando existe comparação entre animais genotipados e não genotipados, a inflação e a deflação são consideradas mais graves, pois, se o preditor está inflacionado, o que aconteceu com a maior parte dos parâmetros estudados, a seleção de animais genotipados pode ser favorecida. Por outro lado, os valores apresentados pelo método LASSO sugerem deflação, na maioria dos casos, indicando que uma determinada variação no EBV equivale a uma menor variação no GEBV.

Tabela 4. Coeficiente de regressão (inflação da predição genômica) do pseudo fenótipo sobre o valor genético genômico $b(y_i, \text{GEBV})$ para as características produção de leite (PL), porcentagem de gordura (PG), porcentagem de proteína (PP) e escore de células somáticas (SCS) obtidas pelos métodos RR-BLUP e LASSO

Característica	RR-BLUP		LASSO	
	$b(y_i, \text{GEBV})$	Desvio padrão	$b(y_i, \text{GEBV})$	Desvio padrão
PL – parâmetro a	0,9677	0,1075	1,4619	0,6726
PL – parâmetro b	0,9148	0,1705	1,0514	0,6096
PL – parâmetro c	-2,165	8,4581	1,4717	10,9114
PG – parâmetro a	0,9824	0,1316	1,0150	0,1637
PG – parâmetro b	0,5387	0,0852	1,0074	0,1873
PG – parâmetro c	-40,4200	63,5809	-14,6083	18,9352
PP – parâmetro a	1,0010	0,0778	1,1210	0,3400
PP – parâmetro b	0,1557	0,0366	1,0885	0,3518
PP – parâmetro c	8,523	40,0980	16,9449	80,5871
SCS – parâmetro a	0,9786	0,2543	1,183497	0,7008
SCS – parâmetro b	0,1393	0,090	1,6042	1,8264
SCS – parâmetro c	-0,0824	0,3939	0,8183	2,9065

Os valores de habilidade de predição apresentaram um incremento razoável ao utilizar os EBVs dos parâmetros estimados pelo modelo de Wilmink como pseudo-fenótipos (Tabela 5). Para o presente estudo, o uso do EBV como pseudo-fenótipo é indicado, pois o preditor do mérito genético dos indivíduos foi mais acurado, quando comparado com o uso do próprio parâmetro proveniente do modelo de Wilmink. Vantagens do EBV como variável-resposta foram confirmadas por Guo et al. (2010) em estudo simulado, em que foi comparado o uso de diferentes pseudo-fenótipos. Os autores observaram que o EBV foi mais acurado que o PYD (desvio de produção da progênie). Entretanto, Garrick et al. (2009) relataram alguns problemas do uso de EBVs como pseudo-fenótipos, entre eles a heterogeneidade das variâncias resultantes de EBVs preditos por diferentes quantidades de informação, como é o caso de bovinos de leite, onde os touros possuem números diferentes de progênie. Os autores propuseram usar o EBV desregredido (dEBV) para estimar os efeitos de marcadores. Como só foram utilizadas fêmeas no presente estudo, optou-se pelo uso do EBV, em detrimento ao dEBV, pois elas apresentaram quantidade de

informação similar, proporcionando a obtenção de EBV predito com acurácia similar (resultados não apresentados).

Tabela 5. Habilidade de predição (acurácia média das 5 repetições – 5-fold) medida pela correlação de Pearson entre o pseudo-fenótipo (EBVs estimados para os parâmetros da curva de Wilmink) e o valor genético genômico $r(y_i, \text{GEBV})$ do pseudo-fenótipo e coeficiente de regressão do pseudo-fenótipo sobre o valor genético genômico $b(y_i, \text{GEBV})$ para as características produção de leite (PL), porcentagem de gordura (PG), porcentagem de proteína (PP) e escore de células somáticas (SCS) obtidas pelo método RR-BLUP

Característica	RR-BLUP			
	$r(y_i, \text{GEBV})$	Desvio padrão	$b(y_i, \text{GEBV})$	Desvio padrão
PL – parâmetro a	0,4260	0,0224	0,5102	0,0346
PL – parâmetro b	0,6014	0,0215	0,6904	0,0387
PL – parâmetro c	0,6203	0,0197	0,7188	0,0371
PG – parâmetro a	0,4685	0,0284	0,5755	0,0318
PG – parâmetro b	0,2698	0,0208	0,3394	0,0370
PG – parâmetro c	0,6693	0,0247	0,7797	0,0400
PP – parâmetro a	0,5385	0,0212	0,6082	0,0294
PP – parâmetro b	0,2920	0,0296	0,3319	0,0361
PP – parâmetro c	0,7233	0,0116	0,7811	0,0201
SCS – parâmetro a	0,7238	0,0255	0,9289	0,0444
SCS – parâmetro b	-0,0139	0,0484	-0,0234	0,0791
SCS – parâmetro c	0,7640	0,0308	0,9670	0,0505

5.3. Modelo de regressão aleatória

As estimativas de herdabilidade para PL, PG e PP obtidas sugerem haver variabilidade genética aditiva suficiente para a obtenção de ganho genético expressivo em qualquer uma das características estudadas (Figura 6).

