MAX BRANDÃO DE OLIVEIRA

BAYES t E BAYES DE: MODELOS BAYESIANOS ROBUSTOS PARA SELEÇÃO GENÔMICA AMPLA

TERESINA-PI

2019

MAX BRANDÃO DE OLIVEIRA

BAYES t E BAYES DE: MODELOS BAYESIANOS ROBUSTOS PARA SELEÇÃO GENÔMICA AMPLA

Tese apresentada ao Programa de Pós-Graduação em Ciência Animal, na área de Produção Animal, como parte dos requisitos para obtenção do título de Doutor.

Área de concentração: Produção Animal.

Orientador: Prof. Dr. José Lindenberg Rocha Sarmento

TERESINA-PI

2019

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Serviço de Processamento Técnico

O48b Oliveira, Max Brandão de.
Bayes t e Bayes DE : modelos bayesianos robustos para seleção genômica ampla / Max Brandão de Oliveira. – 2019.
71 f.

Tese (Doutorado em Ciência Animal) – Universidade Federal do Piauí, Teresina, 2019.
"Orientador: Prof. Dr. José Lindenberg Rocha Sarmento".

1. Características de carcaça. 2. Distribuição t.
3. Ovinos de corte. 4. Robustez. 5. Valor genômico.
I. Título.

CDD 636.31

BAYES t E BAYES DE: MODELOS BAYESIANOS ROBUSTOS PARA SELEÇÃO GENÔMICA AMPLA

MAX BRANDÃO DE OLIVEIRA

Tese aprovada em: 15/03/2019

Banca Examinadora:

Prof. Dr. José Lindenberg Rocha Sarmento (Presidente) / DZO/CCA/UFPI

Prof. Dr. Fábio Barros Britto (Interno) / CCN/UFPI

Prof. Dr. José Ailton Alencar Andrade (Externo) / UFC

Prof. Dr. Luiz Antonio Silva Figueiredo Filho (Externo) / IFMA

Prof. Dr. Luiz Fernando Brito (Externo) / PURDUE UNIVERSITY

"*Nós poderíamos ser muito melhores se não quiséssemos ser tão bons.*"

Sigmund Freud

Aos meus pais, irmãos

e esposa, que me acompanharam nessa jornada.

DEDICO

AGRADECIMENTOS

A *Deus*, que tem me dado fé e proteção a cada momento da minha vida;

A Universidade Federal do Piauí pela oportunidade de realização desse trabalho com etapa de aperfeiçoamento profissional;

Ao meu orientador prof. Dr. José Lindenberg Rocha Sarmento pelo incentivo e pelos valiosos conselhos profissionais;

Aos professores Dr. Ailton Andrade, Dr. Luiz Brito, Dr. Luiz Figueiredo, Dr. Fábio Britto, pela participação na banca examinadora, pela ajuda e sugestões para o enriquecimento do trabalho;

Aos Professores (as) do Programa de Pós-graduação em Ciência Animal da UFPI pela contribuição ao meu aprendizado, pelos ensinamentos e experiências transmitidas;

Agradeço ao professor Antônio de Sousa Júnior, pela parceria e intermédio com ovinocultores dos estados do Piauí e Maranhão, aos quais também sou profundamente grato pela contribuição para a obtenção dos dados fenotípicos utilizados neste trabalho;

Meu muito obrigado às instituições de pesquisa e órgãos financiadores (CNPq, UFPI, INCT-CA e FAPEMA) que possibilitaram a aquisição dos dados moleculares e materiais para a concretização da proposta sugerida neste trabalho;

Ao Grupo de estudos em genética e melhoramento animal (GEMA), pelos momentos de aprendizado e descontração. Em especial aos amigos Luciano Silva, Tatiana Saraiva, Laylson Borges, Bruna Lima e Débora Carvalho, pela amizade e companheirismo nessa caminhada que seguimos juntos;

A minha esposa, Fernanda Abreu Coelho, pela contribuição científica e emocional, pela paciência e pelo companheirismo dedicado integralmente na jornada pelo conhecimento;

Aos professores do pós-graduação em Ciência Animal e em Genética e Melhoramento pelas valiosas orientações e contribuições na minha formação;

Aos meus familiares, pela motivação, educação, paciência e apoio dados em toda minha vida como apoio fundamental na minha vivência como pessoa e profissional;

Enfim, a todas as pessoas que direta ou indiretamente colaboraram para a realização desta tese.

MUITO OBRIGADO!

LISTA DE SIGLAS E ABREVIATURAS

AIC	*Akaike Information Criterium* (Critério da Informação de Akaike)
AOL	Área de Olho de Lombo
ARCO	Associação Brasileira dos Criadores de Ovinos
DE	Dupla-exponencial
DIC	*Deviance Information Criterium* (Critério da Informação Deviance)
DP	Desvio-Padrão
EP	Erro-Padrão
FAO	*Food and Agriculture Organization* (Organização das Nações Unidas para a Alimentação e Agricultura)
FDP	Função Densidade de Probabilidade
GEBV	*Genomic Estimator Breeding Value* (Valor Genético Genômico Estimado)
GWS	*Genomic Wide Selection* (Seleção Genômica Ampla)
IBGE	Instituo Brasileiro de Geografia e Estatística
MCMC	Monte Carlo via Cadeias de Markov
QTL	*Quantitative Traits locus* (Lócus de Característica Quantitativa)
REML	*Maximum Restricted Likelihood* (Máxima Verossimilhança Restrita)
RRBLUP	*Ridge Regression Best Unbiased Linear Estimator* (Regressão de Cumeeira do tipo BLUP)
SAM	Seleção Assistida por Marcadores
SNP	*Single-Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único)
VA	Variável Aleatória
VAC	Variável Aleatória Contínua

LISTA DE TABELAS

Página

CAPITULO 1

CAPITULO 2

LISTA DE FIGURAS

Página

CAPITULO 1

CAPITULO 2

OLIVEIRA, Max Brandão de. **Bayes t e Bayes DE: Modelos bayesianos robustos de Seleção Genômica Ampla**. 2019. Tese (Doutorado em Ciência Animal) – Universidade Federal do Piauí, Teresina, 2019.

RESUMO

Os modelos utilizados nas predições genômicas assumem diferentes distribuições para o efeito dos marcadores, como normal, t e dupla-exponencial, no entanto utilizam apenas a distribuição normal para a variável resposta y_i. A distribuição $t_{(.)}$ apresenta uma simetria equivalente à normal, mas tem caudas pesadas, que confere menor sensibilidade a *outliers*, melhor adaptabilidade e maior variabilidade. Essas características podem favorecer o ajuste de um modelo mais robusto, assim como a dupla-exponencial, sendo que esta última ainda tem uma maior concentração em torno da média. Portanto, o objetivo com esta pesquisa foi desenvolver modelos Bayesianos de seleção genômica ampla usando a distribuição t (Bayes t) e a dupla-exponencial (Bayes DE) para a variável resposta. Para isso, foi proposto e desenvolvido um modelo Bayesiano no *software* R e foram utilizados parâmetros, como DIC, acurácia e análise residual, para quantificar a diferença entre os ajustes dos modelos propostos ao RRBLUP, já consolidado e que apresenta as mesmas características dos propostos. Para validação dos modelos, foram simulados dados genômicos que variaram de acordo com o tamanho da amostra em 1.000 gerações. Os dados foram gerados com três níveis de variância fenotípica: 5, 10 e 15. Além disso, os ajustes foram aplicados em amostras com tamanhos de 300, 1.000 e 2.000 animais. Os resultados apontam que, quando existem *outliers* na amostra, os modelos com distribuições t e Laplace são mais robustos. As médias residuais dos 3 modelos avaliados foram centrados em 0, mas a dispersão dos ajustes propostos foi inferior e, portanto, foram melhores. Em todos os cenários testados, os modelos propostos foram mais acurados em relação ao RRBLUP. O modelo Bayes t foi o mais acurado e qu apresentou menor variabilidade residual, principalmente para tamanhos amostrais menores. Os modelos também foram aplicados em uma amostra de 389 ovinos da raça Santa Inês, com a variável resposta área de olho de lombo. Os resíduos dos 3 se concentraram em torno de 0, de modo que o Bayes t foi o melhor, bem como foi o de menor dispersão residual e o de menor DIC entre os três ajustes. A correlação entre os valores preditos e observados no Bayes t foi 0,8006, enquanto nos modelos RRBLUP e no Bayes DE, foram 0,6835 e 0,6901, respectivamente. Portanto, os métodos propostos surgem como alternativas de modelos robustos para dados com presença de *outliers* e para tamanhos de amostras pequenos, em especial o Bayes t, que se mostrou melhor.

Palavras-chave: Características de carcaça, distribuição t, ovinos de corte, robustez, SNPs, valor genômico

OLIVEIRA, Max Brandão de. **Bayes t and Bayes DE: Bayesian robust models of Genomic Wide Selection.** 2019. Thesis (Animal Science Doctorate) – Universidade Federal do Piauí, Teresina, 2018.

ABSTRACT

The models used in genomic predictions assume different distributions for the effect of markers, such as normal, t and double-exponential, but only use the normal distribution for the response variable y_i. The $t_{(.)}$ Distribution has a symmetry equivalent to normal, but has heavy tails, which gives less sensibility to outliers, better adaptability and greater variability. These characteristics may favor the fit of a more robust model, as well as the double-exponential one, with the latter still having a higher concentration around the average. Therefore, the objective of this research was to develop Bayesian models of Genomic Wide Selection using the t (Bayes t) and double-exponential (Bayes DE) distribution for the response variable. For this, a Bayesian model was proposed and developed in the R software and parameters such as DIC, accuracy and residual analysis were used to quantify the difference between the adjustments of the proposed models to the already consolidated RRBLUP and presenting the same characteristics as those proposed. To validate the models, genomic data were simulated, which varied according to the sample size in 1,000 generations. Data were generated with three levels of phenotypic variance: 5, 10 and 15. In addition, adjustments were applied to samples with sizes of 300, 1,000 and 2,000 animals. The results show that when there are outliers in the sample, the models with t and Laplace distributions are more robust. The residual averages of the 3 models evaluated were centered at 0, but the dispersion of the proposed adjustments was lower and therefore were better. In all scenarios tested, the proposed models were more accurate than RRBLUP. The Bayes t model was the most accurate and showed the lowest residual variability, especially for smaller sample sizes. The models were also applied to a sample of 389 Santa Inês sheep, with the response variable loin eye area. Residues of the 3 were concentrated around 0, so Bayes t was the best, as well as the one with the lowest residual dispersion and the lowest DIC among the three fits. The correlation between the predicted and observed values in Bayes t was 0.8006, while in RRBLUP and Bayes DE models were 0.6835 and 0.6901, respectively. Therefore, the proposed methods appear as alternatives to robust models for outlier data and for small sample sizes, especially Bayes t, which was better.

Key-words: Carcass characteristics, t-Student distribution, cutting sheep, robustness, SNPs, genomic value

SUMÁRIO

1. INTRODUÇÃO

A criação de ovinos é uma das principais atividades pecuárias desenvolvidas no Nordeste brasileiro. Neste cenário, destacam-se os animais de raças deslanadas, que apresentam maior adaptabilidade às condições ambientais dessa região. Apesar do bom desempenho das raças ovinas criadas na região Nordeste, os índices produtivos de algumas destas raças têm sido mantidos abaixo de seu verdadeiro potencial produtivo. De acordo com Lobo (2019), parte da subutilização do potencial produtivo de raças ovinas como a Santa Inês se deve à falta de organização da cadeia produtiva da ovinocultura, em que uma das principais deficiências é a carência de investimentos em melhoria genética dos rebanhos.

Tradicionalmente, a maior parte das avaliações genéticas de ovinos no Brasil tem sido realizada com base apenas em informações fenotípicas e de pedigree (MCMANUS et al., 2010). Atualmente, com o desenvolvimento das tecnologias de genética molecular, tem se tornado possível acrescentar a informação genômica ao melhoramento genético tradicional, o que eleva a resposta à seleção, inclusive para animais que ainda não atingiram maturidade sexual ou que não dispõem de informações fenotípicas (MEUWISSEN; HAYES; GODDARD, 2016; RUPP et al., 2016). O uso da informação genômica também permite avaliar mecanismos genéticos de características de interesse zootécnico (WU et al., 2014).

Neste cenário, é possível identificar variantes causais (genes e/ou *loci* de características quantitativas – QTL – *Quantitative Trait Loci*) no genoma de um conjunto de indivíduos através de marcadores genéticos, principalmente do tipo SNP (Polimorfismo de Nucleotídeo Único). Com o uso de informações de SNPs, pode-se estimar o valor genético dos animais e selecionar diretamente aqueles que são superiores através de métodos estatísticos de seleção genômica ampla (GWS – *Genome Wide Selection*), propostos inicialmente por Meuwissen, Hayes e Goddard (2001).

A GWS consiste em um modelo estatístico complexo que permite avaliar o efeito dos SNPs sobre uma característica de interesse econômico. Desta forma, pode-se ter evidência da influência de uma ou mais variantes casuais sobre a expressão da característica, permitindo identificar quais alelos são benéficos (ou maléficos) para esta e, consequentemente, quais deverão ou não serer transmitidos para as gerações futuras. Através da estimação dos efeitos dos SNPs, é possível predizer os valores genéticos genômicos (GEBVs – *Genomic Estimates Breeding Value*) dos indivíduos para

características de interesse com maior acurácia, em relação à seleção baseada em métodos tradicionais. Além do modelo de GWS proposto por Meuwissen, Hayes e Goddard (2001), outros modelos de GWS foram propostos por Lee et al. (2008), De Los Campos et al. (2009), Li et al. (2011); Segura et al. (2012); Rakitsch et al. (2013) e Li et al. (2017), por exemplo.

As metodologias desenvolvidas para GWS apresentam alta complexidade e funcionalidade em termos computacionais e estatísticos, mas são rígidas quanto à escolha da distribuição associada à variável resposta e quanto ao modelo escolhido, já que todos os modelos propostos utilizam a distribuição normal. A especificação incorreta da distribuição e, consequentemente, do modelo, pode levar a conclusões equivocadas, uma vez que pode afetar fortemente a distribuição *a posteriori* (ANDRADE; OMEY; AQUINO, 2017). Portanto, selecionar uma variável cuja distribuição não seja descrita pelo modelo pode levar à tomada de decisões erradas, como a seleção de variantes indesejadas ou a seleção de animais impróprios para fins de melhoramento genético.

Além da distribuição de probabilidade, o modelo também é sensível a alguns conflitos de informação que podem prejudicá-lo, como por exemplo, *outliers*, que influenciam fortemente os parâmetros (ANDRADE; O'HAGAN, 2011). Lindley (1968) foi o primeiro a identificar esse comportamento e sugeriu o uso da distribuição t de Student para resolver esse conflito, já que a mesma apresenta um fenômeno chamado de cauda pesada e, portanto, conferiria mais robustez ao modelo. A distribuição dupla-exponencial ou Laplace detém a mesma característica e, portanto, também pode favorecer o melhor ajuste do modelo. Desta forma, as duas distribuições são avaliadas como estratégias para conferir mais robustez aos modelos.

Neste sentido, modelos de seleção genômica ampla que possibilitem a utilização das distribuições t e Laplace para características relacionadas à produção animal podem representar uma alternativa promissora para a avaliação genética de ovinos de corte devido ao pequeno volume de dados, normalmente utilizados no Brasil para características de produção de carne, as quais apresentam dificuldade para a mensuração.

Para avaliar o rendimento de carcaça e deposição de músculo, a característica área de olho de lombo (AOL) mensurada por ultrassonografia em tempo real é uma medida apropriada para conduzir estratégias de melhoramento, pois apresenta estimativas de herdabilidade que variam de moderada a alta magnitude quando obtidas com inclusão de dados genômicos (GORDO et al., 2016; SILVA et al., 2017).

Portanto, o objetivo com esta pesquisa é propor um modelo robusto de Seleção Genômica Ampla que utilize as distribuições t de Student e Laplace para a variável resposta (área de olho de lombo - AOL) e produza estimativas mais consistentes a partir de características produtivas do animal.

A Tese é apresentada em Capítulos, de acordo com as Normas do Programa de Pós-Graduação em Ciência Animal da Universidade Federal do Piauí. Em princípio, é apresentada a Revisão de literatura, onde são abordados os conteúdos referentes ao estudo, que fornecem embasamento e subsídio teórico que favoreçam uma melhor compreensão por parte do leitor. Em seguida são apresentados os 2 capítulos, sendo o primeiro apresentando método em dados simulados e o segundo com título "Bayes t e Bayes DE: Modelos robustos para seleção genômica ampla", formatado nas normas da revista *Journal of Animal Breeding and Genetics;* e o segundo, na mesma formatação e para a mesma revista, com o título "Modelos robustos de seleção genômica ampla em ovinos da raça Santa Inês".

2. REVISÃO DE LITERATURA

2.1. Panorama da ovinocultura de corte no Brasil

A produção de carne ovina no Brasil aumentou 2,1% entre os anos de 2016 e 2017. O contingente de ovinos criados no Brasil, segundo o IBGE (2018), é de aproximadamente 13,8 milhões de cabeças, sendo que mais de 9 milhões de animais são criados para produção de carne na região Nordeste, onde o rebanho ovino cresceu 15,94% entre 2006 e 2017. De acordo com a FAO (2018), a produção de carne ovina em 2017 foi de aproximadamente 15 milhões de toneladas. Mesmo diante desse crescimento, o consumo da carne ovina ainda é inferior ao consumo das carnes bovina, suína e de aves (CONSTANTINO et al., 2018). Essa disparidade é fruto do maior investimento aplicado nestes sistemas de criação em comparação à ovinocultura (ALVES et al., 2014).

Enquanto países como China, Austrália e Nova Zelândia têm avançado no melhoramento genético de características de produção em ovinos (ROWE, 2010; AUVRAY et al., 2014; BRITO et al., 2017a, b), no Brasil, os sistemas de produção de ovinos têm a cadeia produtiva atrasada e desorganizada (LOBO, 2019). Esta situação

reflete nos índices de produtividade, na qualidade dos produtos, na falta de regularidade da oferta e na informalidade da comercialização (LUCENA et al., 2018).

No Nordeste do Brasil, os ovinos deslanados representam uma importante fonte de proteína para as populações locais, principalmente na zona rural. Devido às condições climáticas da região, é importante a criação de animais de raças adaptadas às condições de clima tropical e semiárido, como os ovinos da raça Santa Inês, que requerem baixa manutenção, perda mínima de produção durante estresse calórico, alta eficiência produtiva, resistência a doenças, longevidade e baixas taxas de mortalidade (MCMANUS et al., 2011).

2.2. A raça Santa Inês

A raça Santa Inês foi originada na região Nordeste do Brasil a partir de cruzamentos intercorrentes das raças Somalis Brasileira, Bergamácia, Morada Nova e ovinos Sem Raça Definida (ABSI, 2019). Morfologicamente, o ovino Santa Inês (Figura 1) é caracterizado por apresentar pelas orelhas longas, cabeça semi-convexa e mocha (em ambos os sexos), tronco grande e comprido, região dorso-lombar longa e retilínea com boa cobertura muscular, garupa comprida, cauda média, pernas fortes, firmes e com articulações fortes e bons aprumos. As fêmeas possuem peso corporal variando de 40 a 90Kg e os machos podem atingir até 120Kg. São encontradas animais desta raça com pelagem nas cores vermelha, preta, branca e chitada (SOUSA; LOBO; MORAIS, 2003).

