MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CENTRO DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA TROPICAL

EXPEDITO HENRIQUE ULISSES PEREIRA

MÉTODOS DE SELEÇÃO GENÔMICA AMPLA APLICADOS A DADOS SIMULADOS PARA OVINOS COM DISTRIBUIÇÃO GAMA

TERESINA

2020

EXPEDITO HENRIQUE ULISSES PEREIRA

MÉTODOS DE SELEÇÃO GENÔMICA AMPLA APLICADOS A DADOS SIMULADOS PARA OVINOS COM DISTRIBUIÇÃO GAMA

Tese apresentada ao Programa de Pós-Graduação em Zootecnia Tropical da Universidade Federal do Piauí como requisito para obtenção do título de Doutor em Zootecnia Tropical.

Orientador: Dr. José Lindenberg Rocha Sarmento

TERESINA

2020

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Setorial CCA
Serviço de Representação da Informação

P436m Pereira, Expedito Henrique Ulisses.
Métodos de seleção genômica ampla aplicados a dados simulados para ovinos com distribuição gama / Expedito Henrique Ulisses Pereira. -- 2020.
86 f.: il.

Tese (Doutorado) – Universidade Federal do Piauí, Centro de Ciências Agrárias, Programa de Pós-Graduação em Zootecnia Tropical - Teresina, 2023.
"Orientador: Prof. Dr. José Lindenberg Rocha Sarmento."

1. Acurácia. 2. Avaliação genética. 3. Distribuição assimétrica. 4. *Ovis aries*. 5. SNP . I. Sarmento, Linderberg Rocha. II. Título.

CDD 636.3

Bibliotecário: Rafael Gomes de Sousa - CRB3/1163

MÉTODOS DE SELEÇÃO GENÔMICA AMPLA APLICADOS A DADOS SIMULADOS PARA OVINOS COM DISTRIBUIÇÃO GAMA

EXPEDITO HENRIQUE ULISSES PEREIRA

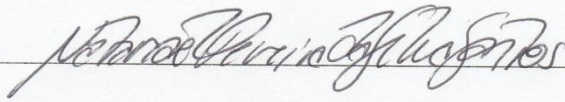
Defesa de tese de Doutorado aprovado em Teresina (PI), 27 de agosto de 2020.

Banca Examinadora:

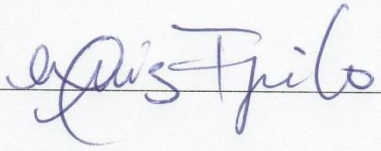
Prof. Dr. José Lindenberg Rocha Sarmento (Presidente) / DZO/CCA/UFPI

Max Brandão de Oliveira

Prof. Dr. Max Brandão de Oliveira (Interno) CCN / UFPI



Prof. Dr. Natanael Pereira da Silva Santos (Interno) / CPCE/UFPI



Dr. Luiz Antônio Silva Figueiredo Filho (Externo) / IFMA

Luciano Silva Sena

Dr. Luciano Silva Sena (Externo)

AGRADECIMENTOS

Primeiramente a Deus e todos os seres de luz que têm me dado força, fé e proteção a cada momento da minha vida.

À Universidade Federal do Piauí e ao Programa de Pós-Graduação em Ciência Animal (PPGCA/UFPI), pela oportunidade. Em especial, agradeço mais especificamente ao professor José Lindenberg Rocha Sarmento, pelo incentivo, dedicação, orientação, paciência e amizade.

Aos professores Dr. Luiz Figueiredo, Dr. Natanael Pereira da Silva Santos, Dr. Max Brandão de Oliveira e Dr. Luciano Silva Sena, pela participação na banca examinadora, pela ajuda e sugestões para o enriquecimento do trabalho.

Agradeço aos amigos e parceiros Max Brandão, Bruna Lima, André Campelo e Luciano Silva por sua ajuda, ensinamentos e parceria no meu trabalho, em uma etapa tão importante da minha vida acadêmica.