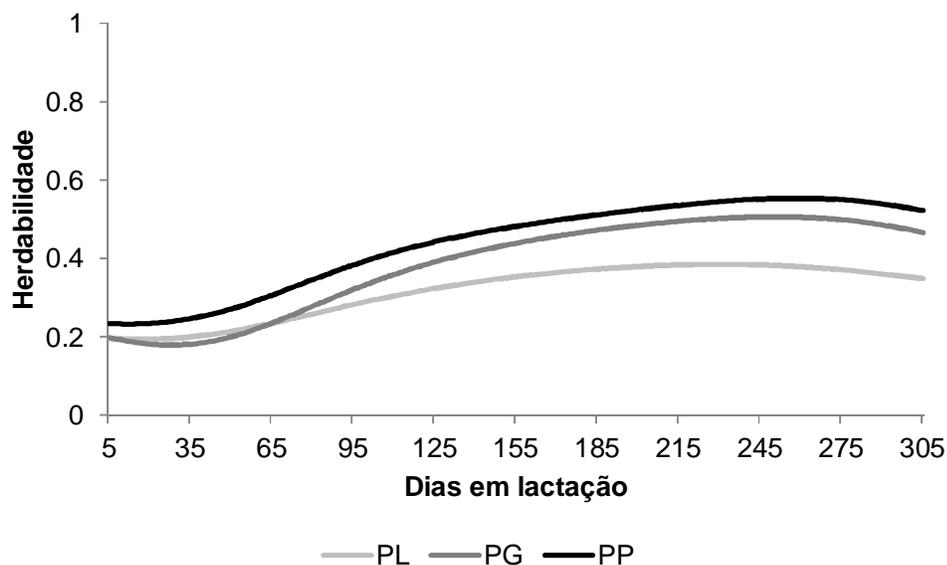


Figura 6. Estimativas de herdabilidade para produção de leite (PL), porcentagem de gordura (PG) e porcentagem de proteína (PP), estimadas por meio de polinômios de Legendre de ordem quadrática para os efeitos genético aditivo e de ambiente permanente

Liu et al. (2000) relataram menores estimativas de herdabilidade para porcentagem de gordura e proteína para vacas da raça Holandesa na Alemanha. Os valores variaram de 0,20 a 0,23 para PG e de 0,16 a 0,25 para PP. As estimativas de herdabilidade para produção de leite foram similares àquelas obtidas neste trabalho.

5.4. Predição genômica – modelo de regressão aleatória

Quando os EBVs referentes a cada um dos coeficientes de regressão aleatória foram utilizados como pseudo-fenótipos, as acurácias encontradas utilizando o procedimento de validação cruzada 5-fold e o método RR-BLUP foram consideradas de magnitude moderada a alta, acima de 0,67, e demonstraram que o uso de coeficientes de regressão aleatória pode ser uma boa alternativa para avaliação de dados longitudinais (Tabela 6). Os valores que indicam inflação ou deflação ($b(y_i, \text{GEBV})$) ficaram próximos de um (Tabela 7).

Tabela 6. Habilidade de predição (acurácia média das 5 repetições – 5-fold) medida pela correlação de Pearson entre o pseudo-fenótipo (coeficientes de regressão aleatória) e o valor genético genômico $r(y_i, \text{GEBV})$ do pseudo-fenótipo para as características produção de leite (PL), porcentagem de gordura (PG) e porcentagem de proteína (PP) obtidas pelos métodos RR-BLUP e LASSO

Característica	RR-BLUP		LASSO	
	$r(y_i, \text{GEBV})$	Desvio padrão	$r(y_i, \text{GEBV})$	Desvio padrão
PL – RRC, intercepto	0,6748	0,0140	0,6856	0,0149
PL – RRC, linear	0,7060	0,0097	0,7128	0,0124
PL – RRC, quadrático	0,7227	0,0140	0,7277	0,0125
PG – RRC, intercepto	0,7164	0,0148	0,7563	0,0104
PG – RRC, linear	0,7062	0,0144	0,7377	0,0117
PG – RRC, quadrático	0,7091	0,0152	0,7343	0,0122
PP – RRC, intercepto	0,7127	0,0120	0,7320	0,0107
PP – RRC, linear	0,7091	0,0095	0,7213	0,0154
PP – RRC, quadrático	0,7295	0,0111	0,7473	0,0113

RRC: coeficiente de regressão aleatória

Tabela 7. Coeficiente de regressão (inflação da predição genômica) do pseudo-fenótipo sobre o valor genético genômico $b(y_i, \text{GEBV})$ para as características produção de leite (PL), porcentagem de gordura (PG) e porcentagem de proteína (PP) obtidas pelos métodos RR-BLUP e LASSO