Figura 1. Fenótipo padrão do ovino Santa Inês (esquerda) e diferentes padrões de pelagens da raça (direita)

Fontes: Foto da esquerda - http://www.arcoovinos.com.br/index.php/mn-srgo/mn-padroesraciais/40-santa-ines; Foto da direita: Luciano Silva Sena

O padrão racial definido pela Associação Brasileira de Criadores de Ovinos (ARCO, 2014) é amplo e sujeito a interpretação, tornando raça susceptível à perda de características associadas à rusticidade e à adaptação em detrimento a ganhos obtidos por meio de cruzamentos com raças comerciais exóticas (McMANUS; PAIVA; ARAÚJO, 2010).

Os animais da raça Santa Inês apresentam notáveis características reprodutivas e de adaptação, devido à sua rusticidade (SOUSA; LÔBO; MORAIS, 2003). Além de rusticidade, os animais desta raça apresentam prolificidade, precocidade e habilidade materna favoráveis. Dentre as raças ovinas deslanadas presentes no Brasil, a raça Santa Inês apresenta maior porte, o que favorece a habilidade materna das matrizes e a ocorrência de partos múltiplos (PAIVA et al., 2005). Essas vantagens promoveram o destaque do ovino Santa Inês na região Nordeste do Brasil e também chamaram a atenção de produtores das regiões Sudeste e Centro-Oeste (MORAIS, 2000). Em função de seu notável tamanho corporal, taxa de crescimento, adaptabilidade, capacidade reprodutiva, entre outros aspectos, a raça Santa Inês passou a ser criada em quase todas as regiões do país. No entanto, a raça ainda apresenta produtividade abaixo de seu verdadeiro potencial, devido à escassez de iniciativas de melhoramento genético (REGO NETO et al., 2018; LOBO, 2019).

Embora a principal finalidade da raça Santa Inês seja a produção de carne, características de carcaça e qualidade de carne ainda têm sido pouco avaliadas como critério de seleção nesta raça. Por vários anos, a seleção dentro da raça Santa Inês tem sido praticada quase que somente com base em tamanho e peso corporal, devido ao fato de que estas características são consideradas de maior interesse econômico pelos criadores (COSTA JÚNIOR et al., 2006; TEIXEIRA NETO et al., 2016). No entanto, para atender às exigências do mercado consumidor, é importante realizar a obtenção de informações de características que indicam de forma mais precisa o rendimento de carcaça em animais de corte, para a implantação de estratégias de melhoramento genético (SOUZA et al., 2016; BRITO et al., 2017b).

Alguns trabalhos conduzidos para avaliação de carcaça com uso de ultrassonografia *in vivo* em ovinos da raça Santa Inês mostraram que características mensuradas no olho de lombo (músculo *Longissimus dorsi*) apresentam variabilidade genética suficiente para serem utilizadas na seleção para melhoria de carcaça nos animais desta raça (SENA et al. 2016; FIGUEIREDO FILHO et al., 2016, 2017).

Há vários anos, a avaliação genética de ovinos Santa Inês tem sido realizada apenas com base em informações fenotípicas e de pedigree (SOUSA et al., 1999; SARMENTO et al., 2006, 2010; CARVALHO et al., 2014; FIGUEIREDO FILHO et al., 2016; SENA et al., 2016). Estudos que incluem a informação genômica ainda são escassos em relação à raça Santa Inês (BIAGIOTTI, 2016; BERTON et al., 2017; REGO NETO, 2017; AMORIM et al., 2018; ALVARENGA et al., 2018; MEIRA et al., 2018; ROVADOSCKI et al., 2018; SANTOS, 2018). Estes estudos têm contribuído para melhor entendimento de mecanismos genéticos de diferentes características na raça Santa Inês, assim como para a seleção mais eficiente de animais desta raça.

2.3. Distribuições de probabilidade

As características envolvidas nos modelos de seleção genômica são tratadas como variáveis aleatórias (VA) contínuas que possuem uma função densidade de probabilidade (FDP) com parâmetro θ desconhecido. Os modelos são utilizados para estimar valores para θ com base na informação amostral, como por exemplo, pode-se estimar os efeitos dos SNPs, do sexo, do grupo contemporâneo e outros. Portanto, é necessário conhecer as principais FDPs empregadas nos principais modelos (BUSSAB; MORETTIN, 2017).

Uma variável aleatória pode ser qualitativa ou quantitativa. Esta segunda divide-se em discreta e contínua. Uma VA é definida como discreta quando o número de valores possíveis que ela assume for finito ou infinito enumerável. Por outro lado, se a VA for contínua, o conjunto dos valores possíveis é infinito não enumerável (MEYER, 2006).

Associada a uma variável aleatória contínua (VAC) X qualquer, tem-se o conceito de função densidade de probabilidade (FDP), que é uma função contínua, não negativa e que satisfaz às propriedades: $f(x) \geq 0 \; \forall \; x \in \; X$ e $\int_{-\infty}^{\infty} f(x)dx = 1$. As densidades têm parâmetros que são normalmente usados para obtenção de medidas importantes da VA, como as de localização e de dispersão, usadas para retratar a variável ou característica de interesse (MONTGOMERY et al., 2011). As principais variáveis aleatórias contínuas utilizadas nos modelos de seleção genômica são: normal, dupla-exponencial, t-Student e qui-quadrado invertida.

Distribuição normal

Uma VAC X tem distribuição normal, denotada por $X \sim N(\mu, \sigma^2)$, se sua FDP for dada por

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], x \in \mathbb{R}$$

em que $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ são os parâmetros da VAC X e representam média e a variância de X, respectivamente. A distribuição normal tem propriedades desejadas em vários aspectos, como por exemplo, é simétrica e o estimador $\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i/n$ é não viesado de variância mínima. Além disso, pelo Teorema do Limite Central (TLC), várias distribuições convergem assintoticamente ($n \to \infty$) em distribuição para a normal.

Distribuição dupla-exponencial (Laplace)

Uma VAC X segue distribuição dupla-exponencial, denotada por $X \sim DE(\mu, \sigma)$, se sua FDP for dada por

$$f(x|\mu,\sigma) = \frac{1}{2\sigma} \exp\left(\left|\frac{x-\mu}{\sigma}\right|\right), x \in \mathbb{R}$$

onde $\mu \in \mathbb{R}$ e $\sigma > 0$ são os parâmetros de localização e de escala de X, respectivamente. Assim, como a normal, a curva que representa a densidade de X é simétrica em torno de μ, mas tem um decaimento bem mais acentuado do que a normal, conferindo caudas mais pesadas à distribuição.

Distribuição t-Student

Uma VAC X segue distribuição t-Student escalonada, denotada por $X \sim t_{(\mu,\sigma,v)}$, se sua FDP for dada por

$$f(x|\mu,\sigma,v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left[\frac{v + \left(\frac{x-\mu}{\sigma}\right)^2}{v}\right]^{-\left(\frac{v+1}{2}\right)} \quad x \in \mathbb{R}$$

onde $\mu \in \mathbb{R}$, $\sigma > 0$ e $v > 0$ são os parâmetros da densidade de X. A média e a variância de X são μ e $\sigma^2\left(\frac{v}{v-2}\right)$, respectivamente. Assim, como a normal e a Laplace, a curva de X é simétrica em torno de μ, mas tem caudas pesadas, que reflete uma maior variabilidade em relação a distribuição normal.

O gráfico apresentado na Figura 2 exibe o comportamento das curvas das distribuições normal, t e dupla- exponencial, deixando claro que as três são simétricas em torno na média, que no caso, em particular, é zero. A dupla-exponencial tem o pico mais alto, seguida pela normal e pela t. Quanto à dispersão, a t e a Laplace tem caudas mais próximas, sendo superiores à normal, expressando maior variabilidade.

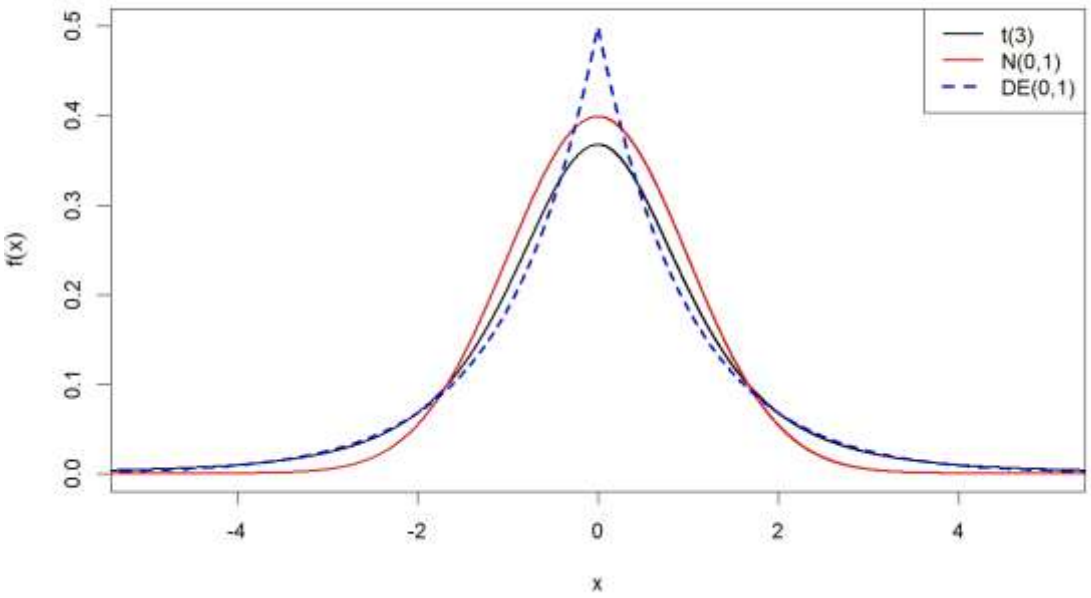

Figura 2. Densidades das distribuições $t_{(3)}$, $N(0,1)$ e $DE(0,1)$ utilizadas como distribuições *a priori* nos modelos de GWS

Devido a esses comportamentos, cada uma é empregada em cenários específicos. A Laplace tem uma maior concentração em torno da média, significando uma quantidade de pontos em torno da média. A distribuição t, por ter caudas mais pesadas, é usada para expressar uma maior dispersão ou falta de informação a respeito de um parâmetro, enquanto a normal detém um comportamento mais geral e conservador.

Distribuição Qui-quadrado invertida escalonada

Uma VAC X segue distribuição Qui-quadrado escalonada, denotada por $X \sim \chi^{-2}_{(\nu,\tau^2)}$, se sua FDP for

$$f(x|v,\tau^2) = \frac{\left(\frac{\tau^2 v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} \frac{\exp\left[-\frac{v\tau^2}{2}\right]}{x^{1+\frac{v}{2}}}, x > 0$$

em que $v > 0$ representa os graus de liberdade e $\tau^2 > 0$ é o parâmetro de escala. Devido a seu suporte não negativo, ela é usada como priori para parâmetros associados a medidas de dispersão, como desvio-padrão ou variância.

As quatro distribuições abordadas são adotadas em modelos Bayesianos nas mais diversas áreas, como climatologia, melhoramento genético animal e vegetal, modelos de valores extremos, psicometria e outros. Para obtenção das estimativas dos parâmetros é necessário usar o processo de inferência Bayesiana.

2.4. Inferência Bayesiana

A necessidade de mais pesquisas e os avanços científicos e tecnológicos contribuíram para o manuseio de grandes volumes de dados. Os estudos passaram a ser conduzidos de forma multivariada, onde são analisadas diversas variáveis e as relações entre elas. A grande quantidade de informação, por outro lado, exige o uso de conceitos e métodos estatísticos e computacionais que sintetizem os resultados pertinentes à pesquisa de forma objetiva. Na estatística, um conceito fundamental que dá suporte às principais metodologias é o de inferência estatística.

A inferência estatística é um conjunto de técnicas e métodos utilizados na estimação de um ou mais valores para um parâmetro desconhecido $\theta \in \Theta$ associado a uma variável aleatória X com função de densidade $f(x|\theta)$ (BOLFARINE; SANDOVAL, 2010). Existem duas abordagens para inferência, a frequentista e a Bayesiana. Na primeira, o conceito de probabilidade envolve uma sequência de repetições para um determinado evento, tratado como um subconjunto de Θ. A teoria baseia-se na regularidade estatística das frequências relativas e sustenta que a probabilidade de um dado acontecimento pode ser medida observando a frequência relativa do mesmo acontecimento, numa sucessão numerosa de experiências idênticas e independentes (BLASCO, 2001). Além disso, o estimador $\hat{\theta}$ assume um valor fixo de acordo com o método de estimação adotado.

A abordagem Bayesiana atribui distribuições de probabilidade a cada parâmetro $\theta_i \sim p(.)\ \forall\ i = 1, \ldots, p$ a ser estimado, de modo que uma estimativa de θ_i seria uma medida de localização da distribuição associada ao parâmetro, como média, moda,

máximo (mínimo) ou mediana. A vantagem dessa estratégia é que o estimador segue uma distribuição que descreve o comportamento do parâmetro com medidas de localização e de dispersão que podem ser controladas e ajustadas de acordo com o conhecimento que se tem sobre θ. Essas distribuições são chamadas de *priori*.

As distribuições *a priori* $p(\theta_i)$ são combinadas com a função de verossimilhança $f(\boldsymbol{y}|\Theta)$, que sintetiza a informação amostral do vetor de observações $\boldsymbol{y}$ condicionada aos parâmetros do modelo Θ. De posse dessas componentes, é utilizado o teorema de Bayes para obter-se a distribuição *a posteriori* $p(\Theta|y) \propto f(\boldsymbol{y}|\Theta)p(\Theta)$, que representa a distribuição de probabilidade dos parâmetros associada à informação de $\boldsymbol{y}$ **(**FARIA et al., 2007**)**.

Essa estratégia é bastante conveniente, já que a *posteriori* fornece estimativas considerando a informação amostral e as distribuições *priori*, que podem acrescentar a subjetividade acerca dos parâmetros. Estas fontes de informação alimentam os métodos empregados na obtenção da *posteriori* e podem trazer vantagens em relação aos métodos frequentistas, como por exemplo, trabalhar com amostras pequenas e ajustar valores de conhecimento prévio nas *priori*, o que favorece às estimativas (GIANOLA; FERNANDO, 1986).

Para a obtenção da *posteriori* são usados métodos iterativos como o de Monte Carlo via Cadeias de Markov (MCMC), um dos mais utilizados. Dentro do MCMC, os dois algoritmos mais adotados para obtenção das estimativas são o amostrador de Gibbs e o Metropolis de Hastings. De acordo com Gianola et al. (1994), os métodos frequentistas, como REML (*Restricted Estimation Maximum Likelihood*), obtêm apenas estimativas pontuais da variância genética com uma única medida de erro, que teria sentido em amostras relativamente grandes.

2.5. Aplicação de dados genômicos em ovinos

Com a implementação da Seleção Genômica Ampla (GWS – *Genome Wide Selection*) por Meuwissen, Hayes e Goddard (2001), passou a ser possível a avaliação genética com uso da informação genômica. Na GWs, o genótipo é representado por painéis densos de polimorfismos de nucleotídeo único (SNPs). Com isso, é possível estimar os efeitos de cada SNP com base em uma amostra de animais e, com estas informações, pode-se predizer os valores genômicos dos candidatos à seleção.

A aplicação de seleção genômica na pecuária tem aperfeiçoado a produção de diferentes espécies em função do aumento da acurácia na obtenção da estimativa do valor genético e da redução dos custos com testes de reprodutores e do intervalo de geração (JONAS; DE KONING, 2015; MEUWISSEN; HAYES; GODDARD, 2016). Com a utilização de informação genômica, Daetwyler et al. (2012) e Brito et al. (2017a), por exemplo, relataram a viabilidade da utilização de informação genômica para o aumento do progresso genético em ovinos de corte, em relação aos métodos tradicionais. Com a inclusão de informação genômica em avaliações genéticas de ovinos, as estimativas de ganhos em acurácia variam entre 0,05 e 0,27 para diferentes características (DAETWYLER et al., 2012;AUVRAY et al., 2014; BALOCHE et al., 2014; RUPP et al., 2016).

As informações genômicas também têm sido utilizadas para Estudos de Associação Genômica Ampla, que consistem basicamente na avaliação dos efeitos de variações existentes no genoma (principalmente SNPs) sobre fenótipos de interesse econômico, bem como os mecanismos genéticos responsáveis pela expressão dessas características (ZHANG et al., 2012). Com a utilização de GWAS, Matika et al. (2016) identificaram regiões genômicas associadas a características de carcaça obtidas por tomografia computadorizada, em ovinos da raça Scottish-Blackface. Em ovinos Santa Inês, Berton et al. (2017) identificaram regiões genômicas relacionadas a atividades do sistema imune e resistência parasitária.

Apesar dos benefícios da GWAS, Zhang et al. (2012) apontam divergências entre resultados de estudos de GWAS para a mesma característica, que podem ser justificadas pelotamanho da população utilizada, densidade dos marcadores, estrutura genética populacional, modelos estatísticos e detecção de falsos positivos. Tais inconsistências, de acordo com Sharma et al. (2015), podem ser minimizadas com o controle de qualidade dos dados, sendo essa etapa considerada uma das mais importantes em estudos que incluem a utilização de informação genômica para GWS e GWAS. Além disso, para a prática de seleção genômica, a escolha do modelo mais apropriado para cada característica em cada população poderá impactar diretamente a acurácia das predições genômicas (DE LOS CAMPOS et al., 2013).

2.6. Principais modelos de GWS

A estrutura de dados é denotada como $\{y_i, x_{r_i}, x_{l_i}\}_{i=1}^{n}$, onde y_i é o fenótipo da característica de interesse medida no indivíduo, x_{r_i} é o vetor de covariáveis que é tratado como uma regressão bayesiana tradicional, com distribuição normal como *priori* e variância comum para todos os parâmetros, e x_{l_i} é o conjunto de covariáveis cujos efeitos são associados aos marcadores SNPs (HADFIELD, 2010). Desta forma, o modelo geral é dado por (1)

$$y_i = \mu + \boldsymbol{x}'_{\boldsymbol{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x}'_{\boldsymbol{l_i}}\boldsymbol{\beta_l} + \varepsilon_i, \quad (1)$$

em que μ é o intercepto, $\boldsymbol{\beta_r}$ e $\boldsymbol{\beta_l}$ são parâmetros associados a $\boldsymbol{x}'_{\boldsymbol{r_i}}$ e $\boldsymbol{x}'_{\boldsymbol{l_i}}$, respectivamente, e ε_i é o resíduo do modelo, considerado independente e identicamente distribuído, e a função de verossimilhança é escrita como (2)

$$p(y \mid \mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} N\left(y_i \mid \mu + \boldsymbol{x}'_{\boldsymbol{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x}'_{\boldsymbol{l_i}}\boldsymbol{\beta_l}, \sigma_\varepsilon^2\right). \quad (2)$$

Seguindo o modelo aditivo denotado em (1), diversas propostas para a predição de valores genômicos através de marcadores SNPs foram desenvolvidas, dentre elas, as mais utilizadas são: a Melhor Predição Linear Não-Viesada Genômica, conhecida como regressão de cumeeira (GBLUP ou RRBLUP); Penalização Bayesiana (*Bayesian shrinkage*), como por exemplo, BayesA, BayesB, BayesCπ, BayesC, LASSO (operador de penalização de menor ângulo e seleção - *least angle shrinkage and selection operator*) e LASSO Bayesiano.