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

A cada um dos integrantes do Grupo de Estudos em Genética e Melhoramento Animal (GEMA - UFPI), manifesto minha mais sincera gratidão pela contribuição dada para o êxito deste trabalho.

A minha família e amigos, em especial meu irmão Rodrigo, minha Mãe Alda Alves Ulisses, minha sobrinha Rafaela e minha mulher Mariana Leopoldino, os quais foram essenciais durante a caminhada e aos quais dedico esta conquista.

Aos amigos e incentivadores Nonato e Tomaz, que passaram por minha vida ajudando no âmbito acadêmico. Aos meus amigos da Universidade Federal do Piauí.

Enfim, a todos que de alguma forma contribuíram para a concretização desta tese. Saibam que tiveram e têm sua importância e que o reconhecimento da minha parte está além das palavras escritas aqui.

MUITO OBRIGADO E MEUS SINCEROS AGRADECIMENTOS!

LISTA DE SIGLAS E ABREVIATURAS

AIC	Critério de informação de Akaike
AOL	Área de olho de lombo
ARCO	Associação Brasileira de Criadores de Ovinos
BLUP	*Best linear unbiased predictor* (Melhor preditor linear não viesado)
DIC	*Deviance Information Criterion* (critério de informação de desvio)
DP	Desvio padrão
EBV	*Estimated breeding value* (valor genético estimado)
EP	Erro padrão
FDP	Função densidade de probabilidade
FP	Função de probabilidade
GBLUP	BLUP genômico
GEBV	*Genomic estimated breeding value* (Valor genético genômico estimado)
GWS	*Genome wide selection* (Seleção genômica ampla)
IBGE	Instituto Brasileiro de Geografia e Estatística
IBLASSO	*Improved Bayesian Lasso* (Lasso Bayesiano aprimorado)
LASSO	*Least angle shrinkage and selection operator* (operador de penalização de menor ângulo e seleção)
MCMC	Monte Carlo via Cadeias de Markov
QTL	*Quantitative trait loci* (Locos de característica quantitativa)
RQR	Regressão Quantílica Regularizada
RRBLUP	Regressão de cumeeira do Melhor Estimador Linear Não Viesado
SAM	Seleção assistida por marcadores
ssGBLUP	*Single-step Genomic BLUP* (BLUP Genômico em Passo Único)
SNP	*Single nucleotide polymorphism* (Polimorfismo de nucleotídeo único)
VA	Variável aleatória
VAC	Variável aleatória contínua
VF	Variância fenotípica

LISTA DE SÍMBOLOS

%	Porcentagem
μ	Mi (média)
α	Alfa
×	Multiplicação
β	Beta
Kg	Quilograma
cM	Centimorgan
k	Kilobase (mil pares de base)
σ^2p	variância fenotípica
σ^2a	variância genética aditiva
σ^2e	variância ambiental
h^2	herdabilidade

LISTA DE ILUSTRAÇÕES

REFERENCIAL TEÓRICO

CAPÍTULO 1

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

RESUMO

PEREIRA, Expedito Henrique Ulisses. **Métodos de seleção genômica ampla aplicados a dados simulados para ovinos com distribuição gama.** 2020. 86p. Tese (Doutorado em Zootecnia Tropical) – Universidade Federal do Piauí, Teresina, 2020.