Característica	RR-BLUP		LASSO	
	$b(y_i, \text{GEBV})$	Desvio padrão	$b(y_i, \text{GEBV})$	Desvio padrão
PL – RRC, intercepto	1,0473	0,0431	1,0438	0,0237
PL – RRC, linear	1,0748	0,0273	1,0686	0,0247
PL – RRC, quadrático	1,0823	0,0439	1,0760	0,0313
PG – RRC, intercepto	0,9916	0,0299	0,9997	0,0315
PG – RRC, linear	1,0149	0,0254	1,0190	0,0272
PG – RRC, quadrático	1,0365	0,0303	1,0370	0,0287
PP – RRC, intercepto	0,9955	0,0265	0,9976	0,0224
PP – RRC, linear	1,0228	0,0202	1,0248	0,0310
PP – RRC, quadrático	0,9743	0,0227	1,0156	0,0260

RRC: coeficiente de regressão aleatória

O método LASSO apresentou maiores valores de habilidade de predição, com maiores diferenças em magnitude para PG e PP (Tabela 6).

Santos et al. (2014) também avaliaram o uso de coeficientes de regressão aleatória como fenótipo para a avaliação genômica de dados de produção de leite e consideraram essa abordagem viável, pois a habilidade de predição destes coeficientes, mensurada por meio da acurácia foi alta, apresentando valores acima de 0,70. No entanto, estes autores não avaliaram o uso de diferentes métodos.

Normalmente, o método BLUP, que assume uma distribuição normal para os efeitos de marcador, mostra-se ligeiramente inferior em relação aos métodos Bayesianos, indicando a presença de genes de moderado a grande efeito. Contudo, o que podemos sugerir é que para a maioria das características relacionadas à produção de leite, o pressuposto do método RR-BLUP, em que há muitos genes de pequeno efeito e poucos ou nenhum de moderado a grande efeito, pode ser próximo da realidade. No caso de PG, onde as diferenças entre os métodos foram ligeiramente maiores para o método LASSO, há um polimorfismo no gene DGAT1 que tem um grande efeito (GRISART et al., 2004).

Como já mencionado anteriormente, Hayes (2009) relatou que a proporção da variância genética explicada pelos marcadores pode variar em função da característica e uma das principais razões está na distribuição *a priori* do efeito dos marcadores, já que para cada modelo foi atribuída uma distribuição diferente. Portanto, aquelas características em que maiores acurácias foram alcançadas utilizando métodos bayesianos, são prováveis de serem afetadas por genes de efeitos moderados a grande (NEVES et al., 2014).

Segundo Daetwyler et al. (2010), a vantagem das metodologias bayesianas sobre o RR-BLUP varia com o número de QTLs afetando a característica, ou seja, a acurácia do RR-BLUP é independente do número de QTLs que afetam a característica, enquanto que a acurácia de métodos bayesianos é maior que a do RR-BLUP quando o número de QTLs é menor que o número de segmentos cromossômicos independentes.

As correlações entre os EBVs e os GEBVs ao longo da lactação foram altas para as três características analisadas e para os dois métodos, demonstrando que em qualquer ponto da curva as predições genômicas poderiam ser utilizadas (Figura 7).