Estes modelos sugerem a distribuição normal para y_i, para as *prioris* dos efeitos sistemáticos e para o mérito genômico do animal; para os componentes de variância, é adotada a distribuição qui-quadrado invertida χ^{-2}; para os marcadores, são utilizadas a normal, a t-Student e a dupla exponencial, todas simétricas, conforme observa-se Figura 2, evidenciando o padrão de comportamento das densidades (CHEN; HUANG, 2014). Destaca-se que a dupla-exponencial é adotada como uma mistura de distribuições (normal e exponencial).

Além dessas diferenças mais perceptíveis, de acordo com Gianola et al. (2009), os modelos BayesA e BayesB têm dependência das *prioris* da variância do marcador, já o modelo BayesCπ é menos sensível à pressuposição *a priori* da variância do marcador, já que os SNPs têm a mesma variância, bem como a proporção de SNPs sem efeito (π) tem a *priori* uniforme (HABIER et al., 2011).

No modelo BayesC, a constante de penalização π é um valor fixo (FERNANDO; GARRICK, 2013), que possibilita a detecção mais acurada de QTLs quando comparado ao BayesCπ, principalmente para QTLs com herdabilidade moderada ou alta (VAN DEN

BERG; FRITZ; BOICHARD, 2013). De acordo com Mehrban et al. (2017), os métodos Bayesianos têm a desvantagem de apresentarem a necessidade de definição as distribuições a *priori*. Essa exigência é resolvida no método LASSO Bayesiano, que necessita de menos informação (LEGARRA et al., 2011). Mesmo com a dificuldade na proposição das *prioris*, os métodos bayesianos são vantajosos, já que os parâmetros assumem uma distribuição que expressa o grau de incerteza acerca da amostra, além de utilizar métodos iterativos eficientes e consistentes na obtenção das estimativas, como por exemplo, o Método de Monte Carlo via Cadeias de Markov (MCMC), que é o mais adotado.

Nos modelos tratados, o fenótipo assume distribuição normal $N(\mu, \sigma^2)$. Tal pressuposição é adotada devido a distribuição ter propriedades ótimas em relação aos estimadores, como invariância, estimadores consistentes e com variância mínima. Além disso, pelo teorema do limite central, vários conjuntos de dados convergem assintoticamente em distribuição para a normal (BOLFARINE; SANDOVAL, 2010). Porém, a distribuição apresenta caudas leves, o que a torna sensível aos conflitos de informação oriundos do banco de dados, como por exemplo, os *outliers*.

A presença de *outliers* no ajuste dos modelos pode prejudicar as estimativas. Desta forma, pode-se optar pela exclusão dos mesmos, levando em consideração que os valores discrepantes podem ser erros provenientes da medição, da coleta ou da amostragem. No entanto, esses valores podem representar justamente os animais superiores presentes na amostra e, portanto, retirá-los da análise pode representar um equívoco no processo de seleção.

A presença e a interferência dos valores discrepantes foram inicialmente observadas por De Finetti (1961). Como alternativa para contornar esse problema, existem distribuições com caudas pesadas, que são mais robustas, expressam mais incerteza acerca da variável de interesse e são menos sensíveis aos conflitos de informação (LINDLEY, 1968). Além disso, as distribuições de caudas pesadas também são adotadas na utilização de bancos de dados com pequenos tamanhos amostrais, já que a maior variabilidade representa a escassez de informação sobre os parâmetros.

Em seguida, Dawid (1973) estabeleceu condições e apresentou indicadores favorecendo o uso da distribuição t-Student como *priori* para parâmetros de localização. Corroborando com o estudo, O'Hagan (1979, 1988, 1990) e O'Hagan e Le (1994), apresentaram parâmetros adicionais que validassem a proposição. Andrade e O'Hagan (2011), O'Hagan e Pericchi (2012), Andrade e Omey (2013) e Andrade e Omey (2016),

por sua vez, deram uma contribuição ao sugerirem modelos robustos a partir da distribuição t. Além disso, a distribuição Laplace, que apresenta simetria semelhante a normal e a t, tem um pico mais elevado em torno da média, que pode ser interessante em cenários com baixa variabilidade.

Nesse contexto, surge a necessidade de estudar e avaliar a suposição da distribuição t de Student e da Laplace para a variável resposta y_i dos métodos de GWS com o propósito de tornar o modelo mais robusto e mais resistente ao conflito de informação. Teugels (1975), Embrechts, Klüppelberg e Milosch (1997) e Goldie e Klüppelberg (1998), apresentaram uma vasta discussão sobre propriedades e aplicações de distribuições de caudas pesadas. Neyman e Scott (1971) mostraram uma relação entre as distribuições de caudas pesadas e a propensão a outliers. Por definição, uma variável aleatória X, com função de distribuição F, tem cauda pesada se $\lim_{x\to\infty} e^{\lambda x}(1 - F(x)) = \infty \ \forall \ \lambda > 0$,que é equivalente a $M_X(s) \to \infty \ \forall \ s > 0$, em que $M_X(s)$ é a função geradora de momentos de associada a F (ASMUSSEN, 2003). Portanto, a ideia consiste em propor 2 modelos de Seleção Genômica Ampla cujas distribuições associadas à variável resposta y_i sejam a t-Student e dupla-exponencial (Laplace), denominados de Bayes t e Bayes DE, respectivamente.

2.7. Comparação entre modelos

Os modelos são comparados de forma tradicional, através de indicadores de diagnóstico dos modelos, como análise residual, DIC (*Deviance Information Criterion*), relação entre valores preditos e observados, entre outros. A análise residual retrata a diferença entre os valores estimados e os observados ($\hat{y} - y$). Neste caso, é interessante que a média residual convirja para 0, enquanto sua dispersão seja a menor possível. Ou seja, espera-se que o erro tenda a 0 com baixa variabilidade (FERNANDO; CARRICK, 2013; ESPIGOLAN et al, 2013; GRASSO, 2014).

O DIC é uma generalização do AIC (*Akaike Information Criterion*) e é recomendado para comparações entre ajustes Bayesianos, em que as distribuições *a posteriori* são obtidas por métodos de simulação via Monte Carlo via Cadeia de Markov (MCMC). O DIC é definido como $D(\boldsymbol{\theta}) = -2\log[p(\boldsymbol{y}|\boldsymbol{\theta})] + C$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos, $\boldsymbol{y}$ representa os dados, $p(\boldsymbol{y}|\boldsymbol{\theta})$ é a função de verossimilhança

e C é uma constante que normalmente é desprezada ao comparar os modelos. O ajuste de menor DIC é considerado o melhor.

A relação entre os valores preditos e observados é utilizada como critério de comparação. O par ordenado é plotado $(\hat{y}, y)$ no plano cartesiano e é desejado que os pontos se comportem de acordo com a reta $y = x$, indicando que o ajuste foi coerente. Um reflexo desse comportamento é a correlação entre $\hat{y}$ e y, de modo que quanto mais próximo de 1, melhor o ajuste. Os resíduos também podem estar diretamente relacionados a estes resultados, já que quanto mais semelhantes são $\hat{y}$ e y, maior a correlação entre eles e mais próximos de 0 são os resíduos (GIANOLA; FERNANDO, 1986, BALAKRISHNAN; RISTIĆ, 2016).

No caso de dados obtidos por meio de simulação, os valores genômicos verdadeiros dos animais são fornecidos. Portanto, uma forma de avaliar o ajuste é através da acurácia, dada pela correlação entre o valor real o estimado pelo modelo. Quanto mais próximo de 1, melhor o modelo e mais acurado. Destaca-se que nem sempre os modelos coincidem nos resultados dos indicadores usados no diagnóstico do ajuste, sendo necessário avaliar isolada e conjuntamente os parâmetros de comparação entre eles (HABIER et al., 2011; GIANOLA; RODRIGUEZ-ZAS; SHOOK, 1994).

O indicador mais importante para comparação entre os modelos é a acurácia. Para os dados simulados, de posse do valor genômico verdadeiro do animal, a acurácia é obtida em função da correlação entre o valor verdadeiro e o valor estimado pelo modelo. Para os dos reais, onde não se sabe o valor verdadeiro, adota-se

$$Acc_j = 1 - \sqrt{\frac{PEV_{ij}}{\sigma^2_{a_j}}},$$

em que Acc_j é a acurácia associada à característica j, PEV_{ij} é o erro padrão de predição para o valor genômico estimado para o animal i e $\sigma^2_{a_j}$ é a variância genética aditiva para a variável j (BIF, 2018).

3. REFERÊNCIAS

ABSI – ASSOCIAÇÃO BRASILEIRA DE SANTA INÊS, 2019. **Origem da raça**. Disponível em: http://www.absantaines.com.br/a-raca/origem.

ALVARENGA, A. B. et al. Linkage disequilibrium in Brazilian Santa Inês breed, *Ovis aries*. **Scientific Reports**, v.8, n.8851, 2018.

ALVES, L. G. C.; OSÓRIO, J. C. S.; FERNANDES, A. R. M.; RICARDO, H. A.; CUNHA, C. M. Produção de carne ovina com foco no consumidor. **Enciclopédia Biosfera**, v. 10, n. 18, p. 2399-2415, 2014.

AMORIM, S. T. et al. Genomic study for maternal related traits in Santa Inês sheep breed. **Livestock Science**, v.217, p. 76–84, 2018.

ANDRADE, J. A. A.; O'HAGAN, A. Bayesian Robustness Modeling Using Regularly Varying Distributions. **Bayesian Analysis**, v. 1, p. 169-188, 2006.

ANDRADE, J. A. A.; O'HAGAN, A. Bayesianro bustness modelling of location and scale parameters. **Scandinavian Journal of Statistics**, v. 38, p. 691-711, 2011.

ANDRADE, J. A. A. AND OMEY, E. Modelling conflicting information using subexponential distributions and related classes, **Annals of the Institute of Statistics and Mathematics**, v. 65, p. 491-511, 2013.

ANDRADE, J. A. A. AND OMEY, E. Resolution of conflict of information using O-regularly varying functions. **Statistica Neerlandica**, 2016.

ANDRADE, J. A. A.; OMEY, E.; AQUINO, C. T. M. Bayesian robustness modelling using the floor distribution. **REVSTAT Statistical Journal**, v. 16, p. 1-17, 2017.

ARCO. **Associação Brasileira de Criadores de Ovinos: Padrões raciais**. Disponível em: http://www.arcoovinos.com.br/index.php/mn-srgo/mn-padroesraciais/40-santa-ines. Acesso em 06.12.2018**.**

ASMUSSEN, S. **Applied Probability and Queues**. Springer, Berlin, 2003.

AUVRAY, B. MCEWAN, J. C.; NEWMAN, A. N.; DODDS, K. G. Genomic prediction of breeding values in the New Zealand sheep industry using a 50K SNP chip. **Journal of Animal Science**, v. 92, n. 10, p. 4375–4389, 2014.

BALOCHE, G. et al. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep**. Journal of Dairy Science**, v. 97, p. 1107-1116, 2014.

BERTON, M. P. et al. Genomic regions and pathways associated with gastrointestinal parasites resistance in Santa Inês breed adapted to tropical climate. **Journal of Animal Science and Biotechnology**, v. 8, n. 73, 2017.

BIAGIOTTI, D. **Associação e seleção genômica ampla em ovinos Santa Inês para características relacionadas a resistência à endoparasitas**. Teresina: Centro de Ciências Agrárias, Universidade Federal do Piauí, 73 p. Tese de Doutorado, 2016.

BLASCO, A. The Bayesian controversy in animal breeding. **Journal of Animal Science**, v. 79, p. 2023-2046, 2001.

BOLFARINE, H. SANDOVAL, M. C. **Introdução à Inferência Estatística**. Editora SBM, 2ª ed., São Paulo, 2010.

BRITO, L. F.; CLARKE, S. M.; MCEWAN, J. C.; MILLER, S. P.; PICKERING, N. K.; BAIN, W. E.; DODDS, K. G.; SARGOLSAEI, M.; SCHENKEL, F. S. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. **BMC Genetics**, v. 18, n. 7, 2017a.

BRITO, L. F.; MCEWAN, J. C.; MILLER, S. P.; BAIN, W. E.; LEE, M.; DODDS, K. G.; NEWMAN, S.A., PICKERING, N.; SCHENKEL, F. S.; CLARKE, S. Genetic parameters for various growth, carcass and meat quality traits in a New Zealand sheep population. **Small Ruminant Research**, n. 154, p. 81–91, 2017b.

BUSSAB, W.O.; MORETTIN, P. A. **Estatística básica**. São Paulo, Editora Saraiva, ed. 9, 2017.

CHEN, C. et al. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. **Journal of Animal Science**, v.89, p. 23-28, 2014.

COSTA JÚNIOR, G.S.; CAMPELO, J.E.G.; AZEVÊDO, D.M.M.R.; MARTINS FILHO, R.; CAVALCANTE, R.R.; LOPES, J.B.; OLIVEIRA, M.E. Caracterização morfométrica de ovinos da raça Santa Inês criados nas microrregiões de Teresina e Campo Maior, Piauí. **Revista Brasileira de Zootecnia**, v.35, n.6, p.2260-2267, 2006.

DAETWYLER, H. D. et al. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. **Genetics Selection Evolution**, v.44, n.33, 2012.

Dawid, A. P. Posterior Expectations for Large Observations. **Biometrika**, 60, 3, 664-667, 1973.

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.

DE LOS CAMPOS, G.; HICKEY, J. M.; WONG, R. P.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**. v. 193, p. 327–345, 2013.

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.

DE FINETTI, B. The Bayesian Approach to the Rejection of Outliers. **Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability**, Volume 1: Contributions to the Theory of Statistics, 199-210, University of California Press, Berkeley, Calif., 1961. https://projecteuclid.org/euclid.bsmsp/1200512167

Espigolan, R.; Baldi, F.; Boligon, A. A.; Souza, F. R.; Gordo, D. G.; Tonussi, R. L.; Cardoso, D. F.; Oliveira, H. N.; Tonhati, H.; Sargolzaei, M. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC Genomics**, 14:305, 2013.

EMBRECHTS, P., KLÜPPELBERG, C., MILOSCH, T. **Modelling Extremal Events**. Springer, New York, 1997.

FAO - FOOD AND AGRICULTURE ORGANIZATION. **Meat Market Review**. Rome, 2018. Disponível em: http://www.fao.org/3/I9286EN/i9286en.pdf. Acesso em: 7 de agosto de 2018.

FARIA, C. U. de; MAGNABOSCO, C. U.; DE LOS REYES, A.; LÔBO, R. B.; BEZERRA, L. A. F. Inferência bayesiana e sua aplicação na avaliação genética de bovinos da raça nelore: revisão bibliográfica. **Ciência Animal Brasileira** , v. 8, n. 1, p. 75-86, 2007.

FERNANDO, R.L.; GARRICK, D. Bayesian methods applied to GWAS. **Methods in Molecular Biology**, v. 1019, p. 237–274, 2013.

FIGUEIREDO FILHO, L. A. S. et al. Genetic parameters for carcass traits and body size in sheep for meat production. **Tropical Animal Health and Production**, v. 48, p. 215–218, 2016.

FIGUEIREDO FILHO, L. A. S.; et al. Estimate of genetic parameters for carcass traits and visual scores in meat sheep using Bayesian inference via threshold and linear models. **Ciência Rural**, v.47, n.3, 2017.

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

GIANOLA, D.; RODRIGUEZ-ZAS, S.; SHOOK, G. E. The Gibbs sampler in the animal model: a primer. In: FOULLEY, J. L.; MOLENAT, H. (Ed.). **SÉMINAIRE MODELE ANIMAL**. INRA Departament de Genetique Animale, La Colle sur Loup, France, p. 47-56, 1994.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. **Genetics**, v.183, n.1, p.347–363, 2009.

GOLDIE, C.M., KLÜPPELBERG, C. Subexponential Distributions. A Practical Guide to Heavy Tails: **Statistical Techniques and Applications**. Birkhauser Boston, Cambridge, 435-459, 1998.

GORDO, D. G. M. et al. Genetic parameter estimates for carcass traits and visual scores including or not genomic information. **Journal of Animal Science**, v. 94, p. 1821–1826, 2016.

GRASSO, A.N.; GOLDBERG, V.; NAVAJAS, E.A.; IRIARTE, W.; GIMENO, D.; AGUILAR, I.; MEDRANO, J.F.; RINCÓN, G.; CIAPPESONI, G. Genomic variation and population structure detected by single nucleotide polymorphism arrays in Corriedale, Merino and Creole sheep. **Genetics and Molecular Biology**, v.37, p.389-395, 2014. DOI: 10.1590/S1415-47572014000300011

HABIER D. et al. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, v. 12, n. 186, 2011.

HADFIELD, J.D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: **The MCMCglmm R Package Journal of Statistical Software,** v.33, p.1-22, 2010.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Pesquisa agropecuária municipal, 2017. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/84/ppm_2017_v45_br_informativo.pdf. Acesso em: 08 de outubro de 2018.

JONAS, E.; DE KONING, D. J. Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. **Frontiers in genetics**, v. 6, n.49, 2015.

LEE, S.H.; VAN DER WERF, J. H. J.; HAYES, B. J.; GODDARD, M. E.; VISSHER, P. M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS Genetics**, v. 4, 2008.

LEGARRA, A. et al. Improved Lasso for genomic selection. **Genetics Research**, v.93, n.1, p.77-87, 2011.

LI, J; DAS, K; FU, G.; LI, R.; WU, R. The Bayesian lasso for genome-wide association studies. **Bioinformatics**. v. 27, p. 516–523, 2011.

LI, H. SU, G.; JIANG, L; BAO, Z. An efficient unified model for genome‑wide association studies and genomic selection. **Genetics Selection Evolution**, v. 49, n. 64, 2017.

Lindley, D. V. The choice of variables in multiple regression, **Journal of the Royal Statistical Society**. Series B, 30, 1, 31, 1968.

LOBO, R.N.B. Opportunities for investiment into small ruminant breeding programmes in Brazil. **Journal of Animal Breeding and Genetics**, v. 136, n.5, p. 313-318, 2019.

LUCENA, C. C.; GUIMARÃES, V. P. **Boletim do Centro de Inteligência e Mercado de Caprinos e Ovinos**. Embrapa Caprinos e Ovinos, Sobral, n. 3, 2018. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/185645/1/BoletimCIM3-2018.pdf. Acesso em: 8 de outubro de 2018.

MATIKA, O. et al. Genome‑wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. **Genetics Selection Evolution**, v. 48, n. 11, 2016.

MCMANUS, C.; PAIVA, S. R.; ARAÚJO, R. O. Genetics and breeding of sheep in Brazil. **Revista Brasileira de Zootecnia**, v. 39, p.236-246, 2010.

MCMANUS, C. LOUVANDINI, H.; PAIM, T. P.; MARTINS, R. S.; BARCELLOS, J. O. J.; CARDOSO, C.; GUIMARÃES, R. F.; SANTANA, O. A. The challenge of sheep farming in the tropics: aspects related to heat tolerance. **Revista Brasileira de Zootecnia**, v.40, p.107-120, 2011.

MEHRBAN, H. et al. Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. **Genetics Selection Evolution**, v. 49, n. 1, 2017.