A distribuição de algumas características de importância econômica em animais apresenta comportamento assimétrico; neste caso, os métodos usuais de seleção genômica, baseados em expectativas condicionais impossibilitam a previsão de todas as distribuições de valores fenotípicos, podem sub ou superestimar os efeitos dos marcadores e, consequentemente, os valores genômicos. Objetivou-se com esta pesquisa verificar a sensibilidade e a capacidade preditiva dos métodos genômicos RR-BLUP, BLASSO e ssGBLUP, quando a pressuposição de normalidade da variável resposta não é atendida, com diferentes tamanhos amostrais, números de animais genotipados e níveis de variância fenotípica, para verificar o efeito dessa premissa sobre a predição de valores genéticos e sobre a acurácia preditiva. Dados genômicos foram simulados no software QMSIM, com uso de um genoma com tamanho idêntico ao genoma real da espécie *Ovis Aries*. Marcadores bialélicos foram distribuídos de acordo com o número de QTLs já conhecidos na literatura com efeito sobre a característica área de olho de lombo (AOL), e supondo no mínimo um QTL por cromossomo caso não houvesse relação com essa característica descrita na literatura. As estimativas de herdabilidade para a característica simulada e para o QTL foram 0,30 e 0,18, respectivamente. Foram considerados três níveis de variação fenotípica (5, 10 e 15) e três tamanhos populacionais (400, 1.000 e 5.000). No capítulo 1, foi realizada a comparação dos métodos RR-BLUP e BLASSO nos diferentes cenários. Quando os dados fenotípicos apresentaram distribuição gama, essa comparação foi feita através da análise de resíduos e da acurácia preditiva. As acurácias apresentaram redução de 6,6 e 10,9% para os métodos RR-BLUP e BLASSO, respectivamente, quando aplicados a dados fenotípicos com distribuição gama, porém analisados assumindo distribuição normal. Uma leve superioridade na acurácia foi verificada com o uso do método RR-BLUP, quando comparado ao BLASSO. No capítulo 2, foi verificada a sensibilidade do método ssGBLUP com diferentes números de animais de genotipados e com dados que apresentavam distribuição gama analisados assumindo distribuição normal. Ao quebrar a pressuposição de normalidade, a capacidade preditiva do método ssGBLUP diminuiu. Quando a pressuposição de normalidade para a variável resposta nos métodos de seleção genômica avaliados não foi atendida, a acurácia preditiva diminuiu.

Palavras-chave: acurácia; avaliação genética; distribuição assimétrica; *Ovis aries*; SNP.

ABSTRACT

PEREIRA, Expedito Henrique Ulisses. **Methods of genome-wide selection applied to data simulated for sheep using gamma distribution** 2020. 86p. Thesis (Doctor of Tropical Animal Science) – Federal University of Piauí, Teresina, 2020.

The distribution of some traits of economic importance in animals presents asymmetrical behavior; in this case, the usual methods of genomic selection, based on conditional expectations, make impossible the prediction of all distributions of phenotypic values, may under- or overestimate marker effects and, consequently, affect the prediction of genomic breeding values. The objective of this study was to verify the sensitivity and predictive capacity of the genomic methods RR-BLUP, BLASSO and ssGBLUP, when the assumption of normality of the response variable is not met, considering different sample sizes, number of genotyped animals and levels of phenotypic variance to verify the effect of this premise on the prediction of breeding values and predictive accuracy. Genomic data were simulated using the QMSIM software, with a genome with length identical to the real genome of the *Ovis Aries* species. Biallelic markers were distributed according to the number of QTLs known in the literature with effect on the loin eye area (LEA), and assuming at least one QTL per chromosome if there was no relationship with LEA described in the literature. The heritability estimates for the simulated trait and QTL were 0.30 and 0.18, respectively. Three levels of phenotypic variance (5, 10, and 15) and three population sizes (400, 1,000, and 5,000) were considered. In chapter 1, the RR-BLUP and BLASSO methods were compared in different scenarios. When the phenotypic data had a gamma distribution, the methods were compared based on residual analysis and predictive accuracy. The accuracies reduced by 6.6 and 10.9% for the RR-BLUP and BLASSO methods, respectively, when they were applied to phenotypic data with gamma distribution that were analyzed assuming normal distribution. A slight superiority in accuracy was verified with the use of the RR-BLUP method, when compared to BLASSO. In chapter 2, the sensitivity of the ssGBLUP method was verified with different numbers of genotyped animals and with data that presented a gamma distribution, but were analyzed assuming normal distribution. Violating the assumption of normality for the response variable decreased the predictive capacity of the genomic methods.