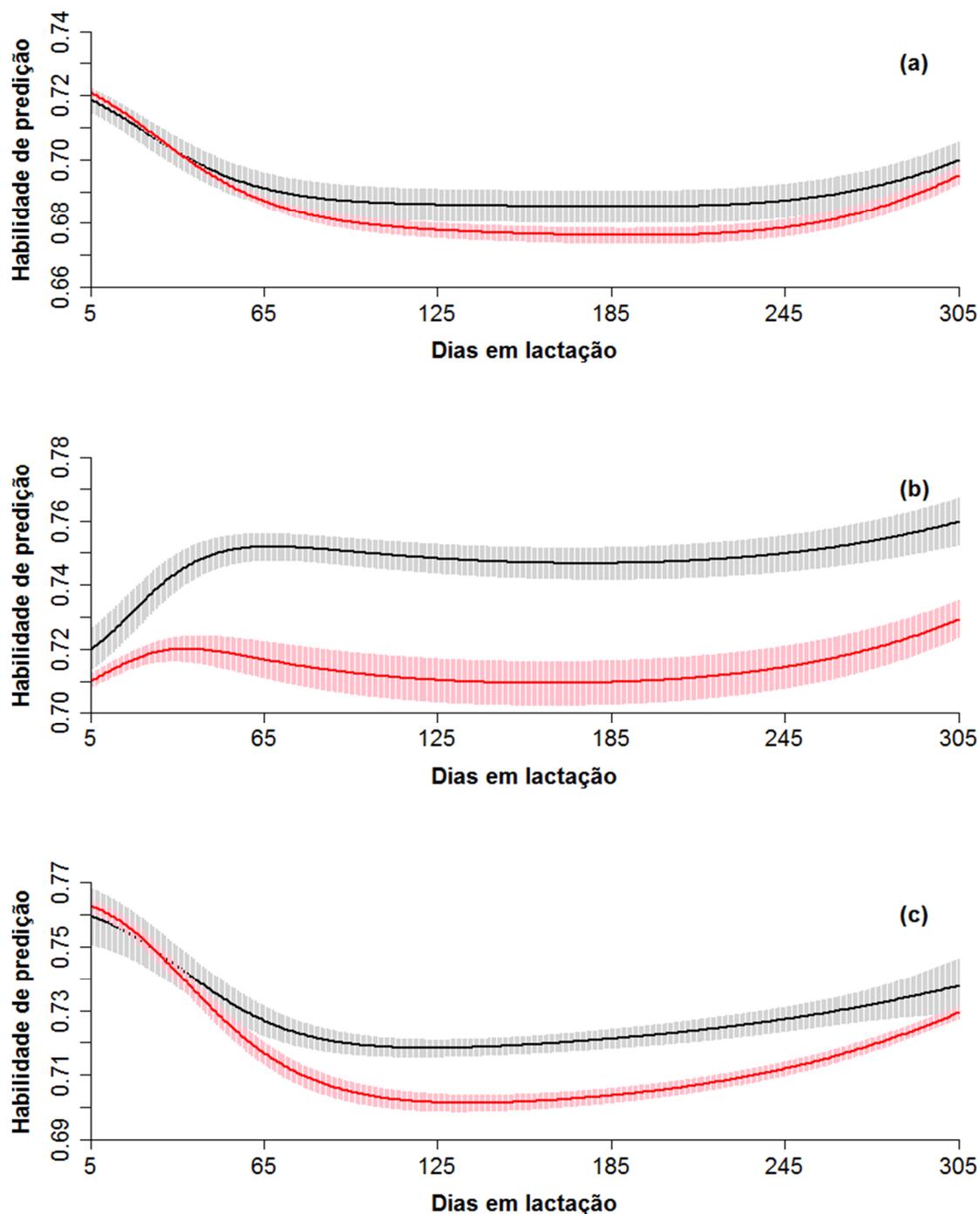


Figura 7. Habilidade de predição medida por meio da correlação de Pearson entre o valor genético (EBV) e o valor genético genômico (GEBV), obtidas pelo método RR-BLUP (vermelho) e pelo método LASSO (preto) e os respectivos intervalos de confiança de 95%, para as características produção de leite (a), porcentagem de gordura (b) e porcentagem de proteína (c).

O método LASSO apresentou maior habilidade preditiva para quase todos os dias em lactação, no entanto, o desvio padrão da habilidade preditiva foi mais elevado em relação ao RR-BLUP. Em função disto, em algumas situações, a diferença entre médias (LASSO vs RR-BLUP) possivelmente não seria significativa.

6. CONCLUSÕES

O uso do modelo de Wilmink proporcionou acurácias relativamente baixas para pelo menos um parâmetro do modelo de Wilmink de todas as características. Vale ressaltar que o banco de dados utilizado para esta análise dispunha de poucas lactações completas, principalmente para as vacas genotipadas, ou seja, seriam necessários outros trabalhos com um banco de dados melhor estruturado para avaliar a adequação desta metodologia para predições genômicas.

Por outro lado, as acurácias de predição obtidas para os coeficientes de regressão aleatória foram de magnitude moderada a alta ($>0,67$), e com valores mais altos quando o método LASSO foi utilizado, demonstrando que o uso de coeficientes de regressão aleatória pode ser uma boa alternativa para predição genômica considerando dados longitudinais.

7. REFERÊNCIAS

- ARDLIE, K. G.; KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the human genome. **Nature reviews. Genetics**, v. 3, n. 4, p. 299–309, 2002.
- BIANCHINI SOBRINHO, E. Estudo da curva de lactação de vacas da raça Gir. Tese de Doutorado, USP, Ribeirão Preto, SP. 1984. 88 p.
- BRODY, S.; TURNER, C.W.; RAGSDALE, A.C. The rate of decline of milk secretion with the advance of the period lactation. **Journal of General Physiology**, v. 5, p. 441, 1923.
- CALUS, M. P. L. Genomic breeding value prediction: methods and procedures. **Animal**, v. 4, n.2, p. 157-164, 2010.
- COBBY, J.M. e LE DU, L.P. On fitting curves to lactation data. **Animal Production**, v. 26, p.127- 133,1978.
- CONGLETON, W.R.J.; EVERETT, R.W. Application of the Incomplete Gamma Function to predict cumulative milk production. **Journal of Dairy Science**, v. 63, p.109-19, 1980.
- DEKKERS, J. C. M. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v. 82 E-Suppl, p. E313–328, 2004.
- DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, p. 375–385, 2009.

DAETWYLER, H. D.; PONG-WONG, R.; VILLANUEVA, B.; WOOLLIAMS, J. A. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, p. 1021-1031, 2010.

Deutscher Holstein Verband e.V. – DHV. Statistic – Facts and figures 2014. Disponível em: <http://www.holstein-dhv.de/facts_and_figures.html>. Acesso em: 29 dez. 2015.