MEIRA, A. N. et al. Single nucleotide polymorphisms in the growth hormone and IGF type-1 (IGF1) genes associated with carcass traits in Santa Ines sheep. **Animal**, v.6, p. 1-9, 2018.

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

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, v. 6, n. 1, 2016.

MEYER, P. L. **Probabilidade: Aplicações à Estatística**. Rio de Janeiro: Livros Técnicos e Científicos, Editora S.A., 2006

MONTGOMERY, D. C.; RUNGER, G. C.; HUBELE, N. F. **Estatística Aplicada à engenharia**. Rio de Janeiro, 2ª ed., LTC 2011.

MORAIS, O. R. Melhoramento Genético dos Ovinos no Brasil: situação e perspectivas. In: **III Simpósio Nacional de Melhoramento Animal**, Belo Horizonte... Anais, Belo horizonte, p.266-271, 2000.

NEYMAN, J., SCOTT, E.T. Outliers Proneness of Phenomena and Related Distributions. **Optimizing Methods in Statistics**. Academic Press, New York, p. 413-430, 1971.

O'Hagan, A. On outlier rejection phenomena in Bayes inference, **Journal of the Royal Statistical Society**. Series B, 41, 3, 358-367, 1979.

O'Hagan, A. Modelling with Heavy Tails, **Bayesian Statistics**, 3, 345-359, 1988.

O'Hagan, A. Outliers and Credence for Location Parameter Inference, **Journal of the American Statistical Association**, v. 85, p. 172-176, 1990.

O'HAGAN, A.; LE, H. **Conflicting information and a class of bivariate heavy-tailed distributions**. In Aspects of Uncertainty: a Tribute to D. V. Lindley, (eds. P. R. Freeman and A. F. M. Smith), New York: Wiley, 311-327, 1994.

O'HAGAN, A.; PERICCHI, L. Bayesian heavy-tailed models and conflict resolution: a review, **Brazilian Journal of Probability and Statistics**, v. 26, p. 372-401, 2012.

PAIVA, S. R. SILVÉRIO, V. C.; PAIVA, D. A. F.; MCMANUS, C.; EGITO, A. A.; MARIANTE, A. S.; CASTRO, S. R.; ALBUQUERQUE, M. S. M.; DERGAN, J. A. Origin of the main locally adapted sheep breeds of Brazil: a rflp-pcr molecular analysis. **Archivos de Zootecnia**, v. 54, p. 395-399, 2005

RAKITSCH, B.; LIPPERT, C.; STEGLE, O.; BORGWARDT, K. A lasso multi-marker mixed model for association mapping with population structure correction. **Bioinformatics**. v. 29, p.206–214, 2013.

REGO NETO, A. **Estrutura genética e associação genômica ampla para características de tamanho corporal em ovinos da raça Santa Inês**. Teresina: Centro de Ciências Agrárias, Universidade Federal do Piauí, 89 p. Tese de Doutorado, 2017.

REGO NETO, A.; SARMENTO, J. L. R.; SANTOS, N. P. S.; BIAGIOTTI, D.; SANTOS, G. V.; CAMPELO, J. E. G.; SENA, L. S.; FIGUEIREDO FILHO, L. A. S. Population genetic structure of Santa Inês sheep in Brazil. **Tropical Animal Health and Production**, v. 50, n.3, p. 503-508, 2018.

ROVADOSCKI, G. A. et al. Estimates of genomic heritability and genome-wide association study for fatty acids profile in Santa Inês sheep. **BMC Genomics**, v. 19, n. 375, 2018.

ROWE, J.B. The Australian sheep industry - undergoing transformation. **Animal Production Science**, v. 50, p. 991–997, 2010.

RUPP, R. MUCHA, S., LARROQUE, H., MCEWAN, J., CONINGTON, J. Genomic application in sheep and goat breeding. **Animal Frontiers**, v.6, n.1, 2016.

SANTOS, G. V. **Estudo genômico aplicado ao melhoramento genético de ovinos tropicais para resistência à endoparasitas**. Teresina: Centro de Ciências Agrárias, Universidade Federal do Piauí, 90 p. Tese de Doutorado, 2018.

SARMENTO, J.L.R. et al. Avaliação genética de características de crescimento de ovinos Santa Inês utilizando modelos de regressão aleatória. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 58, n. 1. p. 68-77, 2006.

SARMENTO, J.L.R. et al. Modelos de regressão aleatória na avaliação genética do crescimento de ovinos da raça Santa Inês. **Revista Brasileira de Zootecnia**, v. 39, n. 8, p. 1723-1732, 2010.

SEGURA, V.; VILHJÁMSSON, B. J.; PLATT, A.; KORTE, A.; SEREN, U.; LONG, Q.; NORDBORG, M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. **Nature Genetics**, v. 44, p. 825–830, 2012.

SENA, L. S. et al. Genetic parameters for carcass traits and body size of meat sheep. **Semina: Ciências Agrárias**, v.37, n. 4, p. 2477-2486, 2016.

SHARMA, A. Stories and Challenges of Genome Wide Association Studies in Livestock — A Review. **Asian Australasian Journal of Animal Science**, v. 28, n. 10, p. 1371-1379, 2015.

SILVA, R. M. O.; STAFUZZA, N. B.; FRAGOMENI, B. O.; CAMARGO, G. M. F.; CEACERO, T. M.; CYRILLO, J. N. S. G.; BALDI, F.; ALBUQUERQUE, L. G. Genome-Wide Association study for carcass traits in an experimental Nelore catle population. PLoS ONE 12(1): e0169860.doi:10.1371/journal. pone.0169860, 2017.

SOUSA, W.H. et al. Estimativas de componentes de (co) variância e herdabilidade direta e materna de pesos corporais em ovinos da raça Santa Inês. **Revista Brasileira de Zootecnia**, v. 28, n. 6, p. 1252-1262, 1999.

SOUSA, W.H.; LÔBO, R.N.B.; MORAIS, O.R. Ovinos Santa Inês: estado de arte e perspectivas. In: SIMPÓSIO INTERNACIONAL SOBRE CAPRINOS E OVINOS DE

CORTE, 2., 2003, João Pessoa. **Anais...** João Pessoa: Empresa Estadual de Pesquisa Agropecuária da Paraíba, 2003. (CD-ROM)

SOUZA, S. F. et al. Aplicação da ultrassonografia para avaliação de condição corporal e acabamento de carcaça em pequenos ruminantes. **Ciência Veterinária nos Trópicos**, v. 19, n. 3, p. 34-42, 2016.

TEIXEIRA NETO, M.R.; CRUZ, J.F.; CARNEIRO, P.L.S.; MALHADO, C.H.M.; SOUZA, L.E.B.; FERRAZ, R.C.N. Evolução da biometria corporal de ovinos da raça Santa Inês. **Agropecuária Científica no Semiárido**, v.12, n.2, p.170-180, 2016. Disponível em: http://revistas.ufcg.edu.br/acsa/index.php/ACSA/article/view/737/pdf.

TEUGELS, J.L. The class of subexponential distributions. **The Annals of Probability**, v. 3 (6), p. 1000-1011, 1975.

VAN DEN BERG, I.; FRITZ, S.; BOICHARD, D. QTL fine mapping with Bayes C(pi): a simulation study. **Genetics Selection Evolution**, v. 45, n.19, 2013.

WU, Y.; FAN, H.; WAN, Y.; ZHANG, L.; GAO, X.; CHEN, Y.; LI, J.; REN, H. Y.; GAO, H. Genome-wide association studies using haplotypes and individual SNPs in Simmental cattle. **Plos One**, v.9, n.10, 2014.

ZHANG, H. Progress of genome wide association study in domestic animals. **Journal of Animal Science and Biotechnology**, v. 3, n. 26, 2012.

CAPÍTULO 1

Bayes t e Bayes DE: modelos bayesianos robustos para Seleção Genômica Ampla

Bayes t e Bayes DE: modelos bayesianos robustos para Seleção Genômica Ampla

Running title: Modelos robustos de Seleção Gênomica Ampla

Max Brandão de Oliveira[(1)] e José Lindenberg Rocha Sarmento[(2)]

[(1)]Pós Graduando do Programa de Pós-Graduação em Ciência Animal – UFPI. e-mail: maxbrandao@gmail.com

[(2)]Professor da UFPI/DZO e Bolsista do CNPq, e-mail: sarmento@ufpi.edu.br

Resumo: O objetivo com esta pesquisa foi desenvolver modelos de Seleção Genômica Ampla adotando as distribuições t-Student e dupla-exponencial para a variável resposta, Bayes t e Bayes DE, respectivamente. Para isso, foi desenvolvido um modelo Bayesiano no software R e foram utilizados diferentes critérios (por exemplo, DIC, análise residual e acurácia) para quantificar a diferença entre os ajustes dos modelos propostos com o RRBLUP, já consolidado e descrito na literatura. Os modelos foram aplicados a dados simulados gerados pelo software QMSim com diferentes níveis de dispersão para variável resposta (5, 10 e 15) e com diferentes tamanhos amostrais (300, 1.000, 2.000 e 4.000). Os resultados apontaram que, quando existem *outliers* na amostra, os modelos Bayes t e Bayes DE são mais robustos que o RRBLUP. Os três modelos avaliados foram centrados em 0, mas a dispersão residual dos modelos propostos foi melhor. Além disso, em todos os cenários testados, os modelos Bayes t e Bayes DE foram mais acurados em relação ao RRBLUP. O modelo Bayes t foi o mais acurado e apresentou menor variabilidade residual do que o Bayes DE, principalmente para tamanhos amostrais menores. Desta forma, os modelos propostos mostraram-se mais atrativos para dados de maior dispersão e menor tamanho amostral, conferindo mais robustez em relação ao modelo RRBLUP.

Palavras-chave: Avaliação genética genômica, Distribuição t-Student, outliers, predição genômica, robustez, SNPs.

1 | INTRODUÇÃO

Os processos de sequenciamento e genotipagem permitem acrescentar a informação genômica no melhoramento genético animal. Os dados genômicos são incorporados nos modelos de Seleção Genômica Ampla (GWS – *Genome Wide Selection*), propostos por Meuwissen, Hayes e Goddard (2001), que utilizam marcadores do tipo SNP (*single nucleotide polymorphism*), com larga abrangência em espécies cujo genoma esteja sequenciado.

Outros métodos de GWS foram propostos na mesma linha, como por exemplo, De Los Campos et al. (2009), Habier, Fernando, Kizilkaya e Garrick (2011), Fernando e Garrick (2013), Van Den Berg, Fritz e Boichard (2013), Misztal, Legarra e Aguilar (2014) e Bowless (2015). A predição na GWS é simultânea e direta, permitindo a estimação do

mérito genômico dos animais e do efeito de cada SNP em função da característica de interesse (Gianola et al., 2009).

Embora apresentem várias vantagens, os métodos de GWS são complexos em termos computacionais e estatísticos, e são rígidos quanto à distribuição associada à variável resposta (gaussiana ou ordinal) (Resende, Lopes, Silva, & Pires, 2008; Meuwissen, Hayes, & Goddard, 2016). A adoção incorreta de uma distribuição ou de um modelo pode levar a conclusões equivocadas, uma vez que pode afetar fortemente a distribuição *a posteriori* (ANDRADE; OMEY; AQUINO, 2017).

Além da escolha da distribuição adequada, os modelos estatísticos são sensíveis a conflitos de informação, como *outliers*, que prejudicam o ajuste(Andrade & O'Hagan, 2011). Esse fenômeno foi identificado e a distribuição t de Student foi sugerida como solução, já que ela tem caudas pesadas e, portanto, poderia conferir mais robustez ao modelo (De Finetti, 1961; Lindley, 1968). Além da distribuição t-Student, a dupla-exponencial tem um comportamento análogo quanto à simetria e às caudas pesadas, mas tem um pico mais elevado e concentrado em torno da média. A adoção destas distribuições para variável resposta $y_i$dos modelos de GWS pode conferir mais robustez aos ajustes, como menor sensibilidade à presença de *outliers* ou a pequenos tamanhos amostrais.

Portanto, com esta pesquisa objetivou-se propor modelos bayesianos robustos de Seleção Genômica Ampla que utilizem a distribuição t-Student e a dupla-exponencial para variável respostay_i e produzam estimativas mais consistentes.

2 | MATERIAL E MÉTODOS

2.1 | Dados simulados

Foi utilizado o *software* livre QMSim para simular os dados genômicos (Sargolzaei & Schenkel, 2009). Este *software* desenvolvido para simular arquiteturas genéticas e estruturas populacionais na pecuária. Dados genômicos em larga escala e pedigrees complexos podem ser simulados eficientemente. Além disso, também pode levar em conta características evolutivas predefinidas, como desequilíbrio de ligação, mutação, gargalos genéticos e expansões (Hill & Robertson, 1968).

A simulação é realizada em duas etapas. Na primeira, uma população histórica é simulada para estabelecer o equilíbrio mutação-deriva e, na segunda, são geradas estruturas populacionais recentes. O *software* QMSim permite incorporar diversos parâmetros nos modelos de simulação para produzir dados apropriados bem próximos a um cenário real, com a vantagem de produzir o valor genético verdadeiro para os animais.

A estrutura dos dados foi gerada a partir da simulação de 1000 gerações de uma população inicialmente composta por 200 machos e 1800 fêmeas com prolificidade média igual a duas progênies. Foi adotada equiprobabilidade entre os sexos (0,5 para cada). Na última geração, foi sugerida uma quantidade de 200 machos para favorecer um tamanho efetivo coerente com o propósito de evitar endogamia.

Foi utilizada uma herdabilidade de 0,3 e três níveis de variância fenotípica: 5, 10 e 15. Tais quantidades foram usadas para avaliar o comportamento e a sensibilidade do modelo proposto em relação a presença de mais *outliers* na amostra, já que, quanto maior a variância, mais valores discrepantes se fazem presentes na amostra. Além disso, foram utilizados 4 tamanhos amostrais nas simulações, $n = 300, 1.000, 2.000$ e 4.000, com o propósito de avaliar como os modelos se comportam para diferentes valores de n.

Em termos genômicos, foram considerados 10 cromossomos com 400 cM cada. Entre cada cM havia um marcador sujeito a ocorrência de mutação. A taxa de mutação adotada foi de $m = 2.5 \times 10^{-7}$ por *locus* e por geração, sendo aleatória. Foram utilizados 4000 marcadores moleculares distribuídos de forma aleatória em 10 QTLs (*Quantitative Trait Loci*) por cromossomo, de modo que a frequência alélica por QTL é aleatória.

A taxa de marcadores genotipados perdidos foi de 1.0×10^{-5}. Quanto ao erro na genotipagem, a taxa de erro foi definida como 5.0×10^{-5}. A taxa de mutação dos marcadores na população histórica foi considerada como 2.5×10^{-3}, enquanto a taxa de mutação da QTL na população histórica foi de 2.5×10^{-5}, sendo estes dois últimos padrões do QMSim. Os marcadores e os QTLs foram aleatorizados no genoma.

2.4 | Bayes t e Bayes DE

Os modelos propostos são puramente bayesianos e utilizam distribuições a *priori* equivalentes aos modelos já propostos, isto é, normal para os efeitos sistemáticos e para o mérito genômico dos animais, qui-quadrado invertida para os componentes de variância e para os efeitos dos marcadores, serão usadas duas distribuições, a normal e a t de Student.

Os ajustes dos modelos foram implementados com a inclusão de dados fenotípicos e genotípicos simulados. Desta forma, assume-se modelo aditivo geral dado por (1)

$$y_i = \mu + \boldsymbol{x}'_{\boldsymbol{f}_i}\boldsymbol{\beta_f} + \boldsymbol{x}'_{\boldsymbol{m}_j}\boldsymbol{\beta_m} + \varepsilon_i \; \forall i = 1,2,\dots,n\,; j = 1,2,\dots,m, \quad (1)$$

em que μ é o intercepto, $\boldsymbol{\beta_f}$ é o vetor de parâmetros associado aos efeitos sistemáticos; $\boldsymbol{\beta_m}$ é o vetor associado ao efeitos dos marcadores genéticos; $\boldsymbol{x}'_{\boldsymbol{f}_i}$ e $\boldsymbol{x}'_{\boldsymbol{m}_j}$, respectivamente, são as matrizes de incidência dos efeitos sistemáticos e aleatórios dos SNPs; e ε_i é o resíduo do modelo, considerado independente e identicamente distribuído. A função de verossimilhança, com base no modelo tradicional com distribuição normal, é escrita como (2)

$$p(y \mid \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} N(y_i \mid \mu + \boldsymbol{x}'_{\boldsymbol{f}_i}\boldsymbol{\beta_f} + \boldsymbol{x}'_{\boldsymbol{m}_i}\boldsymbol{\beta_m}, \sigma_\varepsilon^2). \quad (2)$$

Os modelos inicialmente propostos sugerem a distribuição normal para y_i e para as *prioris* dos efeitos sistemáticos. Para os componentes de variância, é adotada a distribuição qui-quadrado invertida. Já para os marcadores, são utilizadas a normal, a t de Student e a dupla exponencial, sendo essa última tratada como uma mistura de distribuições (normal e exponencial). Portanto, as distribuições a *priori* são especificadas como (3)

$$\begin{aligned} p(\mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_\varepsilon^2, \sigma_m^2) &= N(\mu|0, \sigma_\mu)\, N(\boldsymbol{\beta_f} \mid \mathbf{0}, \boldsymbol{I}\sigma_f^2) \\ \times\, \chi^{-2}(\sigma_\varepsilon^2 \mid S_\varepsilon, d.f._{\varepsilon})\, &\chi^{-2}(\sigma_m^2 \mid S_m, d.f._{m}) \\ \times\, p(\boldsymbol{\beta_m}|\mathbf{0}, \boldsymbol{I}\sigma_m^2)\; &\forall\, i = 1,2,\dots,n\,; j = 1,2,\dots,p, \end{aligned} \quad (3)$$

onde $p(\boldsymbol{\beta_m}|\mathbf{0}, \boldsymbol{I}\sigma_m^2)$ tem distribuição $N(\mu, \sigma^2), t_k$ ou $DE(\mu, \sigma)$; σ_μ^2 e σ_m^2 são as variâncias de μ e β_{mj}, respectivamente; $d.f.$ e S são o grau de liberdade e o parâmetro de escala correspondentes às distribuições χ^{-2}; o índice i se refere ao número de animais; e j à quantidade de marcadores (Meuwissen, Hayes, & Goddard, 2001; De Los Campos et al., 2009; Gianola et al., 2009; Habier et al., 2007, 2011; Fernando & Garrick, 2013; Misztal, Legarra, & Aguilar,2014; Wang et al., 2014).