Keywords: accuracy; asymmetric distribution; genetic evaluation; *Ovis aries*; SNP.

SUMÁRIO

1 INTRODUÇÃO

O rebanho mundial de ovinos contava com aproximadamente 1,2 bilhão de cabeças em 2014, distribuídas em todos os continentes, com destaque para China, Austrália, Índia, Iran e Sudão como os maiores produtores. Nesse período, uma queda no efetivo de rebanho foi verificada nos países da América do Sul (FAOSTAT, 2017), embora o rebanho do Brasil tenha aumentado nos anos seguintes, evidenciando o crescimento da ovinocultura no país. Segundo dados do mais recente censo agropecuário, o rebanho ovino brasileiro tem aproximadamente 18,9 milhões de cabeças (IBGE, 2018).

No contexto do Brasil, a região Nordeste foi responsável por 66,7% do total do efetivo de ovinos estimado em 2018, com mais de 12,6 milhões de cabeças (IBGE, 2018). Essa região destaca-se por apresentar animais mais adaptados ao ambiente tropical, como por exemplo, ovinos da raça Santa Inês, que apresentam atributos de destaque para a produção de carne. No entanto, variáveis como rendimento de carcaça e cortes nobres não têm sido usadas como critérios de seleção nesta raça, devido à dificuldade de mensuração dessas características, principalmente em larga escala (MEIRA et al., 2019).

A identificação de indivíduos geneticamente superiores para características de interesse econômico é o principal objetivo da seleção de animais de produção. Desta forma, a busca de ferramentas que proporcionem melhor avaliação para esses indivíduos é imprescindível para que, com menor custo, sejam obtidos animais mais resistentes, com melhor qualidade de carcaça e com maior velocidade de crescimento.

Os métodos usados na identificação desses indivíduos têm evoluído continuamente, incluindo desde uma simples avaliação visual, passando por avaliações genéticas que usam informações de fenótipo e de pedigree (valores genéticos tradicionais), até modelos estatísticos mais avançados que incluem informações genômicas na predição do valor genético genômico dos indivíduos. Na predição de valores genéticos tradicionais, informações fenotípicas e de pedigree (genealogia) são utilizadas na inferência sobre os efeitos genéticos dos indivíduos. Em relação à predição genômica, as informações genotípicas, de pedigree e fenotípicas são utilizadas para realizar a inferência de futuros valores fenotípicos, ou seja, valores genéticos genômicos

preditos. Melhorias nessas predições têm sido possíveis devido à evolução computacional e da biotecnologia, como a identificação de marcadores genéticos.

Com esses avanços, novas formas de acessar marcadores em grande escala foram desenvolvidas. Dentre outros marcadores, destacam-se os SNPs (polimorfismos de um único nucleotídeo, *Single Nucleotide Polymorphisms*). Os polimorfismos no DNA são as origens da variação genética. Marcadores genéticos em desequilíbrio de ligação com variantes causais, ou seja, *loci* de características quantitativas (QTLs - *Quantitative Trait Loci*) podem ser utilizados para identificação de indivíduos superiores candidatos à seleção. Utilizando modelos estatísticos que incluíram esses marcadores amplamente distribuídos pelo genoma, Meuwissen, Hayes e Goddard (2001) idealizaram a seleção genômica ampla (GWS, *Genome Wide Selection*).

Com a aplicação da GWS, em resumo, é possível realizar a avaliação genética com uso de dados fenotípicos e genotípicos oriundos de painéis densos de SNPs. Para isto, os efeitos de SNPs são estimados à partir de uma população de animais com informação fenotípica e genotípica (população de referência ou treinamento), para em seguida ser realizada a predição de valores genéticos de animais candidatos à seleção, com base em seus genótipos (MEUWISSEN; HAYES; GODDARD, 2001).