DRAPER, N.M.; SMITH, H. **Applied Regression Analysis**, 2.ed., New York, Wiley, 1981.

DURÃES, M. C.; TEIXEIRA, N. M.; FREITAS, A. F. Curvas de lactação de vacas da raça Holandesa mantidas em confinamento total. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.43, n.5, p.447-458, 1991.

ERBE, M.; PIMENTEL, E.C.G.; SHARIFI, A.R.; H. SIMIANER. Assessment of Cross-validation Strategies for Genomic Prediction in Cattle. In: 9th World Congress on Genetics Applied to Livestock Production, 2010, Leipzig. **Proceedings** of 9th World Congress on Genetics Applied to Livestock Production, 2010.

EUROSTAT Statistics Explained. Milk and milk products statistics. 2015. Disponível em: <http://ec.europa.eu/eurostat/statistics-explained/index.php/Milk_and_milk_product_statistics>. Acesso em: 30 dez. 2015.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to quantitative genetics**. 4.ed. Edinburgh : Longman Group Limited, 1996. 464p.

GAINES, W.L. Persistence of lactation in dairy cows. Agriculture Experiment Station, 1927. (Bulletin, 288).

- GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics, Selection, Evolution**. p. 41-55. 2009.
- GIANOLA, D.; DE LOS CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive genetic variability and the Bayesian alphabet. **Genetics**, v. 183, p. 347-363, 2009.
- GODDARD, M. E.; HAYES, B. J. Genomic Selection. **Journal of Animal Breeding and Genetics**, v. 124, p. 323-330, 2007.
- GRISART, B.; FARNIR, F.; KARIM, L.; CAMBISANO, N.; KIM, J. J.; KVASZ, A.; MNI, M.; SIMON, P.; FRÈRE, J. M.; COPPIETERS, W.; GEORGES, M. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. **Proceedings of the National Academy of Sciences. USA**, v. 101, p. 2308-2403, 2004.
- GUO, G.; LUND, M. S.; ZHANG, Y.; SU, G. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. **Journal of Animal Breeding and Genetics**, v. 127, n.6, p.423–432. 2010.
- HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of genetic relationship on genome-assisted breeding values. **Genetics**, v. 117, p. 2389-2397, 2007.
- HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, v. 12, n. 186. 2011.

- HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J., GODDARD, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, v. 92, p. 433-443, 2009.
- HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, v. 32, p. 69-83, 1976.
- HENDERSON, C. R. **Applications of Linear Models in Animal Breeding**. University of Guelph, 1984.
- HENDERSON Jr., C. R. Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. **Biometrics**, v.38, p.623-640, 1982.
- HENSHALL, J. M.; GODDARD, M. E. Comparison of marker assisted selection using mixed model (BLUP) and mixed model with a test for the QTL. **Proceedings of 12th Conference Association for the Advancement of Animal Breeding and Genetics**, v. 12, p. 217-221, 1997.
- HUTCHISON, J. L.; COLE, J. B.; BICKHART, D. M. Short communication: Use of young bulls in the United States. **Journal of Dairy Science**, v. 97, p. 3213-3220, 2014.
- IBÁÑEZ-ESCRICHE, N.; BLASCO, A. Modifying growth curve parameters by multitrait genomic selection. **Journal of Animal Science**, v. 89, p. 661-668, 2011.
- KIRKPATRICK, M.; HECKMAN, N. A quantitative genetic model for growth, shape and other infinite-dimensional characters. **Journal of Mathematical Biology**, v.27, p.429-450, 1989.

- KIRKPATRICK, M.; LOFSVOLD, D.; BULMER, M. Analysis of the inheritance, selection and evolution of growth trajectories. **Genetics**, v.124, p.979-993, 1990.
- KOIVULA, M.; STRANDÉN, I.; PÖSÖ, J.; AAMAND, G. P.; MÄNTYSAARI, E. A. Single Step Genomic Evaluations for the Nordic Red Dairy Cattle Test Day Data. Interbull Bulletin n°. 46. Cork, Ireland, 2012.
- LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743–756, 1990.
- LEGARRA, A.; RICARD A.; FILANGI O. **GS3 Genomic Selection — Gibbs Sampling — Gauss Seidel (and BayesCπ)**. Disponível em: <http://snp.toulouse.inra.fr/~alegarra/manualgs3_last.pdf>. Acesso em: 20 jul. 2015.
- LEGARRA, A.; ROBERT-GRANIE, C.; MANFREDI, E.; ELSEN, J-M. Performance of genomic selection in mice. **Genetics**, v. 180, p. 611–618, 2008.
- LEGARRA, A.; ROBERT-GRANIE, C.; CROISEAU, P.; GUILLAUME, F.; FRITZ, S. Improved Lasso for genomic selection. **Genetics Research**, v. 93, p. 77–87, 2011.
- LEGARRA, A. **Bases for Genomic Prediction**. Version 0.7.1. 77p. 2014.
- LIU, Z.; REINHARDT, F.; REENTS, R. Estimating Parameters of a Random Regression Test Day Model for First Three Lactation Milk Production Traits Using the Covariance Function Approach. Interbull Bulletin 01/2000; 25.
- LUND, M. S.; SAHANA, G.; KONING, D.; SU, G.; CARLBORG, O. Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. **BMC Proceedings**, v.3(Suppl 1):S1, 2009.

- MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819–1829, 2001.
- MEYER, K.; HILL, W. G. Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. **Livestock Production Science**, v.47, p.185-200, 1997.
- MEYER, K. Estimating covariance functions for longitudinal data using random regression model. **Genetics Selection Evolution**, v.30, p.221-240, 1998.
- MEYER, K. Scope for random regression model in genetic evaluation of beef cattle for growth. **Livestock Production Science**, v.86, p.69-83, 2004.
- MISZTAL, I.; TSURUTA, S.; STRABEL, T.; AUVRAY, B.; DRUET, T.; LEE, D. BLUPF90 and related programs (BGF90). In: 7th World Congress on Genetics Applied to Livestock Production, Montpellier. **Proceedings** of 7th World Congress on Genetics Applied to Livestock Production, Communication No. 28–07. INRA, Castanet-Tolosan, France, 2002.
- MOSER, G.; TIER, B.; CRUMP, R. E.; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution**, v. 41, p. 56, 2009.
- MRODE, R. A. Linear models for the prediction of animal breeding values. 2.ed. Wallingford: CABI Publishing, 2005. 344p.
- NEVES, H. H. R.; CARVALHEIRO, R.; O'BRIEN, A. M. P.; UTSUNOMIYA, Y. T.; CARMO, A. S.; SCHENKEL, F. S.; SÖLKNER, J.; MCEWAN, J.C.; VAN TASSEL, C. P.; COLE, J. B.; SILVA, M. V. G. B.; QUEIROZ, S. A.; SONSTERGARD, T. S.; GARCIA, J. F. Accuracy of genomic predictions in *Bos indicus* (Nellore) cattle. **Genetics Selection Evolution**, v. 46, n. 17, 2014.

PRYCE, J. E.; DAETWYLER, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. **Animal Production Science**, v. 52, p. 107-114, 2012.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007.

RESENDE, M. D. V. 2008 Genômica quantitativa e seleção no melhoramento de plantas perenes e animais. Colombo: Embrapa Florestas, 330 p.

RESENDE, M. D. V.; RESENDE Jr., M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. 2010 Computação da Seleção Genômica Ampla (GWS). Colombo: Embrapa Florestas, 79 p.

ROSO, V.M.; SCHENKEL, F.S. AMC – A computer program to assess the degree of connectedness among contemporary groups. In: 8th World Congress of Genetics Applied to Livestock Production, 2006, Belo Horizonte. **Proceedings** of 8th World Congress of Genetics Applied to Livestock Production, 2006. p.26. (Communication, 27).

SANTOS, D. J.; BOISON, S. A.; UTSUNOMYA, A. T.; PEIXOTO, M. G. C. D.; TONHATI, H.; SÖLKNER, J.; SILVA, M. V. An approach to genomic analysis of longitudinal data using random regression. In: 10th World Congress of Genetics Applied to Livestock Production, 2014, Vancouver, **Proceedings** of 10th World Congress of Genetics Applied to Livestock Production, 2014.

SAS User's Guide: Statistics. SAS Institute Inc., Cary, North Caroline, 2013.

- SEGELKE, D.; CHEN, J.; LIU, Z.; REINHARDT, F.; THALLER, G.; REENTS, R.
Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. **Journal of Dairy Science**, v.95, p. 5403-5411, 2012.
- SHOOK, G. E. Approaches to summarizing somatic cell count which improve interpretability. In: ANNUAL MEETING NATIONAL MASTITIS COUNCIL, 2., 1982, Arlington. **Proceedings...** Arlington: 1982. p.150.
- SILVA, K. M.; GUIMARÃES, S. E. F.; LOPES, P. S., NASCIMENTO, C. S.; LOPES, M. S.; AMARAL, M. M. D. C. Mapeamento de locos de características quantitativas para desempenho no cromossomo 4 de suínos. **Revista Brasileira de Zootecnia**, v. 38, p. 474-479, 2009.
- SILVA, F. F.; RESENDE, M. D. V.; ROCHA, G. S., DUARTE, D. A.; LOPES, P. S.; BRUSTOLINI, O. J.; THUS, S.; VIANA, J. M.; GUIMARÃES, S. E. Genomic growth curves of an outbred pig population. **Genetics and Molecular Biology**, v. 36, n. 4, p. 520-527, 2013.
- SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. Genomic selection using different marker types and densities. **Journal of Dairy Science**, v. 86, 2447–2454, 2008.
- SWALVE, H. H. Theoretical basis and computational methods for different test-day genetic evaluation methods. **Journal of Dairy Science**, v. 83, p. 1115-1124, 2000.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society, Series B**, v. 58, p. 267-288, 1996.
- UNITED STATES DEPARTMENT OF AGRICULTURE - USDA. **Dairy: World Markets and Trade**. Foreign Agricultural Service. Dec. 2015. 21p.