Portanto, pode-se representar hierarquicamente o modelo inicial por

$$y_i|\, \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_m^2, \sigma_e^2 \sim N(\mu + \boldsymbol{x}'_{\boldsymbol{f}_i}\boldsymbol{\beta_f} + \boldsymbol{x}'_{\boldsymbol{m}_i}\boldsymbol{\beta_m}, \sigma_\varepsilon^2)$$

$\mu \sim N(0, \sigma_\mu^2)$: Intercepto

$\boldsymbol{\beta_f} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_f^2)$: Efeitos sistemáticos do modelo

$\boldsymbol{\beta_m} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_m^2)\; \forall\, j = 1,2,3,\dots,m$: Efeito do marcador genético

$\sigma_\varepsilon^2 \sim \chi^{-2}(\sigma_\varepsilon^2 \mid S_\varepsilon\,, df_\varepsilon)$: Variância residual

$\sigma_m^2 \sim \chi^{-2}(\sigma_m^2 \mid S = S_m\,, df = df_m)$: Variância associada a $\boldsymbol{\beta_m}$

Desta forma, a proposta consiste em ajustar modelos cuja variável resposta do modelo (y_i) tem distribuição t de Student e dupla-exponencial (Laplace), com densidades (4) e (5), respectivamente.

$$f(x|v,\mu,\sigma) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left[\frac{v+\left(\frac{x-\mu}{\sigma}\right)^2}{v}\right]^{-\left(\frac{v+1}{2}\right)}, x,\mu \in \mathbb{R}, \sigma \geq 0, v \in \mathbb{N} \quad (4)$$

$$f(x|\mu,\sigma) = \frac{1}{2\sigma} exp\left(\left|\frac{x-\mu}{\sigma}\right|\right), x,\mu \in \mathbb{R}, \sigma > 0 \quad (5)$$

A distribuição t pode ser obtida em função da razão entre uma distribuição normal e uma qui-quadrado através da transformação

$$X \sim t_{(v)} = \frac{Z}{\sqrt{V/v}},$$

em que $Z \sim N(0,1)$ e $V \sim \chi^2_{(v)}$ (Chen & Huang, 2014). No entanto, apontam que a distribuição $t_{(.)}$ (4), com baixos graus de liberdade, tem mais robustez, principalmente em comparação a distribuição normal (Andrade & O'hagan, 2006; Andrade & Omey, 2013, 2016). Além disso, estudos apontam que os conflitos de informação prejudicam as inferências dos parâmetros de localização (Barnett & Lewis, 1978; Desgagné & Angers, 2007; Balakrishnan & Ristić, 2016).

As distribuições adotadas foram a $t_{(v)}$ com v graus de liberdade e a Laplace com média μ e desvio-padrão σ para a variável resposta y_i do modelo, que continua a ser especificado da mesma forma de (1), mas as verossimilhanças passam a ser escritas como (6) e (7)

$$p\left(y \mid \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_\varepsilon^2\right) = \prod_{i=1}^{n} t\left(y_i \mid \mu + \boldsymbol{x'_{f_i}\beta_f} + \boldsymbol{x'_{m_i}\beta_m}, \sigma_\varepsilon^2, d.f.\right) \quad (6)$$

$$p\left(y \mid \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_\varepsilon^2\right) = \prod_{i=1}^{n} DE\left(y_i \mid \mu + \boldsymbol{x'_{f_i}\beta_f} + \boldsymbol{x'_{m_i}\beta_m}, \sigma_\varepsilon^2\right). \quad (7)$$

Mantendo a mesma configuração com relação às distribuições a *priori*, tem-se a mesma estrutura de (3). Portanto, pode-se representar hierarquicamente os modelos propostos por

$$y_i|\, \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_m^2, \sigma_e^2 \sim t_v(\mu + \boldsymbol{x'_{f_i}\beta_f} + \boldsymbol{x'_{m_i}\beta_m}, \sigma_\varepsilon^2)$$

$$y_i|\, \mu, \boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_m^2, \sigma_e^2 \sim DE(\mu + \boldsymbol{x'_{f_i}\beta_f} + \boldsymbol{x'_{m_i}\beta_m}, \sigma_\varepsilon^2)$$

$\mu \sim N\left(0, \sigma_\mu^2\right)$: Intercepto;

$\boldsymbol{\beta_f} \sim N\left(\mathbf{0}, \boldsymbol{I}\sigma_f^2\right)$: Efeitos sistemáticos do modelo;

$\boldsymbol{\beta_m} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_m^2)\ \forall\ j = 1,2,3, \ldots, p$: Efeito genético do marcador;

$\sigma_{\varepsilon}^2 \sim \chi^{-2}(\sigma_{\varepsilon}^2 \mid S_{\varepsilon}, df_{\varepsilon})$: Variância residual;

$\sigma_m^2 \sim \chi^{-2}(\sigma_m^2 \mid S = S_m, df = df_m)$: Variância associada a $\boldsymbol{\beta_m}$.

Utilizando as funções de verossimilhanças em (6) e (7), utiliza-se a condicional completa

$$p(\boldsymbol{\beta_f}, \boldsymbol{\beta_m}, \sigma_m^2, \sigma_{\varepsilon}^2) = p(\boldsymbol{\beta_f}|\sigma_{\varepsilon}^2)p(\sigma_{\varepsilon}^2)p(\boldsymbol{\beta_m}|\sigma_m^2)p(\sigma_m^2)$$

$$= \prod_{i=1}^{n} N(\beta_{f_i}, \sigma_{\varepsilon}^2) \times \chi^{-2}(\sigma_{\varepsilon}^2|d.f., S) \times$$

$$\prod_{j=1}^{p} N\left(\beta_{m_j}, \sigma_m^2\right) \times \chi^{-2}(\sigma_m^2|d.f., S_m), \qquad (8)$$

em ambos os casos. Estes modelos são denominados de Bayes t e Bayes DE, sugeridos de acordo com as distribuições adotadas. Na Figura 1 são apresentadas as 3 distribuições mencionadas, a t de Student com 3 graus de liberdade, a normal padrão e a dupla exponencial padrão. Percebe-se que a $t_{(3)}$ apresenta caudas mais pesadas do que a distribuição $N(0,1)$, conferindo a mesma uma maior variabilidade e expressando uma ignorância acerca do parâmetro ou fenômeno de interesse. Por outro lado, a $DE(0,1)$ tem uma forte concentração em torno da média e caudas semelhantes a t.

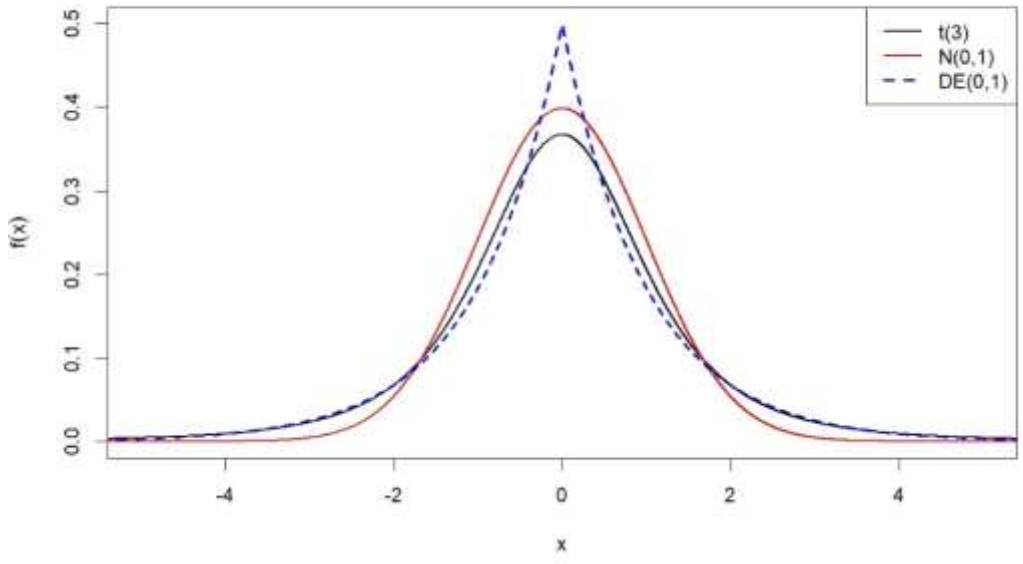

FIGURA 1 Densidades das distribuições $N(0,1), t_{(3)}$ e $DE(0,1)$ utilizadas para modelar a variável resposta y_i nos ajustes RRBLUP, Bayes t e Bayes DE

2.5 | Desenvolvimento dos modelos propostos

Para implementação dos ajustes foi utilizado o *software* R versão 2.5 (R Core Team, 2018). Para tanto, foi usado o pacote BGLR (Perez & De Los Campos, 2014) para aplicar o RRBLUP, método já consolidado e presente na rotina do BGLR (Meuwissen, Hayes, & Goddard, 2001). A função, com o mesmo nome do pacote, estimou o efeito do sexo como fixo e os efeitos dos marcadores como aleatórios, todos com o mesmo componente aleatória σ_m^2.

Para os ajustes propostos, foram utilizados os pacotes metRology, para a distribuição t-Student escalonada com v graus de liberdade, média μ e desvio-padrão σ; mvtnorm, para utilização da distribuição normal multivariada para a priori de $\boldsymbol{\beta_m}$, LaplacesDemon, para utilizar a distribuição dupla-exponencial com média μ e variância σ; e mcmc para implementação do método iterativo através do algotirmo Metropolis-Hastings.

A ideia partiu da elaboração de 3 funções, que representam a verossimilhança, a *priori* e a *posteriori*. A função de verossimilhança continha todos os parâmetros do modelo e associava a distribuição t ou dupla-exponencial à variável resposta y_i. A função a priori continha as distribuições associadas a cada parâmetro, isto é, qui-quadrado invertida para σ_ε^2 e σ_m^2, e normal para os demais efeitos sistemáticos e aleatórios. Por fim, a *posteriori* era definida em função da verossimilhança e da *priori.*

As iterações foram iniciadas com a função metrop do pacote mcmc, cujos argumentos são a função da *posteriori*, os vetores de inicialização dos parâmetros do modelo e valores correspondentes ao número de interações, de *burn in* e os vetores da variável resposta y_i e do sexo dos animais e a matriz dos marcadores SNPs, codificadas como 0, 1 e 2.

Foram utilizados os modelos Bayes t e Bayes DE em comparação ao RRBLUP, já que os 3 apresentam as mesmas *prioris* e diferenciam-se apenas na distribuição de y_i. Em termos de comparações, foram adotados os 3 níveis de variância fenotípica (5, 10 e 15) e 3 tamanhos amostrais (300, 1.000, 2.000 e 4.000) da última geração, com o propósito de avaliar o comportamento dos 3 modelos em situações distintas. Os *outliers* foram identificados e selecionados em cada cenário ($\sigma^2 = 5, 10$ e 15).

As cadeias tiveram um total de 1.500.000 de iterações com descarte de 100.000 e um intervalo de amostragem de 100. A comparação entre os modelos foi efetuada

principalmente pelo DIC (*Deviance Information Criterion*), pela acurácia (correlação de Pearson entre o valor genômico verdadeiro e o predito pelo modelo) e pela análise residual. Além disso, para confirmar a robustez, foram introduzidos aleatoriamente 10% de *outliers* nos tamanhos amostrais dos cenários testados com o propósito de avaliar se os critérios adotados ainda se apresentaram melhores (Andrade, Omey, & Aquino, 2017).

2 | RESULTADOS E DISCUSSÃO

Os diagramas de caixa dos três fenótipos utilizados como variável resposta dos modelos de acordo com suas variabilidades, destacando a presença de *outliers* na amostra são apresentados na Figura 2. A acurácia e o DIC dos três modelos (RRBLUP, Bayes t e Bayes DE) estão apresentados na Tabela 1. Observa-se que, mantendo o mesmo tamanho amostral observado dentro de cada ajuste, o DIC aumenta à medida que $\sigma^2_{fen_i}$ cresce, já a acurácia tende a diminuir de $\sigma^2_{fen} = 5$ para 10, e a aumentar de 10 para 15. Este fenômeno se repetiu nos três ajustes de forma geral.

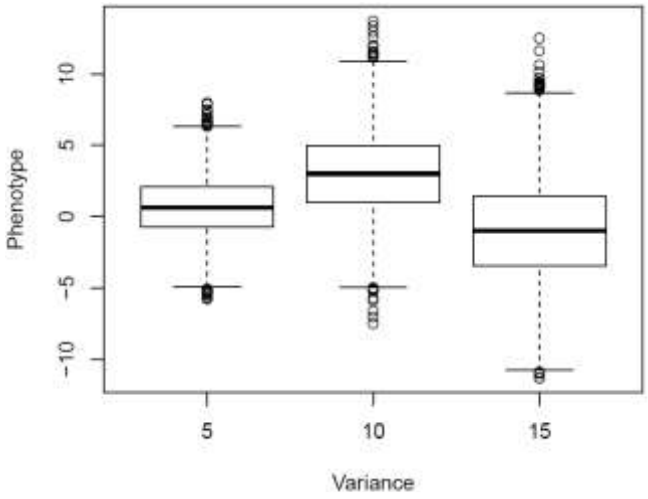

Figura 2 Diagramas de caixa para valores observados usados como variáveis respostas dos modelos com variâncias fenotípicas de 5, 10 e 15

Fixando o ajuste e a variância, o DIC diminui de acordo com a redução do tamanho amostral, o que faz sentido, visto que a informação amostral da verossimilhança diminui, mas a quantidade de parâmetros continua a mesma. Mantendo o tamanho amostral e σ^2_{fen} constantes e variando o ajuste (Tabela 1), percebe-se que o DIC aumenta do RRBLUP em direção ao Bayes t, de modo que a diferença entre o Bayes DE e o Bayes t é menor

em relação ao RRBLUP. Este comportamento pode ser justificado pela simulação dos dados, já que o mesmo obtém os resultados fenotípicos através de uma distribuição normal e isso poderia favorecer na obtenção da log-verossimilhança usada no cálculo do DIC.

Em contrapartida, a acurácia dos modelos propostos é superior ao RRBLUP, principalmente o Bayes t, que se mostrou mais acurado do que o Bayes DE, confirmando a hipótese testada. A acurácia do RRBLUP está de acordo com Clark et al. (2012) quando $n = 1.000$. Por outro lado, o Bayes t e o Bayes DE apresentaram acurácias menores que aquelas relatadas por Brito et al. (2017) em alguns cenários, porém estes autores aplicaram as técnicas de GWS em rebanhos organizados de acordo com métodos de agrupamento. O mesmo comportamento se repete em relação ao estudo de Habier, Fernando e Dekkers (2007), onde as acurácias dos modelos Bayes t e Bayes DE foram superiores.

TABELA 1 DIC e acurácia dos modelos (RRBLUP, Bayes t e Bayes DE) para tamanhos diferentes tamanhos amostrais (n) e variâncias fenotípicas (σ^2_{fen})

Parâmetro	DIC			Acurácia		
Variância	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$
RRBLUP						
$n = 300$	1265,017	1429,183	1570,522	0,1286	0,2413	0,0014
$n = 1000$	4165,727	4899,451	5253,897	0,3759	0,2607	0,4409
$n = 2000$	8603,218	9999,669	10846,960	0,4174	0,4262	0,4454
$n = 4000$	17510,350	19895,350	20274,900	0,4773	0,4443	0,4729
Bayes t						
$n = 300$	1306,329	1468,185	1626,677	0,2069	0,3024	0,0076
$n = 1000$	4288,222	5021,515	5376,565	0,3935	0,3625	0,4969
$n = 2000$	8840,680	10237,360	11071,160	0,4315	0,5306	0,5154
$n = 4000$	17991,850	20350,590	20274,130	0,5197	0,5866	0,5328
Bayes DE						
$n = 300$	1298,422	1462,750	1618,385	0,1307	0,2556	0,0029
$n = 1000$	4303,296	5028,455	5393,182	0,3850	0,2714	0,4617
$n = 2000$	8874,890	10257,130	11085,680	0,4297	0,4677	0,4652
$n = 4000$	18034,870	20380,370	20795,100	0,4886	0,4957	0,4934

Note. DIC: Deviance Information Criterion, RRBLUP: Ridge Regression Best Linear Unbiased Prediction

Destaca-se que, embora os dados fenotípicos tenham sido simulados a partir de uma distribuição normal, os ajustes com distribuições diferentes promoveram melhores resultados em termos de acurácia, o que é uma propriedade desejada na predição dos valores genômicos. Além disso, com o aumento da amostra, a acurácia aumentou, levando a crer que os resultados seriam ainda melhores em rebanhos maiores. Para avaliar a robustez, foram introduzidos aleatoriamente *outliers* em 10% dados fenotípicos nos diversos cenários seguindo distribuição normal com média igual a 20 e desvio-padrão igual a 1.

A introdução dos *outliers* naturalmente prejudica a análise em termos de acurácia, como observado na Tabela 2. Os valores do DIC e da acurácia, como esperado, foram influenciados de forma prejudicial, aumentando o DIC e reduzindo a acurácia. No entanto, apesar disto, os modelos Bayes t e o Bayes DE apresentaram comportamento análogo ao apresentado na Tabela 1, nas devidas proporções, os quais se mostraram melhores que o RRBLUP em todos os cenários. Destaca-se ainda que, à medida que o tamanho amostral aumenta, os métodos propostos se distanciam e apresentam melhores resultados.

TABELA 2 DIC e acurácia dos modelos (RRBLUP, Bayes t e Bayes DE) para diferentes tamanhos amostrais (n) e variância fenotípica (σ^2_{fen}) na presença de *outliers*

Parâmetro	DIC			Acurácia		
Variância	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$
RRBLUP						
$n = 300$	2212,081	2247,090	2269,747	0,0346	0,0161	0,0111
$n = 1000$	7342,035	7420,268	7337,984	0,0447	0,0823	0,0989
$n = 2000$	14650,250	14907,550	14729,940	0,0835	0,1095	0,1151
$n = 4000$	28842,100	28809,000	29548,190	0,1322	0,1663	0,1711
Bayes t						
$n = 300$	2211,177	2246,352	2273,126	0,0702	0,0555	0,0442
$n = 1000$	7342,186	7415,924	7313,869	0,0858	0,1120	0,1409
$n = 2000$	14650,010	14829,670	14697,120	0,1040	0,1926	0,2166
$n = 4000$	28842,130	28627,760	29353,100	0,1812	0,2662	0,2907

Bayes DE						
$n = 300$	2175,805	2218,218	2242,263	0,0502	0,0264	0,0195
$n = 1000$	7198,863	7324,893	7206,895	0,0581	0,1027	0,1161
$n = 2000$	14371,930	14653,870	14534,520	0,0923	0,1314	0,1464
$n = 4000$	28160,720	28276,600	29003,040	0,1521	0,2159	0,2204

Note. DIC: Deviance Information Criterion, RRBLUP: Ridge Regression Best Linear Unbiased Prediction

Observando as medidas residuais dos modelos (Tabela 3), todos os ajustes flutuaram em torno de 0, mas os modelos Bayes t e o Bayes DE apresentaram uma menor dispersão em todos os cenários comparativos. Isto é, os ajustes propostos têm mais consistência na predição dos estimadores, convergindo para um valor predito mais próximo do real.

Esse fenômeno justifica o uso, em especial, da distribuição t para pequenas amostras, pois essa propriedade confere maior variabilidade e expressa a ignorância acerca do parâmetro de interesse, tornando o modelo mais robusto, menos sensível à presença de *outliers* e ao tamanho da amostra (Haro-Lópes & Smith, 1999).

Fica evidente que o modelo Bayes t apresenta melhores resultados em relação aos modelos RRBLUP e Bayes DE principalmente quando a amostra é pequena (n = 300). Sabe-se que a distribuição t converge para a distribuição normal assintoticamente, mas para pequenas amostras, principalmente na presença de outliers, a distribuição t é melhor e confere mais robustez ao modelo. Assim, para amostras pequenas, recomenda-se a utilização da distribuição t.