A execução e implantação da GWS são acompanhadas de dificuldades estatísticas e computacionais. Por conta disso, vários métodos foram desenvolvidos, com diferenças basicamente no tipo de pressuposição assumida. A escolha do melhor método pode estar relacionada com o tipo de distribuição de efeitos genéticos da característica quantitativa em estudo e as pressuposições do efeito dos SNPs.

Entre os métodos propostos para a predição de valores genéticos com base em dados genômicos, se destacam os seguintes: Melhor Predição Linear Não-Viesada Genômica (conhecida como regressão de cumeeira GBLUP ou RRBLUP); Penalização Bayesiana ou *Bayesian shrinkage*, por exemplo, Bayes A e Bayes B (MEUWISSEN; HAYES; GODDARD, 2001; GIANOLA et al., 2009); Bayes C, Bayes D, Bayes Cπ e Bayes Dπ (Habier et al., 2011); operador de penalização de menor ângulo e seleção (LASSO, *least angle shrinkage and selection operator*) Bayesiano (BLASSO) e suas variações (DE LOS CAMPOS et al., 2009; PARK; CASELLA, 2008) e Melhor

Predição Linear Não-Viesada Genômica em passo único (ssGBLUP, *single-step genomic* BLUP).

As principais diferenças entre esses modelos são as pressuposições em relação à distribuição dos efeitos dos marcadores genéticos. Porém, em todos os casos, a escolha da distribuição associada à variável dependente (resposta) é inflexível, pois em todos os modelos apresentados acima esta variável assume distribuição normal.

A especificação incorreta da distribuição e do método pode levar a conclusões equivocadas, uma vez que isto pode afetar fortemente a distribuição a *posteriori* (ANDRADE; OMEY; AQUINO, 2017). Portanto, o uso de métodos que se adaptam à distribuição com dados assimétricos e seu comportamento em relação à média devem ser observados visando aumentar a acurácia.

Para melhor compreensão dos problemas relatados acima, objetivou-se com esta pesquisa verificar a sensibilidade dos métodos genômicos RRBLUP, BLASSO e ssGBLUP, com o uso de diferentes números de animais genotipados, para verificar o efeito dessa premissa sobre a predição de valores genéticos sobre a acurácia preditiva e, dessa forma, identificar estimativas mais plausíveis para a característica de interesse econômico, com diferentes perfis de variação e tamanho amostral.

O presente estudo está estruturado de acordo com as normas do Manual de Normalização de Monografia, Dissertação e Tese da Universidade Federal do Piauí. Primeiramente, tem-se o Referencial Teórico, que engloba os temas abordados nos Capítulos, bem como as respectivas Referências Bibliográficas utilizadas. Em seguida, é apresentado o Capítulo I, intitulado "Sensibilidade dos métodos RR-BLUP e BLASSO para dados com distribuição assimétrica gama com o uso de diferentes tamanhos amostrais e níveis de variabilidade fenotípica". Na sequência, temos o Capítulo 2, denominado "Método ssGBLUP aplicado a dados simulados em *Ovis aries* com distribuição Gama". Ambos os capítulos seguem as normas de publicação da revista *Animal Biotechnology* (ISSN: 1532-2378). A pesquisa se encerra com as Considerações Finais, onde são apresentados aspectos de relevância para continuidade do estudo.

2 REFERENCIAL TEÓRICO

2.1 Contexto da ovinocultura de corte no Brasil

Com a crise internacional da lã na década de 1990, a demanda por raças especializadas na produção de carne em todo Brasil estimulou a criação de ovinos deslanados, principalmente na região Nordeste (VIANA, 2008). Desde então, a ovinocultura de corte no Brasil está em expansão, impulsionada pelo aumento da produção de animais, cotação de cortes nobres e adesão de grandes empresas na comercialização. A produção de carne é o maior insumo da ovinocultura na região Nordeste e tem sido observado um contínuo crescimento no número efetivo de ovinos na região.