VAN DOORMAAL, B. Impact of Genomics on Genetic Selection and Gain. Canadian Dairy Network, p. 1–3, 2014.

WILMINK, J.B.M. Adjustment of test day milk, fat and protein yields for age, season and stage of lactation. **Livestock Production Science**, v.16, p.335-348, 1987.

WOOD, P.D.P. Algebraic model of the lactation curve in cattle. **Nature**, v.216, p.164-165, 1967.

APÊNDICE

Tabela 1a. Número de lactações e frequência de lactações de vacas da raça Holandesa, ajustados pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), de acordo com o R^2_A , para produção de leite

R^2_A	Wood		Cobby e Le Du		Wilmink	
	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)
$R^2_A \geq 0,50$	2.202	46,70	2077	44,05	2209	46,85
$R^2_A \geq 0,70$	1.354	28,72	1241	26,32	1336	28,34
$R^2_A \geq 0,80$	840	17,82	750	15,91	842	17,86
$R^2_A \geq 0,90$	270	5,73	270	5,73	284	6,02

Tabela 2a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987).

Parâmetro	Média	SD	CV (%)	Máximo	Mínimo
Wood 4.720 vacas					
a	31,9318	6,3371	19,8458	75,0491	10,6732
b	0,2537	0,3196	125,9707	2,7424	-2,8093
c	0,0838	0,0884	105,4960	0,9503	-0,7248
Cobby e Le Du 4.672 vacas					
a	36,1116	24,5444	67,9683	1.624,55	14,8063
b	0,0348	0,0631	181,3153	3,6550	-0,1151
c	1,5261E71	1,0214E73	6.692,99	6,98E74	0,0028
Wilmink 4.720 vacas					
a	35,6471	6,3594	17,8398	66,0422	6,0655
b	-0,0335	0,0325	-96,9222	0,1729	-0,2663
c	-12,8066	22,3124	-161,6066	251,8145	-260,4067

Tabela 3a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987)

Parâmetro	Média	SD	CV (%)	Mínimo	Máximo
Wood					
PP (4.313)	33,1034	5,2287	15,7951	9,4288	54,1491
TP (4.715)	46,8376	2620,59	5.595,05	-173.958,60	29224,89
Cobby e Le Du					
PP (4.156)	43,0503	25,8415	60,0262	2,5729E-73	221,5192
TP (4.156)	8,6909	16,1743	186,1066	0,2067	179,5871
Wilmink					
PP (3.865)	33,4363	5,1116	15,2877	16,4573	52,3665
TP (3.865)	62,2730	21,2221	34,0792	-66,5722	191,9300

PP: pico de produção (kg); TP: tempo para o pico de produção (dias).

Tabela 4a. Número de lactações e frequência de lactações de vacas Holandesas para porcentagem de gordura no leite, ajustados pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), de acordo com o R^2_A

R^2_A	Wood		Cobby e Le Du		Wilmink	
	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)
$R^2_A \geq 0,50$	1.990	42,10	1	0	1.924	40,70
$R^2_A \geq 0,70$	1.180	24,96	0	0	1.103	23,33
$R^2_A \geq 0,80$	728	15,40	0	0	654	13,84
$R^2_A \geq 0,90$	276	5,84	0	0	236	4,99

Tabela 5a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para porcentagem de gordura

Parâmetro	Média	SD	CV (%)	Máximo	Mínimo
Wood 4.727 vacas					
a	8,7348	33,8154	387,1326	1990,98	0,0488
b	-0,1603	0,2118	-132,0935	1,3761	-1,8721
c	-0,0020	0,0022	-110,6056	0,0206	-0,0189
Cobby e Le Du 3.631 vacas					
a	3623583199	2140346699	59,0671328	269467206199	431580,70
	8	6			
b	3623,53	2.140,33	59,0676427	26946,54	0,0010
c	1E-7	0	0	1E-7	1E-7
Wilmink 4.727 vacas					
a	3,2213	0,6482	20,1215	7,2859	-1,4940
b	0,0038	0,0040	104,7239	0,0393	-0,0300
C	2,1078	2,9942	142,0510	29,4456	-19,3320

Tabela 6a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para porcentagem de gordura

Parâmetro	Média	SD	CV (%)	Mínimo	Máximo
Wood					
PP (4.302)	3,5585	0,5023	14,1147	1,8653	7,7203
TP (4.727)	33,6580	5329,68	15834,80	-320074,91	168134,36
Cobby e Le Du					
PP (3.631)	43907,81	1318114,34	3002,00	62,8062	41371450,42
TP (3.631)	0,0044	0,1318	3002,00	6,2806E-6	4,1371
Wilmink					
PP (3.929)	3,4858	0,4610	13,2247	1,4426	5,6202
TP (3.929)	67,0482	22,0045	32,8189	-58,7653	180,2080

PP: pico de produção (kg); TP: tempo para o pico de produção (dias).