TABELA 3 Média e erro-padrão dos modelos (RRBLUP, Bayes t e Bayes DE) para diferentes tamanhos amostrais (n) e variâncias fenotípicas (σ^2_{fen})

Variância	$\sigma^2_{fen} = 5$		$\sigma^2_{fen} = 10$		$\sigma^2_{fen} = 15$	
Estimativa	$\bar{X}$	EP	$\bar{X}$	EP	$\bar{X}$	EP
RRBLUP						
$n = 300$	-0,0022	1,8783	0,0005	2,4735	-0,0007	3,1406
$n = 1000$	0,0002	1,7631	0,0002	2,5815	0,0008	2,9798
$n = 2000$	0,0001	1,8858	0,0007	2,7041	0,0006	3,3577
$n = 4000$	0,0000	2,0053	-0,0004	2,6834	0,0008	2,7975
Bayes t						

$n = 300$	-0,0001	1,7614	0,0000	2,3307	0,0005	3,0043
$n = 1000$	0,0000	1,6545	0,0001	2,4315	0,0015	2,5767
$n = 2000$	0,0000	1,8059	0,0002	2,5877	-0,0009	3,0367
$n = 4000$	0,0000	1,9660	0,0000	2,6273	0,0006	2,5768
Bayes DE						
$n = 300$	0,0018	1,8349	-0,0004	2,4328	0,0038	3,1039
$n = 1000$	0,0001	1,7076	0,0006	2,5119	0,0017	2,8823
$n = 2000$	0,0001	1,8459	0,0000	2,6532	-0,0007	3,2952
$n = 4000$	0,0000	1,9854	-0,0001	2,6570	0,0002	2,7700

Note. RRBLUP: Ridge Regression Best Linear Unbiased Prediction

Na presença de *outliers* (Tabela 4), os resultados se mostraram melhores para pequenas amostras ($n = 300$), já que especialmente o modelo Bayes t apresentou uma diferença maior em relação aos ajustes RRBLUP e Bayes DE tanto em relação à média quanto ao desvio-padrão. Para as amostras superiores a 1.000, o comportamento foi semelhante aos observados na Tabela 3. O desvio-padrão foi menor no modelo Bayes t, seguido pelo Bayes DE e pelo RRBLUP, indicando um menor erro nas estimativas, contudo a diferença foi pequena.

TABELA 4 Média e erro-padrão dos modelos (RRBLUP, Bayes t e Bayes DE) para diferentes tamanhos amostrais (n), e variâncias fenotípicas (σ^2_{fen}) e com a inclusão de *outliers*

Variância	$\sigma^2_{fen} = 5$		$\sigma^2_{fen} = 10$		$\sigma^2_{fen} = 15$	
Estimativa	$\bar{X}$	EP	$\bar{X}$	EP	$\bar{X}$	EP
RRBLUP						
$n = 300$	0,0037	9,1289	-0,0024	9,6239	-0,0041	10,0684
$n = 1000$	0,0004	8,8827	-0,0012	9,1295	0,0004	8,8148
$n = 2000$	0,0000	8,7669	0,0008	9,4132	0,0010	9,0793
$n = 4000$	0,0002	8,4232	0,0009	8,4055	0,0001	9,2057
Bayes t						
$n = 300$	0,0012	9,0435	-0,0010	9,3393	-0,0019	9,8304
$n = 1000$	0,0002	8,8021	-0,0007	8,6456	-0,0003	8,3835
$n = 2000$	0,0001	8,5976	0,0003	9,1334	0,0005	8,8810
$n = 4000$	-0,0001	8,2243	-0,0005	8,2784	0,0000	9,0591

Bayes DE						
$n = 300$	0,0021	9,0134	0,0019	9,4721	0,0037	9,9896
$n = 1000$	-0,0002	8,8461	0,0010	8,9306	-0,0004	8,6345
$n = 2000$	0,0000	8,6272	-0,0006	9,2807	-0,0009	8,9945
$n = 4000$	0,0001	8,3584	0,0007	8,3447	-0,0000	9,1329

Note. RRBLUP: Ridge Regression Best Linear Unbiased Prediction

A correlação entre os valores preditos e observados nos 3 modelos aumenta à medida que o tamanho amostral cresce (Tabela 5). Mantendo tamanho amostral e os modelos constantes, observa-se que a correlação diminui quando a variância fenotípica aumenta, um comportamento esperado em função da maior dispersão do fenótipo. Comparando os modelos de acordo com os cenários, os Bayes t novamente apresentou melhores resultados em relação ao RRBLUP e ao Bayes DE.

TABELA 5 Correlações entre valores preditos e observados obtidos com uso dos modelos RRBLUP, Bayes t e Bayes DE para diferentes tamanhos amostrais (n) e variâncias fenotípicas (σ^2_{fen})

vVariância	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$
RRBLUP			
$n = 300$	0,3347	0,2885	0,2560
$n = 1000$	0,6053	0,5658	0,5116
$n = 2000$	0,6264	0,6066	0,5947
$n = 4000$	0,6794	0,6622	0,5869
Bayes t			
$n = 300$	0,4768	0,4539	0,3371
$n = 1000$	0,6682	0,6396	0,6069
$n = 2000$	0,6821	0,6635	0,6597
$n = 4000$	0,7374	0,7124	0,6874
Bayes DE			
$n = 300$	0,3943	0,2873	0,2732
$n = 1000$	0,6326	0,6054	0,6433
$n = 2000$	0,6445	0,6129	0,6107
$n = 4000$	0,6879	0,6964	0,5956

Note. RRBLUP: Ridge Regression Best Linear Unbiased Prediction

Os valores de correlação entre os valores preditos e o reais estão de acordo com Meuwissen et al. (2001), $0{,}732 \pm 0{,}030$, de modo que os 3 métodos apresentaram comportamento esperado. Isto é, para maiores valores de n, maior a consistência nas estimativas e menor a distância entre o estimado e o observado, bem como, quanto maior a dispersão, menor o valor da correlação, já que o modelo tem menor precisão quando a variabilidade dos dados é alta. Porém, ainda assim, o Bayes t mostrou-se melhor de forma geral.

Na presença de *outliers* (Tabela 6), o cenário foi equivalente aos demais. Ao se comparar os modelos, mais uma vez o Bayes t foi superior ao Bayes DE e ao RRBLUP, bem como, quando o tamanho amostral cresce, maior a correlação entre os valores preditos e os observados e, quanto maior a variância, menor a correlação.

Considerando o DIC, a diferença entre os ajustes é pequena, mas favorável ao uso do modelo RRBLUP em relação aos outros dois modelos avaliados, mesmo nos diferentes cenários. Quanto ao resíduo, os três modelos foram centrados em torno de 0, mas os métodos Bayes t e Bayes DE apresentaram dispersão inferior ao RRBLUP. Isso pode ser justificado devido às propriedades de caudas pesadas apresentadas pela distribuições t e dupla-exponencial (LE & O'HAGAN, 1998).

O resultado mais importante está presente na acurácia dos modelos propostos neste estudo, em comparação ao RRBLUP. Os dois modelos propostos mostraram-se mais acurados em relação ao RRBLUP, em todos os cenários avaliados, de modo que o modelo Bayes t foi superior ao Bayes DE. Esta diferença se destaca ainda mais quando a amostra é maior. Além disso, para uma amostra menor ($n = 300$), o modelo Bayes t ainda se mostra mais acurado e com menor dispersão residual.

Esse comportamento sugere que se deve adotar a distribuição t com baixos graus de liberdade em relação às distribuições normal e dupla-exponencial para a variável resposta, em modelos de Seleção Genômica. Além disso, como o processo de genotipagem ainda é considerado caro em países subdesenvolvidos ou emergentes, seria mais interessante utilizar modelos mais robustos para aplicação em amostras com tamanhos menores e com *outliers*.

TABELA 6 Correlações entre valores preditos e observados obtidos com uso dos modelos RRBLUP, Bayes t e Bayes DE com diferentes tamanhos amostrais (n), e variâncias fenotípicas (σ^2_{fen}) e com *outliers* na variável resposta

Variância	$\sigma^2_{fen} = 5$	$\sigma^2_{fen} = 10$	$\sigma^2_{fen} = 15$
RRBLUP			
$n = 300$	0,2878	0,3146	0,2640
$n = 1000$	0,5283	0,5641	0,5739
$n = 2000$	0,5858	0,5816	0,5420
$n = 4000$	0,5449	0,5478	0,5495
Bayes t			
$n = 300$	0,3437	0,3903	0,3361
$n = 1000$	0,5984	0,6232	0,6263
$n = 2000$	0,6854	0,6172	0,5723
$n = 4000$	0,6867	0,5659	0,5892
Bayes DE			
$n = 300$	0,3227	0,3587	0,2868
$n = 1000$	0,5509	0,5894	0,5965
$n = 2000$	0,6038	0,5988	0,5545
$n = 4000$	0,5959	0,5565	0,5596

Note. RRBLUP: Ridge Regression Best Linear Unbiased Prediction

4 | CONCLUSÃO

As distribuições t e dupla-exponencial promoveram menor sensibilidade à presença de *outliers* na amostra, o que conferiu maior robustez aos modelos bayesianos de seleção genômica propostos, Bayes t e Bayes DE. Com este estudo, há evidências que apontam uma diferença entre os ajustes propostos, de modo que os modelos Bayes t e Bayes DE são mais vantajosos do que o modelo RRBLUP. Dentre os modelos testados, o Bayes t apresentou-se melhor em todos os cenários avaliados, indicando benefícios significativos da utilização de distribuições mais robustas em modelos de seleção genômica ampla.

5 | REFERÊNCIAS

Andrade, J. A. A.; O'Hagan, A. (2006). Bayesian Robustness Modeling Using Regularly Varying Distributions. *Bayesian Analysis*, v. 1, p. 169-188

Andrade, J. A. A.; O'Hagan, A. (2011) Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, v. 38, p. 691-711

Andrade, J. A. A.; Omey, E. (2013). Modelling conflicting information using subexponential distributions and related classes, *Annals of the Institute of Statistics and Mathematics*, v. 65, p. 491-511

Andrade, J. A. A.; Omey, E. (2016). Resolution of conflict of information using O-regularly varying functions. *Statistica Neerlandica*

Andrade, J. A. A.; Omey, E.; Aquino, C. T. M. (2017). Bayesian robustness modelling using the floor distribution. *REVSTAT Statistical Journal*, v. 16, p. 1-17

Balakrishnan, N.; Ristić, M. M. (2016) Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal. *Journal of Multivariate Analysis*, v.143 , p.194–207

Barnett, V.; Lewis, T. (1978). *Outliers in statistical data*. Wiley: New York, 1978.

Bowless, D. (2015). Recent advances in understanding the genetic resources of sheep breeds locally-adapted to the UK uplands: opportunities they offer for sustainable productivity. *Frontiers in Genetics*, v.6, n.24

Brito, L. F., McEwan, J.C., Millera, S., Bain, W., Lee, M., Dodds, K., Newman, S.A., Pickering, N., Schenkel, F.S, & Clarke, S. (2017). Genetic parameters for various growth, carcass and meat quality traits in a New Zealand sheep population. *Small Ruminant Research*, n. 154, p. 81–91

Chen, S.; Huang, J. (2014). Rates of convergence of extreme for asymmetric normal distribution. *Statistics and Probability Letters*, v.84, P. 158–168

Clark, S. A.; Hickey, J. M.; Daetwyler, H. D.; Julius HJ; Van Der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, v. 44, p. 1-4

De finetti, B. (1961). The Bayesian Approach to the Rejection of Outliers. Proceedings of the 4th *Berkeley Symposium on Mathematical Statistics and Probability*. Volume I, 99-210. Berkeley, California: University of California Press

De Los Campos, G.; Naya, H.; Gianola, D.; Crossa, J. Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J. M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*, v. 182, p. 375–385

Desgagné, A. ; Angers, J. F. (2007). Conflicting information and location parameter inference. *Metron*, v. 65, p. 67–97

Fernando, R.L.; Garrick, D. (2013). Bayesian methods applied to GWAS. *Methods in Molecular Biology*, v. 1019, p. 237–274

Gianola, D.; De Los Campos, G.; Hill, W.G.; Manfredi, E.; Fernando, R.L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363

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

Habier D.; Fernando, R. L.; Kizilkaya, K.; Carrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, v. 12, n. 186

Haro-López, R. A.; Smith, A. F. M. (1999). On robust Bayesian analysis for location and scale parameters. *Journal of Multivariate Analysis*, v. 70, no. 1, p. 30–56

Hill, W.G.; Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet.*, v. 38, p.226-231

Le, H.; O'Hagan, A. (1998). A class of bivariate heavy-tailed distributions. *Sankhya***,** v. 60, p. 82–100

Lindley, D. V. (1968). The choice of variables in multiple regression. *Journal of the Royal Statistical Society*. Series B, v. 30

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

Meuwissen, T.; Hayes, B. J.; Goddard, M. E. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, v.6, n.1, p. 6-14

Misztal, I.; Legarra, A.; Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, v. 97, p. 3943–3952

Perez, P., De Los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics*, v. 198 (2), p. 483-495

R Core Team R. (2018). A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria

Resende, M. D. V.; Lopes, P. S.; Silva, R,L.; Pires, I.E. (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, v. 56, p. 63-78

Sargolzaei, M.; F. S. Schenkel. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, v. 25, p. 680-681. First published January 28

Van Den Berg, I.; Fritz, S.; Boichard, D. (2013). QTL fine mapping with Bayes C(pi): a simulation study. *Genetics Selection Evolution*, v. 45, n.19

Wang, Z.; Zhang, Z.; Yang, H; Wang, S.; Rong, E.; Pei, W.; Li, H.; Wang, N. (2014). Genome-wide association study for wool production traits in a Chinese Merino sheep population. *Plos One*, v.9, n.9

CAPÍTULO 2

Modelos bayesianos robustos para seleção genômica ampla de ovinos da raça Santa Inês para melhoramento de carcaça

Modelos bayesianos robustos para seleção genômica ampla de ovinos da raça Santa Inês para características de carcaça

Running title: Modelos robustos de Seleção Gênomica Ampla em ovinos da raça Santa Inês

Max Brandão de Oliveira[(1)] e José Lindenberg Rocha Sarmento[(2)]

[(1)]Pós-Graduando do Programa de Pós-Graduação em Ciência Animal – UFPI. e-mail: maxbrandao@gmail.com

[(2)]Professor da UFPI/DZO e Bolsista do CNPq, e-mail: sarmento@ufpi.edu.br

Resumo: O objetivo com esta pesquisa é aplicar e avaliar modelos de seleção genômica ampla usando as distribuições t (Bayes t) e Laplace (BayesDE) para a característica Área de Olho de Lombo (AOL) em ovinos da raça Santa Inês. Para isso, foram implementados modelos bayesianos de Seleção Genômica Ampla no *software* R e comparados ao modelo RRBLUP, já consolidado e descrito na literatura. Os modelos foram aplicados em uma amostra de 389 ovinos da raça Santa Inês, com a variável resposta AOL. Os resíduos dos três modelos testados se concentraram em torno de 0, mas o modelo Bayes t obteve o valor mais próximo de 0, bem como apresentou a menor dispersão residual e o menor valor de DIC. A correlação entre os valores preditos e observados no Bayes t foi 0,8459, enquanto nos modelos RRBLUP e Bayes DE, foram 0,7855 e 0,8054, respectivamente. Da mesma forma, a acurácia para o Bayes t foi 42,80%, enquanto para RRBLUP e Bayes DE foram 37,04% e 40,93%. Os indicadores apontaram Bayes t como um modelo superiora os demais. Ao selecionar 10% dos animais com maiores valores genômicos, a coincidência entre Bayes t e RRBLUP foi de 85,71%, indicando que, para fins de seleção, alguns animais seriam diferentes, possibilitando um equívoco e um prejuízo em termos de ganho genético. Portanto, os métodos propostos surgem como alternativas de modelos robustos para dados com presença de outliers e para tamanhos de amostras pequenos, com destaque para o Bayes t, que apresentou melhores resultados para a estrutura de dados testada.

Palavras-chave: características de carcaça, distribuição t, ovinos de corte, robustez, seleção genômica ampla

1 | INTRODUÇÃO

Os avanços nas pesquisas em genética molecular favoreceram o conhecimento do genoma de algumas espécies de interesse econômico, a partir dos processos de sequenciamento e genotipagem com uso de marcadores genéticos, em especial do tipo SNP (*single nucleotide polymorphism*). Essa informação é inserida nos chamados modelos de Seleção Genômica Ampla (GWS – *Genome Wide Selection*), empregados na estimação dos valores genômicos dos animais para fins de seleção. Os primeiros métodos de GWS surgiram com Meuwissen, Hayes & Goddard (2001), deu origem a vários outros (De Los

Campos et al., 2009; Gianola et al., 2009; Habier et al., 2011; Fernando & Garrick, 2013;Van Den Berg, Fritz, & Boichard, 2013; Misztal, Legarra, & Aguilar, 2014; Wang et al., 2014).

Uma das principais vantagens da GWS é que a seleção pode ocorrer de forma simultânea para uma grande quantidade de marcadores, permitindo a estimação do mérito genômico dos animais e o efeito de cada SNP em função da característica de interesse (Wang et al., 2014; Wu et al., 2014). Além disso, o processo de seleção se tornava direto, sendo possível estimar o efeito de genes específicos sob uma característica de interesse para fins de melhoramento (Resende et al., 2008; Bowless, 2015).

Devido à possibilidade que a GWS apresenta de agregar fontes de informação fenotípica e genotípica, esta técnica tem sido aplicada com eficiência para a predição do mérito genômico em animais de diferentes espécies (Meuwissen, Hayes, & Goddard, 2016). Em ovinos de diferentes raças já foram realizados estudos que comprovaram as vantagens da seleção genômica na avaliação genética de animais para diferentes características de carcaça e qualidade de carne (por exemplo, Clark, Hickey, Daetwyler & van der Werf, 2012; Daetwyler et al., 2010, 2012; Brito et al., 2017a, 2017b).

Os estudos relatados foram conduzidos em países desenvolvidos, como Austrália e Nova Zelândia. Nestes países, há maior disponibilidade de recursos financeiros e maior estruturação da cadeia produtiva da ovinocultura. Assim, a quantidade de animais com informações fenotípicas variou de aproximadamente 2.000 animais a mais de 14.000 indivíduos, dependendo da característica, com quantidades semelhantes de animais genotipados.

No Brasil, o cenário é bem diferente. Embora o efetivo do rebanho ovino do país seja de aproximadamente 14 milhões de cabeças (IBGE, 2017), a aplicação de seleção genômica em pesquisas com ovinos ainda é bem limitada, devido a limitações financeiras e organizacionais que dificultam a disponibilidade de maior número de informações. Desta forma, a quantidade de animais genotipados ainda é pequena, principalmente em função dos poucos recursos financeiros disponibilizados para a pesquisa nesta área (Lobo, 2019). Portanto, uma alternativa interessante para compensar a carência de dados consiste na utilização de modelos de GWS robustos, em que os resultados não sejam comprometidos devido a limitações no número de amostras.

Os modelos de GWS implementados em diferentes estudos são sensíveis a alguns conflitos de informação, como por exemplo, *outliers*, que influenciam fortemente os parâmetros (Andrade & O'Hagan, 2011; Andrade, Omey & Aquino, 2017). A

Distribuição t-Student com baixos graus de liberdade (De Finetti, 1961; Lindley, 1968) foi apontada para resolver esse conflito, pois tem caudas pesadas e expressa melhor a dispersão. Além da t-Student, a dupla-exponencial apresenta um comportamento semelhante, mas com uma maior concentração em torno da média.