Contrastando com esse cenário, existe ainda grande demanda no mercado consumidor para o aumento na produção de carne ovina no Brasil, visto que, quase 60% da carne consumida nacionalmente é importada, sendo o Uruguai praticamente o único fornecedor para o país (ALVES et al., 2014). O consumo da carne ovina pelo brasileiro é insignificante, quando comparado ao consumo de carne bovina, suína e aves (CONSTANTINO et al., 2018). A disparidade fica explícita quando se compara o consumo nacional per capita anual das carnes suína (14,1 kg), bovina (37,4 kg) e de aves (43,9 kg) com a quantidade de aproximadamente de 0,6 kg de carne ovina consumida por pessoa por ano (OECD/FAO, 2018).

Esta conjuntura pode ser explicada pelo baixo nível de organização da ovinocultura e desinteresse de investimento por parte dos produtores de ovinos no Brasil. Portanto, para atender o mercado consumidor interno crescente, deve-se melhorar a rede produtiva. Para isto, deve-se incluir o sistema de criação, comercialização, melhoramento genético e escolha de raças que atendam à demanda de qualidade e, ao mesmo tempo, a realidade do produtor.

Neste sentido, é importante a criação de animais de raças adaptadas às condições de clima tropical, como os ovinos da raça Santa Inês. Animais desta raça são caracterizados pela manutenção ou perda mínima de produção durante estresse térmico e escassez de alimentos, boa eficiência produtiva, resistência a doenças, longevidade e baixas taxas de mortalidade (MCMANUS et al., 2011), quando comparados a raças exóticas ovinas criadas no Brasil.

2.2 Melhoramento genético da Raça Santa Inês para característica de carcaça

De acordo com a Associação Brasileira de Criadores de Ovinos (ARCO), a raça Santa Inês foi desenvolvida no Nordeste, resultante do cruzamento intercorrente das raças Bergamácia, Morada Nova, Somalis e ovinos sem raça definida (SRD), de modo que suas características atuais são resultantes de seleção natural e fixação através de trabalhos técnicos de criadores. Os ovinos Santa Inês são deslanados, com pelos curtos (nas cores vermelha, preta, branca e chitada), sem chifres e de grande porte. Os machos adultos pesam entre 80 a 110 kg e as fêmeas de 60 a 70 kg. A carne desses ovinos apresenta excelente qualidade e baixo teor de gordura (ARCO, 2018).

Os programas de melhoramento de ovinos tiveram seu início com o PROMOVI (Programa de Melhoramento Genético em Ovinos), entre 1977 e 1995 no Rio grande de Sul, visando à melhoria na produção de carne e lã, se estendendo depois para outros estados (MCMANUS; PAIVA; ARAÚJO, 2010). Atualmente, os programas de melhoramento disponíveis no Brasil são geridos por órgãos públicos, privados e associações de criadores. A maioria desses programas é centralizada na melhoria de raças deslanadas, como a Santa Inês, especializadas na produção de carne.

Algumas medidas de dimensões corporais são normalmente usadas como critério de seleção em ovinos. Porém, estudos que abordem características que avaliam além da produtividade a qualidade da carne, como área do músculo *Longissimus Dorsi*, gordura intramuscular (marmoreio) e espessura de gordura subcutânea são importantes indicadores de desenvolvimento muscular (SOUZA et al., 2016). Tais características apresentam variabilidade genética suficiente para o melhoramento genético (SENA et al., 2016; FIGUEIREDO FILHO et al., 2016).

Estudos sobre avaliação genética de ovinos da raça Santa Inês foram realizados, por muito tempo, com base apenas em informações fenotípicas e de pedigree (SOUSA et al., 1999; SARMENTO et al., 2006; FIGUEIREDO FILHO et al., 2016; SENA et al., 2016). O uso da genômica em ovinos vem trazendo novas perspectivas para as cadeias produtivas, com a possibilidade de aplicação de resultados em ações de melhoramento genético animal. Estudos genômicos com ovinos Santa Inês são recentes (BIAGIOTTI, 2016; AMORIM et al., 2018; ALVARENGA et al., 2018; MEIRA et al., 2019; ROVADOSCKI et al., 2018; SANTOS, 2018; SENA et al., 2020) e têm contribuído para o entendimento de características de interesse econômico.