Tabela 7a. Número de lactações e frequência de lactações de vacas Holandesas para porcentagem de proteína no leite, ajustados pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), de acordo com o R^2_A

R^2_A	Wood		Cobby e Le Du		Wilmink	
	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)
$R^2_A \geq 0,50$	3149	66,50	2198	46,42	3238	68,38
$R^2_A \geq 0,70$	2237	47,24	1379	29,12	2341	49,44
$R^2_A \geq 0,80$	1555	32,84	870	18,37	1614	34,09
$R^2_A \geq 0,90$	631	13,33	331	6,99	679	14,34

Tabela 8a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para porcentagem de proteína no leite

Parâmetro	Média	SD	CV (%)	Máximo	Mínimo
Wood 4.735 vacas					
a	3,3514	1,1471	34,2268	14,6080	0,1879
b	-0,0148	0,1084	-731,2554	0,7409	-0,4333
c	-0,0009	0,0011	-114,7322	0,0061	-0,0062
Cobby e Le Du 4.719 vacas					
a	3,0595	0,2530	8,2690	4,4829	2,1771
b	-0,0021	0,0015	-69,2989	0,0084	-0,0105
c	1,7802	6,1911E71	3477,75	3,0631E73	0,0201
Wilmink 4.735 vacas					
a	2,9241	0,2857	9,7716	4,1855	1,7433
b	0,0029	0,0016	55,0681	0,0166	-0,0064
c	0,2422	1,2391	511,5974	7,0825	-12,6656

Tabela 9a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para porcentagem de proteína no leite

Parâmetro	Média	SD	CV (%)	Mínimo	Máximo
Wood					
PP (3.634)	3,1880	0,3349	10,5049	2,5401	6,8473
TP (4.735)	-55,6979	6744,95	-12109,88	-423038,85	102473,56
Cobby e Le Du					
PP (293)	28,7530	46,4148	161,4255	5,8145E-72	187,6391
TP (293)	26,9518	36,0620	133,8019	2,9687	178,1038
Wilmink					
PP (3.298)	3,0574	0,2116	6,9208	2,0859	3,8943
TP (3.298)	48,0051	22,2559	46,3616	-134,2824	167,3677

PP: pico de produção (kg); TP: tempo para o pico de produção (dias).

Tabela 10a. Número de lactações e frequência de lactações de vacas Holandesas para escore de células somáticas, ajustados pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), de acordo com o R^2_A

R^2_A	Wood		Cobby e Le Du		Wilmink	
	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)	Número de lactações	Frequência (%)
$R^2_A \geq 0,50$	988	20,99	242	5,14	971	20,63
$R^2_A \geq 0,70$	545	11,58	100	2,12	509	10,82
$R^2_A \geq 0,80$	322	6,84	49	1,04	308	6,54
$R^2_A \geq 0,90$	136	2,89	17	0,003	122	2,59

Tabela 11a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para escore de células somáticas

Parâmetro	Média	SD	CV (%)	Máximo	Mínimo
Wood 4.706 vacas					
a	-1,6286E51	1,0824E53	-6645,76	5,1933E31	-7,4210E51
b	0,0241	2,5722	10688,18	26,4960	-41,0343
c	1502917250	1,0310E13	6860,03	7,0727E14	-0,3686
	73				
Cobby e Le Du 3.625 vacas					
a	8711,77	500210,11	5741,77	30056103,75	-531406,78
b	-0,0115	1,2971	-11259,73	10,7246	-76,5320
c	1,9241	7,6792E72	3991,00	4,1786	-0,2163
Wilmink 4.706 vacas					
a	1,9729	1,8801	95,2974	12,4652	-5,6946
b	0,0019	0,0100	532,8249	0,0769	-0,0780
c	1,8152	7,4479	410,3060	58,3704	-103,2317

Tabela 12a. Média, desvio-padrão (SD), coeficiente de variação (CV), valores mínimos e máximos das estimativas de parâmetros para as lactações individuais, ajustadas pelos modelos de Wood (1967), Cobby e Le Du (1978) e Wilmink (1987), para escore de células somáticas

Parâmetro	Média	SD	CV (%)	Mínimo	Máximo
Wood (
PP (4.151)	2,4387	5,9698	244,7916	-0,9083	374,2324
TP (4.706)	-54,0433	9765,37	-18069,52	-603561,18	189684,41
Cobby e Le Du					
PP (2.647)	35,3678	65,5408	185,3119	-987,8765	617,1762
TP (2.647)	20,8228	31,6270	151,8864	-29,9353	177,2752
Wilmink					
PP (3.448)	2,0379	1,4244	69,8973	-0,9337	8,4712
TP (3.448)	70,2569	27,6261	39,3215	-99,5512	223,5604

PP: pico de produção (kg); TP: tempo para o pico de produção (dias).