Portanto, com este estudo objetivou-se comparar o modelo RRBLUP a modelos bayesianos robustos de Seleção Genômica Ampla que utilizem a distribuição t-Student (Bayes t) e a dupla-exponencial (Bayes DE) na variável resposta (y_i) para predição de valores genéticos genômicos em ovinos da raça Santa Inês para a característica área de olho de lombo medida por ultrassonografia.

2 | MATERIAL E MÉTODOS

2.1 | População amostrada

A amostra foi composta por informações de 389 ovinos da raça Santa Inês criados em rebanhos localizados nos estados do Piauí e Maranhão, Brazil, registrados na Associação Brasileira de Criadores de Ovinos (ARCO) ou pertencentes ao núcleo de conservação de caprinos e ovinos da Embrapa Meio-Norte (Campo Maior, Piauí). O DNA utilizado para a genotipagem desses animais foi extraído de amostras de sangue. As informações fenotípicas utilizadas na pesquisa são referentes a coletadas realizadas entre os anos de 2013 e 2018. Os animais usados eram criados em sistema semi-intensivo, recebendo suplemento nutricional na estação seca, cuja pastagem era escassa, ou em sistema intensivo, em que a alimentação era oferecida apenas nos cochos e sem acesso a pastagem.

Os animais eram predominantemente do sexo feminino (92,3%) e foram distribuídos nos estados do Piauí (84,4%) e do Maranhão, situados na região Nordeste, cujo clima é o semi-árido. A raça foi selecionada devido a sua adaptabilidade ao clima e a sua rusticidade, que têm despertado a atenção dos criadores da região há décadas.

2.2 | Estrutura de dados

2.2.1 | Dados fenotípicos

A característica de carcaça avaliada foi a área de olho de lombo (AOL), medida em cm^2, obtida por meio de imagens ultrassonográficas do corte transversal do músculo *Longissimus dorsi* entre a 12ª e 13ª vértebras lombares. A variável AOL, não negativa, foi utilizada como variável resposta nos modelos de Seleção Genômica Ampla.

Foram utilizadas informações sobre o sexo do animal ("M" – macho; "F" - fêmea), o estado de origem ("PI" – Piauí; "MA" – Maranhão) e a estação em que a medida foi aferida ("CH" – chuvosa, de janeiro a maio; "SE" – Seca, de junho a dezembro), que podem influenciar as características do animal de acordo com Liu et al. (2014). Estas variáveis foram usadas como efeitos sistemáticos nos modelos comparados.

A variável grupo contemporâneo (GC) foi proposta em função das variáveis estado chuvoso (de janeiro a maio Chuvosa (C); caso contrário, seca (S)), sexo (macho (M); fêmea (F)) e fazenda (6 propriedades – (A, B, C, D, E, F)), totalizando 13 combinações possíveis.

2.2.2 | Dados genômicos e controle de qualidade

Os 389 animais foram genotipados com o painel de alta densidade Ovine-SNP50k BeadChip BeadChip (Illumina Inc., San Diego, CA, EUA), que contém 54.241 SNPs uniformemente espaçados nos cromossomos do genoma ovino (Bush & Moore, 2012). De posse da base de dados genômicos, foi realizado o controle de qualidade com base nos seguintes parâmetros: *Call Rate* para amostras $\geq 0{,}85$ e para SNP $\geq 0{,}95$; menor frequência alélica (MAF) $> 0{,}05$; e pontuações GC $\geq 0{,}20$.

Os valores adotados nos critérios de qualidade estão acordo com Kemper et al. (2011) e Gianola et al. (2009). O controle de qualidade foi realizado no software R, utilizando os pacotes SNPStats, para leitura e formatação dos dados brutos fornecidos pela empresa de genotipagem, e HapEstXXR, que contém as funções responsáveis pela obtenção dos parâmetros de qualidade (Knueppel & Rhode, 2015). Após o controle de qualidade, restaram 29.834 SNPs.

2.3 | Bayes t e Bayes DE

Os modelos foram puramente bayesianos e utilizaram distribuições a *priori* equivalentes aos modelos já propostos, isto é, normal para os efeitos sistemáticos e para o efeito dos SNPs e qui-quadrado invertida para os componentes de variância.

Os ajustes dos modelos foram implementados com a inclusão de dados fenotípicos e genotípicos. Desta forma, assumiu-se o modelo aditivo (1)

$$y_i = \mu + \boldsymbol{x}'_{\boldsymbol{r}_i}\boldsymbol{\beta_r} + \boldsymbol{x}'_{\boldsymbol{l}_j}\boldsymbol{\beta_l} + \varepsilon_i \; \forall i = 1,2,\dots,n\,; j = 1,2,\dots,m, \tag{1}$$

em que μ é o intercepto, $\boldsymbol{\beta_r}$ é o vetor de parâmetros associado aos efeitos sistemáticos; $\boldsymbol{\beta_l}$ é o vetor associado ao efeitos dos marcadores genéticos; $\boldsymbol{x}'_{\boldsymbol{r}_i}$ e $\boldsymbol{x}'_{\boldsymbol{l}_j}$, respectivamente, são as matrizes de incidência dos efeitos sistemáticos e aleatórios dos SNPs; e ε_i é o resíduo do modelo, considerado independente e identicamente distribuído. A função de verossimilhança, com base no modelo tradicional com distribuição normal, foi escrita como (2)

$$p(y \mid \mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} N\left(y_i \mid \mu + \boldsymbol{x}'_{\boldsymbol{r}_i}\boldsymbol{\beta_r} + \boldsymbol{x}'_{\boldsymbol{l}_i}\boldsymbol{\beta_l}, \sigma_\varepsilon^2\right). \tag{2}$$

Desta forma, as distribuições a *priori* foram especificadas como (3)

$$\begin{aligned} p(\mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, u, \sigma_\varepsilon^2, \sigma_l^2) &= N\left(\mu|0, \sigma_\mu\right) N(\boldsymbol{\beta_r} \mid \mathbf{0}, \boldsymbol{I}\sigma_r^2) \\ &\times \chi^{-2}(\sigma_\varepsilon^2 \mid S_\varepsilon, d.f._\varepsilon)\, \chi^{-2}(\sigma_l^2 \mid S_l, d.f._l) \\ &\times N(\boldsymbol{\beta_l}|\mathbf{0}, \boldsymbol{I}\sigma_l^2) \; \forall\, i = 1,2,\dots,n\,; j = 1,2,\dots,m, \end{aligned} \tag{3}$$

onde σ_μ^2 e σ_r^2 são as variâncias de μ e β_{r_i}, respectivamente; $d.f.$ e S são o grau de liberdade e o parâmetro de escala correspondentes às distribuições χ^{-2} de σ_ε^2 e σ_l^2; e o índice i se refere ao número de animais e j à quantidade de marcadores.

Portanto, pôde-se representar hierarquicamente o modelo por (4) e (5)

$$y_i|\, \mu, \beta_r, \beta_l, \sigma_l^2, \sigma_e^2 \sim N(\mu + \boldsymbol{x}'_{\boldsymbol{r}_i}\boldsymbol{\beta_r} + \boldsymbol{x}'_{\boldsymbol{l}_i}\boldsymbol{\beta_l}, \sigma_\varepsilon^2) \tag{4}$$

$\mu \sim N\left(0, \sigma_\mu^2\right)$: Intercepto

$\boldsymbol{\beta_r} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_r^2)$: Efeitos sistemáticos do modelo

$\boldsymbol{\beta_{l_j}} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_l^2) \; \forall\, j = 1,2,3,\dots,m$: Efeito do marcador genético (5)

$\sigma_\varepsilon^2 \sim \chi^{-2}(\sigma_\varepsilon^2 \mid S_\varepsilon\,, df_\varepsilon)$: Variância residual

$\sigma_l^2 \sim \chi^{-2}(\sigma_l^2 \mid S = S_l\,, df = df_l)$: Variância associada a $\boldsymbol{\beta_l}$

No entanto, alguns estudos apontam que os conflitos de informação, como *outliers*, prejudicam as inferências dos parâmetros de localização (Barnett & Lewis, 1978; Desgagné & Angers, 2007; Balakrishnan & Ristíc, 2016). Portanto, a proposta consistiu em ajustar modelos cuja variável resposta do modelo (y_i) tenha distribuição t-Student e dupla-exponencial, já que a distribuição t com baixos graus de liberdade tem mais robustez quando comparada a distribuição normal (Andrade & O'Hagan, 2006; Andrade & Omey, 2013, 2016).

Desta forma, foi adotada a distribuição $t_{(v)}$ com v graus de liberdade e a dupla-exponencial com média μ e desvio-padrão σ para a variável resposta y_i do modelo, que continua a ser especificado da mesma forma de (1), mas as verossimilhanças passam a ser escritas como (6) e (7), respectivamente,

$$p(y \mid \mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} t\left(y_i \mid \mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l}, \sigma_\varepsilon^2\ , d.f.\right) \quad (6)$$

$$p(y \mid \mu, \boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_\varepsilon^2) = \prod_{i=1}^{n} DE\left(y_i \mid \mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l}, \sigma_\varepsilon^2\ \right). \quad (7)$$

Mantendo a mesma configuração com relação às distribuições a *priori*, tem-se a mesma estrutura de (3). Portanto, pode-se representar hierarquicamente os modelos por

$$y_i \mid \mu, \beta_r, \beta_l, u, \sigma_l^2, \sigma_u^2, \sigma_e^2 \sim t_v(\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l} + u_i, \sigma_\varepsilon^2)$$

$$y_i \mid \mu, \beta_r, \beta_l, u, \sigma_l^2, \sigma_u^2, \sigma_e^2 \sim DE\left(\mu + \boldsymbol{x'_{r_i}}\boldsymbol{\beta_r} + \boldsymbol{x'_{l_i}}\boldsymbol{\beta_l} + u_i, \sigma_\varepsilon^2\right),$$

de modo que as distribuições a *priori* são semelhantes a (5) e as condicionais completas são

$$\begin{aligned} p(\boldsymbol{\beta_r}, \boldsymbol{\beta_l}, \sigma_l^2, \sigma_\varepsilon^2) &= p(\boldsymbol{\beta_r}|\sigma_\varepsilon^2)p(\sigma_\varepsilon^2)p(\boldsymbol{\beta_l}|\sigma_l^2)p(\sigma_l^2) \\ &= \prod_{i=1}^{n} N\left(\beta_{r_i}, \sigma_\varepsilon^2\right) \times \chi^{-2}(\sigma_\varepsilon^2|d.f., S) \times \\ &\quad \prod_{j=1}^{m} N\left(\beta_{l_j}, \sigma_m^2\right) \times \chi^{-2}(\sigma_l^2|d.f., S_l), \end{aligned} \quad (8)$$

A Figura 1 apresenta as três distribuições mencionadas, a t de Student com 3 graus de liberdade, a normal padrão e a dupla exponencial padrão. Destaca-se que a $t_{(3)}$ apresenta caudas mais pesadas do que a distribuição $N(0,1)$, conferindo a mesma uma maior variabilidade e expressando uma ignorância acerca do parâmetro ou fenômeno de interesse. Por outro lado, a $DE(0,1)$ tem uma forte concentração em torno da média, bem como sua cauda se assemelha a $t_{(3)}$. Isto é, a Laplace pode ser usada em cenários com uma maior robustez e mais concentrados em torno da média.

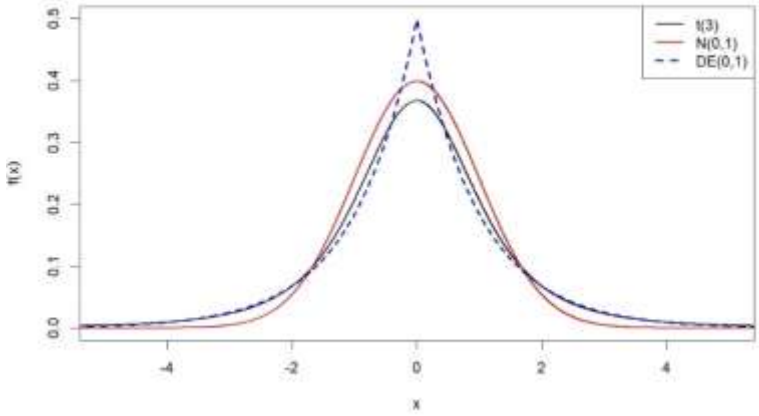

Figura 1 Densidades das distribuições $t_{(3)}$, $N(0,1)$ e $DE(0,1)$, utilizadas para ajustar a variável resposta dos modelos Bayes t, RRBLUP e Bayes DE, respectivamente

2.4 | Implementação dos ajustes

Os modelos foram compilados no *software* livre R (R CORE TEAM, 2018) através dos pacotes mcmc, usado para implementar o método iterativo Metropolis-Hastings, para obtenção da *posteriori* dos parâmetros via método de Monte Carlo via Cadeias de Markov, que de acordo com Chen & Huang (2014) tem propriedades ótimas de convergência. Para tanto, foram elaboradas três funções que representam a função de verossimilhança, as distribuições a *priori* e a *posteriori*, que foram inseridas na função *metrop* do pacote mcmc.

Além do mcmc, foram usados os pacotes invgamma para a utilização da distribuição Qui-quadrado invertida como priori; mvtnorm para a implementação da distribuição normal multivariada; LaplacesDemon para o uso da distribuição Laplace (dupla-exponencial). Para comparar com os modelos propostos, foi utilizado o método RRBLUP (Meuwissen, Hayes, & Goddard, 2001), do pacote BGLR, se assemelha aos ajustes implementados.

Para as estimativas da distribuição *a posteriori* dos componentes de variância e dos demais parâmetros genéticos para o banco de dados, foi utilizada uma cadeia de 1.500.000 iterações, com descarte inicial de 200.000 (*burn-in*) e intervalo de amostragem a cada 100 ciclos. Os valores iniciais dos parâmetros dos efeitos aleatórios e sistemáticos foram definidos com base em De Los Campos et al. (2009) e Meuwissen, Hayes, & Goddard (2001), sugerindo distribuições a *priori* que expressam alta variabilidade.

Os modelos foram comparados de acordo como o proposto por Li et al. (2017), onde estatísticas referentes aos valores genômicos (GEBVs) obtidos dos animais foram utilizadas. Foram implementados os três modelos com o fim de avaliar quais fornecem indicadores consistentes de um ajuste mais adequado para os dados reais. Os indicadores propostos envolvem: (i) correlação de Pearson entre valores preditos e observados; (ii) inclinação da regressão dos valores estimados e os obsevados; (iii) análise residual, como medida de ajuste do modelo aos dados. (Resende et al., 2008). Ainda foi calculada a acurácia, considerada o indicador mais importante da qualidade do ajuste, dada por

$$Acc_i = 1 - \sqrt{\frac{PEV_{ij}}{\sigma^2_{a_j}}},$$

em que PEV_{ij} é o erro padrão de predição para o valor genômico estimado para o animal i e $\sigma^2_{a_j}$ é a variância genética aditiva para a característica j (BIF, 2018).

3 | RESULTADOS E DISCUSSÃO

Percebe-se que a característica área de olho de lombo (AOL), medida em cm^2, tem uma leve assimetria à direita, confirmada pelo coeficiente de assimetria de 1,25. O diagrama de caixa aponta a presença de animais com valores mais acima do intervalo interquartílico, evidenciando a presença de outliers (Figura 2).

A média da AOL foi de 8,44 cm^2, enquanto a mediana é 7,92 cm^2 e o desvio-padrão é 3,03 cm^2. Tais medidas de média e mediana reforçam a assimetria positiva, já que $\bar{X} > Md$. Este fenômeno é bastante comum ao se trabalhar com melhoramento genético, visto que é desejável selecionar os animais superiores, cujas características se destacam entre os demais, sejam acima ou abaixo da média.

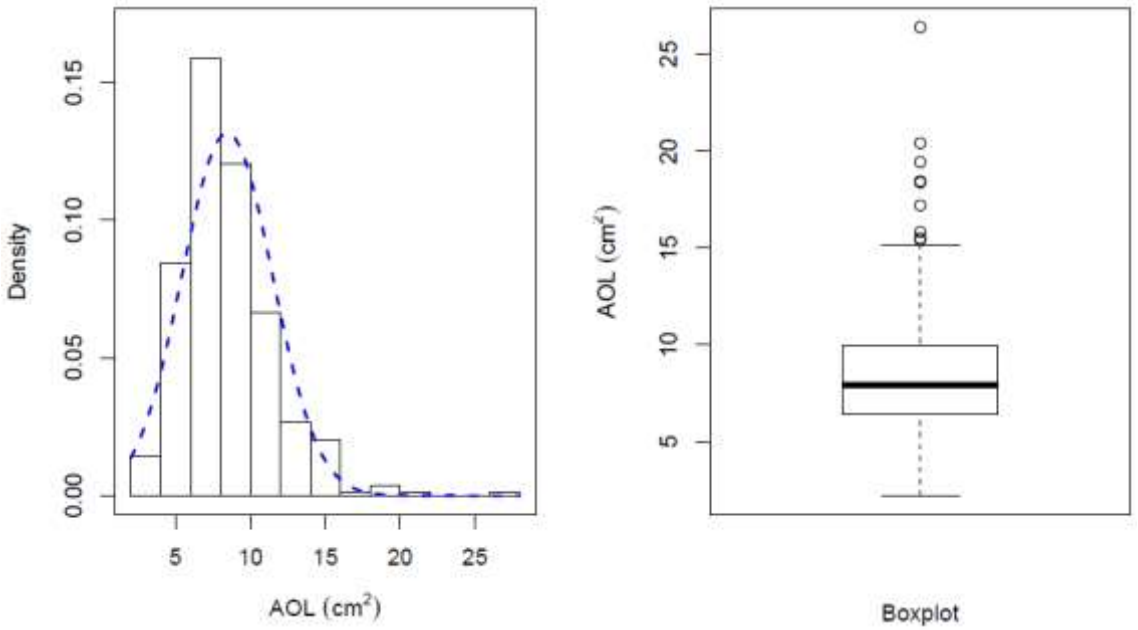

Figura 2 Histograma e *boxplot* da AOL (cm^2) dos 389 ovinos da raça Santa Inês estudados

Foram utilizados três modelos para base de comparação, tal que a principal diferença encontra-se na distribuição atribuída a y_i, adotada no modelo de GWS. A estratégia tem o propósito de avaliar as diferenças entre os ajustes, considerando apenas a substituição da distribuição normal pela t-Student e pela dupla-exponencial associada a y_i, principalmente em termos da robustez, justificada pelo pequeno tamanho da amostra e pela presença de *outliers*.

As médias para AOL foram 12,73 cm^2 para os machos, 8,08 cm^2 para as fêmeas, 8,57 cm^2 para a estação seca e 8,38 cm^2 para chuvosa. O menor DIC dos três ajustes foi obtido para o modelo RRBLUP, seguido pelo Bayes t e pelo Bayes DE. Mesmo com uma pequena diferença entre eles, o RRBLUP se mostrou melhor nesse critério (Tabela 1).