O uso de informações genômicas aperfeiçoou os programas modernos de melhoramento genético e contribuiu substancialmente para o aumento do progresso genético para uma variedade de características economicamente importantes. Porém, para melhor compreensão sobre as metodologias desenvolvidas, o entendimento das informações biológicas e estatísticas utilizadas é de fundamental importância.

2.3 Seleção Genômica

O melhoramento genético convencional tem por base apenas informações fenotípicas dos indivíduos, que geralmente estão viesadas devido à grande influência de fatores como o ambiente. Dessa forma, existe a necessidade de técnicas mais elaboradas e confiáveis. Neste sentido, vários procedimentos têm sido empregados envolvendo a biologia molecular, sobretudo com o surgimento de novas formas de acessar marcadores moleculares, auxiliando os melhoristas a obterem resultados mais acurados.

Uma proposição com enfoque molecular foi descrita por Lande e Thompsom (1990), por meio da seleção auxiliada por marcadores (SAM), caracterizada por usar informações genômicas e fenotípicas em conjunto. Essa metodologia depende da identificação de associação entre marcadores genéticos e *loci* de Características Quantitativas (QTL), assim como da associação entre marcadores, enquanto que esta depende da distância entre marcadores e genes-alvo (IBTISHAM et al., 2017).

A SAM tem potencial para incremento da produção animal pelo aumento da confiabilidade, mas a tecnologia de marcadores disponíveis inicialmente para sua execução não apresentou alta eficiência, pois os marcadores genéticos utilizados não apresentavam boa distribuição pelo genoma e tinham baixa associação com os genes alvo (SIMIANER, 2016). Isto resultava em pouco ganho para a maioria das características de interesse pecuário, que são governadas por muitos genes.

O desenvolvimento de novas metodologias de sequenciamento mudou radicalmente o cenário descrito acima, baixando custos e acelerando a velocidade na qual o trabalho pode ser realizado. As chamadas tecnologias de sequenciamento de segunda geração, como Roche 454 (MARGULIES et al., 2005), Solexa-illumina (BENNETT, 2004) e ABI Solid (VALOUEV et al., 2008) são capazes de produzir vastos conjuntos de dados (da ordem de milhões de bases sequenciadas) em um único experimento altamente automatizado, com duração de 48 horas. A alta cobertura do genoma com marcadores torna possível particionar a variabilidade genética aditiva de

uma característica por todo o genoma, permitindo estimar o valor de substituição de alelo em cada um dos *loci* envolvidos com o fenótipo.

Os dados gerados com uso das tecnologias mencionadas possibilitam a estimação do valor genético de um indivíduo com base nos genótipos de todos os marcadores associados com uma característica de interesse, utilizando informações diretamente do DNA na seleção, possibilitando alta eficiência e baixo custo na seleção. Essa variância genética pode ser mais bem explicada pelos marcadores genéticos do tipo SNP (*single nucleotide polymorphism*), devido a sua abundância, baixa taxa de mutação e facilidade de genotipagem (RESENDE et al., 2008).

Diante desse cenário, a seleção genômica ampla (GWS, *Genome Wide Selection*), idealizado por Meuwissen, Hayes e Goddard (2001), tornou-se atraente para o melhoramento genético. De acordo com Resende et al. (2008), a GWS é definida como a seleção simultânea para centenas ou milhares de marcadores, que cobrem todo o genoma de maneira densa, de forma que todos os genes de uma característica quantitativa estejam em desequilíbrio de ligação (LD) com pelo menos uma parte dos marcadores. Os marcadores em LD com os QTLs, tanto de grandes quanto pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo.