TABELA 1 DIC, média e desvio-padrão residuais, correlação entre os valores preditos e observados, acurácia e herdabilidade dos modelos RRBLUP, Bayes t e Bayes DE dos ovinos da raça Santa Inês

Modelo	DIC	Média Residual	SD residual	Cor($\hat{y}, y$)	Acurácia (%)	Herdabilidade (%)
RRBLUP	1.803,80	0,0011	2,1135	0,7855	37,0484	19,3422
Bayes t	1.821,55	-0,0001	1,8712	0,8459	42,8028	24,2453
Bayes DE	1.832,79	-0,0002	1,9672	0,8054	40,9327	22,8759

Quanto à análise residual, os modelos Bayes t e Bayes DE apresentaram valores mais próximos de 0 do que o RRBLUP e o mesmo aconteceu para o desvio-padrão, pois ambos os ajustes propostos se mostraram com resíduos menos dispersos, em especial o Bayes t (Tabela 1). A Figura 3 exibe os resíduos dos três modelos e deixa evidente a menor dispersão do Bayes t.

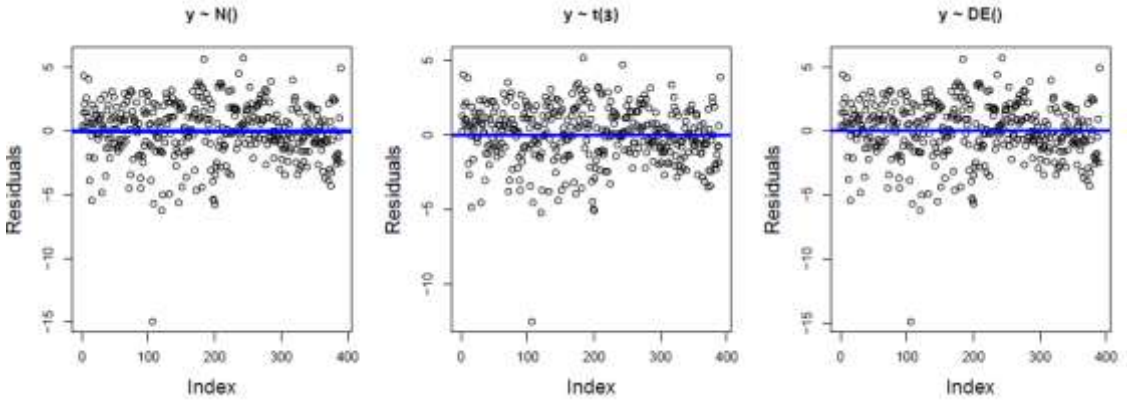

Figura 3 Resíduos dos modelos RRBLUP, Bayes t e Bayes DE para AOL (cm^2) dos 389 ovinos da raça Santa Inês

A correlação entre os valores preditos e observados no modelo Bayes t foi 0,8459, enquanto nos modelos RRBLUP e Bayes DE os valores foram 0,7855 e 0,8054, respectivamente (Tabela 1). Estes valores são próximos aos que foram apresentados por Meuwissen, Hayes & Goddard (2001) em relação ao RRBLUP, de modo que o modelo Bayes t foi superior. O mesmo acontece quando os valores são comparados àqueles obtidos por De Los Campos et al. (2009), que aplicaram um método de GWS mais complexo (LASSO Bayesiano). Apesar disto, os valores obtidos no presente estudo são aproximados. Quanto à acurácia, os modelos Bayest e Bayes L apresentaram-se melhores do que o RRBLUP, de modo que o Bayes t foi superior em relação aos demais.

As estimativas de herdabilidade assumiram valores condizentes com alguns resultados descritos na literatura (Sena et al., 2016; Figueiredo Filho et al., 2016; Figueiredo Filho et al., 2017). O quadrado do efeito dos SNPs para os três modelos ajustados está apresentado na Figura 4. Observa-se que o modelo Bayes t resultou em mais pontos com efeito sobre AOL, indicando que o Bayes t detectou mais SNPs em relação aos demais, já que apresenta mais picos em relação ao efeito geral dos SNPs.

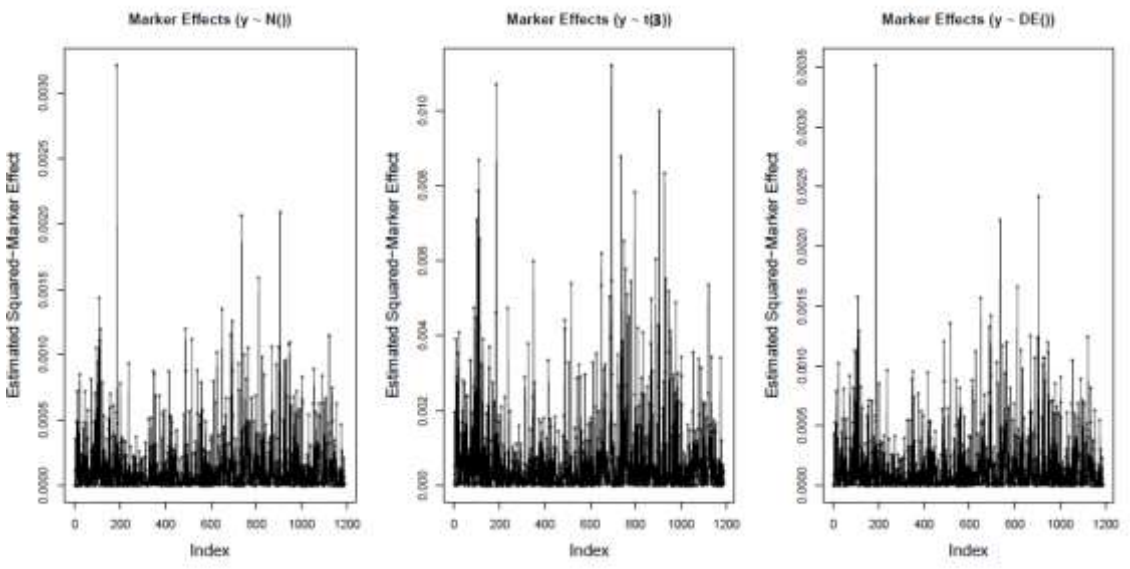

Figura 4 Quadrado do efeito dos SNPs dos modelos RRBLUP, Bayes t e Bayes DE para AOL (cm^2) dos 389 ovinos da raça Santa Inêsavaliados

A Figura 5 exibe a relação entre os valores preditos e os observados, onde novamente pode-se perceber que o Bayes t obteve melhores resultados em relação ao RRBLUP e ao Bayes DE, que apresentaram resultados semelhantes. O resultado se confirma pelas correlações (Tabela 2), que foram inferiores às obtidas por Clark et al. (2012) e semelhantes aos resultados de Bhuiyan et al. (2017).

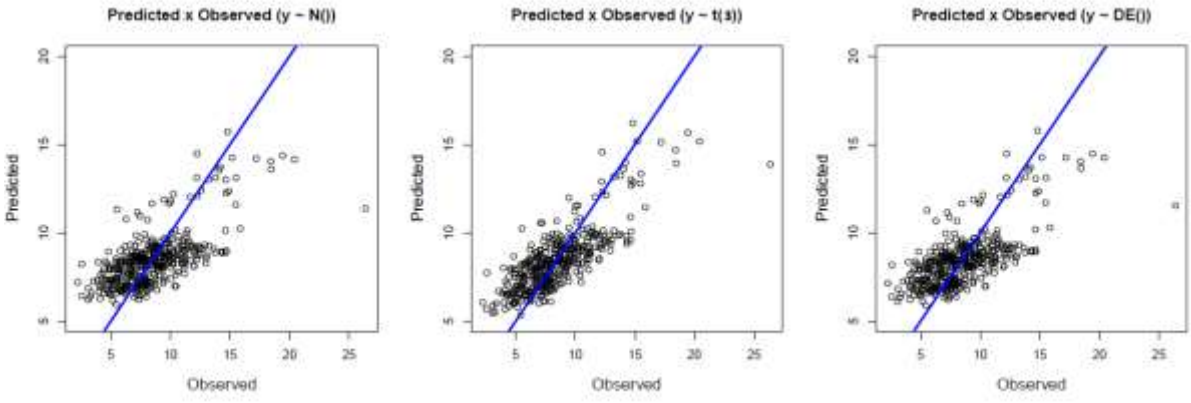

Figura 5 Valores preditos e observados dos modelos RRBLUP, Bayes t e Bayes DE para AOL (cm^2) dos 389 ovinos da raça Santa Inês

Quanto ao efeito dos marcadores SNPs, as correlações entre os efeitos dos três modelos são cor($\hat{\beta}_{\text{RRBLUP}}$, $\hat{\beta}_{\text{Bayes t}}$) = 0,9203, cor($\hat{\beta}_{\text{RRBLUP}}$, $\hat{\beta}_{\text{Bayes DE}}$) = 0,9775, cor($\hat{\beta}_{\text{Bayes DE}}$, $\hat{\beta}_{\text{Bayes t}}$) = 0,9411, indicando que os métodos propostos têm resultados semelhantes ao

já existente e consolidado RRBLUP, bem como as correlações entre os méritos genômicos dos animais cor($\widehat{GEBV}_{RRBLUP}$, $\widehat{GEBV}_{Bayes\ t}$) = 0,9529, cor($\widehat{GEBV}_{RRBLUP}$, $\widehat{GEBV}_{Bayes\ DE}$) = 0,9981, cor($\widehat{GEBV}_{Bayes\ DE}$, $\widehat{GEBV}_{Bayes\ t}$) = 0,9602 validam os modelos.

Quanto ao mérito genômico, um total de 98 animais se posicionou acima do terceiro quartil nos três modelos. Destes três grupos, 84 animais coincidiram entre o RRBLUP e o Bayes t, que corresponde a 85,71%. Quanto ao RRBLUP e o Bayes DE, 98,97% coincidiram, ou seja, 97 animais foram comuns aos dois grupos. Já entre o Bayes t e o Bayes DE, a interseção foi de 83 animais, representando 84,69% de semelhança. Tais resultados já eram esperados, tendo em vista a coincidência entre os modelos RRBLUP e Bayes DE (Vissher, Hill, & Wray, 2008).

Destaca-se ainda que os métodos propostos comportaram-se de forma esperada e mostraram-se melhores em quesitos, como acurácia, correlação entre valores preditos e observados e análise residual. Portanto, os modelos propostos foram razoáveis e podem ser alternativas viáveis de modelos de GWS. Ambos os resultados sugerem que os modelos propostos estimaram os valores de forma coerente, inclusive na seleção dos animais, em que houve pelo menos 84% de coincidência dos animais selecionados nos três modelos. Isto indica que, na prática, os animais selecionados, seriam diferentes, podendo levar a um equívoco na seleção.

Outro resultado importante está na correlação entre os valores preditos e observados, que se posicionou acima de 0,80 para o Bayes t, como esperado. Desta forma, dentre os modelos avaliados, o Bayes t se mostrou mais robusto a conflitos de informação, como *outliers* e dados com tamanhos amostrais pequenos. Tais conflitos são comuns em bancos de dados reais e, portanto, estes modelos podem ser usados como alternativas viáveis para a avaliação genômica com estrutura de dados limitadas.

4 | CONCLUSÃO

Foi constatada diferença na implementação dos ajustes com diferentes distribuições para a variável resposta AOL do modelo de seleção genômica ampla. Pela análise residual, o Bayes t se mostrou melhor, mais concentrado em torno 0 e menos disperso, enquanto o RRBLUP teve sua média mais distante de 0 e apresentou uma maior variabilidade. Exceto pelo DIC, todos os critérios adotados apontaram os modelos propostos foram melhores do que o RRBLUP, que pode ter sido reflexo da presença de *outliers* e pelo tamanho da

amostra, indicando uma alternativa viável para cenários com poucos animais genotipados e com dados com *outliers*.

5 | REFERÊNCIAS

Andrade, J. A. A.; O'Hagan, A. (2006). Bayesian Robustness Modeling Using Regularly Varying Distributions. *Bayesian Analysis*, v. 1, p. 169-188

Andrade, J. A. A.; O'Hagan, A. (2011) Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, v. 38, p. 691-711

Andrade, J. A. A.; Omey, E. (2013). Modelling conflicting information using subexponential distributions and related classes, *Annals of the Institute of Statistics and Mathematics*, v. 65, p. 491-511

Andrade, J. A. A.; Omey, E. (2016). Resolution of conflict of information using O-regularly varying functions. *Statistica Neerlandica*

Andrade, J. A. A.; Omey, E.; Aquino, C. T. M. (2017). Bayesian robustness modelling using the floor distribution. *REVSTAT Statistical Journal*, v. 16, p. 1-17

ARCO. (2018). *Associação Brasileira de Criadores de Ovinos: Padrões raciais*. Disponível em: http://www.arcoovinos.com.br/index.php/mn-srgo/mn-padroesraciais/40-santa-ines. Acesso em 06.12.2018

Balakrishnan, N.; Ristić, M. M. (2016) Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal. *Journal of Multivariate Analysis*, v.143 , p.194–207

Barnett, V.; Lewis, T. (1978). *Outliers in statistical data*. Wiley: New York, 1978.

Beef Improvement Federetion. (2018) *Guidelines for Uniform Beef Improvement Programs*. 9th ed. https://beefimprovement.org/library-2/bif-guidelines (Acessado em 15 de Abril de 2019)

Bowless, D. (2015). Recent advances in understanding the genetic resources of sheep breeds locally-adapted to the UK uplands: opportunities they offer for sustainable productivity. *Frontiers in Genetics*, v.6, n.24

Brito, L. F., McEwan, J.C., Millera, S., Bain, W., Lee, M., Dodds, K., Newman, S.A., Pickering, N., Schenkel, F.S, & Clarke, S. (2017a). Genetic parameters for various growth, carcass and meat quality traits in a New Zealand sheep population. *Small Ruminant Research*, n. 154, p. 81–91

Brito, L. F.; Clarke, S. M.; McEwan, J. C.; Miller, S. P.; Pickering, N. K.; Bain, W. E.; Dodds, K. G.; Sargolsaei, M.; Schenkel, F. S. (2017b) Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. *BMC Genetics*, v. 18, n. 7

Bhuiyan, M. S. A.; Kim, H. J.; Lee, D. H.; Lee, S. H.; Cho, S. H.; Yang, B. S.; Kim, S. D.; Lee, S. H. (2017). Genetic parameters of carcass and meat quality traits in

diferente muscles (longissimus dorsi and semimembranosus) of Hanwoo (Korean cattle). *Journal of Animal Science*, v. 95, p. 3359-3369

Chen, S.; Huang, J. (2014). Rates of convergence of extreme for asymmetric normal distribution. *Statistics and Probability Letters*, v.84, P. 158–168

Clark, S. A.; Hickey, J. M.; Daetwyler, H. D.; Julius HJ; Van Der Werf, J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, v. 44, p. 1-4

Daetwyler H. D.; Hickey J. M.; Henshall, J. M.; Dominik, S.; Gredler, B.; Van Der Werf, J. H. J.; Hayes, B. J. (2010) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Animal Prod. Sci*., v. 50, p. 1004-1010

Daetwyler, H. D.; Swan, A. A.; Van Der Werf, J. H.; Hayes, B. J. (2012). Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multibreed sheep data assessed by cross-validation. *Genetic Selection Evolution*, v. 44., p.33

De finetti, B. (1961). The Bayesian Approach to the Rejection of Outliers. Proceedings of the 4th *Berkeley Symposium on Mathematical Statistics and Probability*. Volume I, 99-210. Berkeley, California: University of California Press

De Los Campos, G.; Naya, H.; Gianola, D.; Crossa, J. Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J. M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*, v. 182, p. 375–385

Desgagné, A. ; Angers, J. F. (2007). Conflicting information and location parameter inference. *Metron*, v. 65, p. 67–97

Fernando, R.L.; Garrick, D. (2013). Bayesian methods applied to GWAS. *Methods in Molecular Biology*, v. 1019, p. 237–274

Figueiredo Filho, L. A. S.; Do, Ó; A. O.; Sarmento, J. L. R.; Santos, N. P. S.; Torres, T. S. (2016). Genetic parameters for carcass traits and body size in sheep for meat production. *Trop. Anim Health Prod*, v. 48, p. 215–218

Figueiredo Filho, L. A. S.; Sarmento, J. L. R.; Do Ó; A. O.; Santos, N. P. S.; Sena, L. S.; Sousa Júnior, A. (2017). Estimate of genetic parameters for carcass traits and visual scores in meat sheep using Bayesian inference via threshold and linear models. *Ciência Rural*, Santa Maria, v. 47

Gianola, D.; De Los Campos, G.; Hill, W.G.; Manfredi, E.; Fernando, R.L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363

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

Habier D.; Fernando, R. L.; Kizilkaya, K.; Carrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, v. 12, n. 186

Haro-López, R. A.; Smith, A. F. M. (1999). On robust Bayesian analysis for location and scale parameters. *Journal of Multivariate Analysis*, v. 70, no. 1, p. 30–56

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. (2017). *Censo agropecuário : resultados preliminares*. Acesso em: 12 de maio de 2018.

Kemper K. (2011). The distribution of SNP marker effects for faecal worm egg count in sheep, and the feasibility of using these markers to predict genetic merit for resistance to worm infections. *Genetics Research Cambridge*, v.93, p.203-219

Knueppel, S.; Rohde, K. (2015) *Package 'HapEstXXR'*. Disponível em: http://cran.rproject.org/web/packages/HapEstXXR/HapEstXXR.pdf

Li, H. Su, G.; Jiang, L; Bao, Z. (2017). An efficient unified model for genome‑wide association studies and genomic selection. **Genetics Selection Evolution**, v. 49, n. 64

Liu, T.; Qu, H.; Luo, C.; Shu, D.; Wang, J.; Lund, M. S.; Su, G. (2014). Accuracy of genomic prediction for growth and carcass traits in Chinese triple-yellow chickens. *BMC Genetics*, v. 15, n. 110

Lobo, R.N.B. (2019). Opportunities for investiment into small ruminant breeding programmes in Brazil. *Journal of Animal Breeding and Genetics*, v. 136, n.5, p. 313-318

Lindley, D. V. (1968). The choice of variables in multiple regression. *Journal of the Royal Statistical Society*. Series B, v. 30

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

Meuwissen, T.; Hayes, B. J.; Goddard, M. E. (2016). Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, v.6, n.1, p. 6-14

Misztal, I.; Legarra, A.; Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, v. 97, p. 3943–3952

R Core Team R. (2018). A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria

Resende, M. D. V.; Lopes, P. S.; Silva, R,L.; Pires, I.E. (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, v. 56, p. 63-78

Sena, L. S.; Santos, G. V.; Torres, T. S.; Sousa Júnior, A.; Rego Neto, A. A.; Sarmento, J. L. R.; Biagiotti, D. (2016). Genetic parameters for carcass traits and body size of meat sheep. *Semin. Ciências Agrárias*, Londrina, v. 37, n. 4, suplemento 1, p. 2477-2486

Van Den Berg, I.; Fritz, S.; Boichard, D. (2013). QTL fine mapping with Bayes C(pi): a simulation study. *Genetics Selection Evolution*, v. 45, n.19

Visscher, P. M.; Hill, W. G.; Wray, N. R. (2008). Heritability in the genomics era – concepts and misconceptions. *Genetics*, v. 9, p. 255-266

Wang, Z.; Zhang, Z.; Yang, H; Wang, S.; Rong, E.; Pei, W.; Li, H.; Wang, N. (2014). Genome-wide association study for wool production traits in a Chinese Merino sheep population. *Plos One*, v.9, n.9

Wu, Y.; Fan, H.; Wang, Y.; Zhang, L.; Gao, X.; Chen, Y.; Li, J.; Ren, H. Y.; Gao, H. (2014). Genome-Wide Association Studies Using Haplotypes and Individual SNPs in Simmental Cattle. *Plos One*, v.9, n.10