A estimação de parâmetros genéticos só pode ser mensurada capitalizando o LD entre marcadores e QTLs ligados. O desequilíbrio de ligação ou de fase gamética é uma medida que avalia a dependência ou não entre dois ou mais *loci*. Em um grupo de indivíduos, se dois alelos são encontrados juntos com frequência maior que a esperada, com base no produto de suas frequências, infere-se que tais alelos estão em desequilíbrio de ligação (RESENDE et al., 2008). Ainda de acordo com esse autor, a variação dos valores do desequilíbrio de ligação é entre 0 e 1, em que valores próximos de zero indicam equilíbrio ou independência entre alelos de diferentes genes e próximos de um indicam desequilíbrio ou dependência (ligação) entre alelos de diferentes genes.

Após a capitalização do desequilíbrio de ligação entre marcadores e QTLs, predições derivadas de dados fenotípicos e de genótipo (SNPs em alta densidade) em uma geração são utilizadas para obtenção de valores genéticos genômicos (GEBV, *Genomic Estimated Breeding Values*) dos indivíduos de qualquer geração subsequente, tendo por base os seus próprios genótipos marcadores. Os GEBVs equivalem à soma dos efeitos de cada alelo e de cada gene de caráter quantitativo. Existem várias equações para predição de GEBVs, de modo que a melhor metodologia é aquela que consegue captar com maior precisão os fatores envolvidos na característica em estudo.

Na prática da seleção genômica ampla, são definidas três populações: população de descoberta ou estimação; população de validação; e população de seleção. Essas populações podem exercer duas funções simultaneamente (uma só população utilizada para estimação e validação) ou as três funções ao mesmo tempo.

Na população de descoberta ou estimação, um grande número de marcadores é avaliado em um número moderado de indivíduos, os quais devem ter seus fenótipos avaliados para os vários caracteres de interesse. Nesta população, obtêm-se equações de predição de GEBVs. Essas equações associam a cada intervalo marcador o seu efeito no fenótipo de interesse (RESENDE et al., 2008). Assim, são detectados os marcadores que explicam os locos que controlam os caracteres e seus efeitos são estimados.

De acordo com Almeida et al. (2016), a população de validação possui um conjunto de dados menor que a de descoberta, quando usadas populações diferentes, e contempla os indivíduos avaliados para os marcadores SNPs para vários caracteres de interesse. Para calcular a acurácia, os GEBVs são preditos (usando os efeitos estimados na população de descoberta ou estimação) e submetidos à análise de correlação com seus valores fenotípicos observados. Os erros dos GEBVs e valores fenotípicos são independentes e a correlação entre esses valores são de natureza genética, devido ao fato de que a amostra de validação não está envolvida na predição do efeito dos marcadores.

Na população de seleção, existe apenas a genotipagem dos candidatos à seleção e as equações de predição derivadas na população de descoberta são então utilizadas na predição dos valores genéticos genômicos ou fenótipos futuros dos candidatos à seleção (CAVALCANTI et al., 2012). Nesta etapa, não é necessária a avaliação dos fenótipos, pois estes serão preditos com base nos valores estimados na população de referência.

Vários métodos estatísticos já foram propostos para predição de valores genéticos genômicos. Os métodos ideais para GWS devem contemplar alguns atributos, diferem entre si na suposição do modelo genético associado ao caráter quantitativo e, em grande parte deles, estão embasados no teorema de Bayes.

2.4 Métodos Bayesianos na seleção genômica ampla

2.4.1 Inferência Bayesiana

A inferência estatística lida com o problema de tirar conclusões sobre quantidades não observadas a partir de dados numéricos (quantidades observadas). As

quantidades não observadas podem ser de duas naturezas: quantidades que não são diretamente observáveis, tais como parâmetros que governam o processo hipotético que produz os dados observados e quantidades potencialmente observáveis, tais como observações futuras de um processo. 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).

Estudos podem ser feitos utilizando tanto a inferência clássica (ou frequentista) quanto a inferência Bayesiana. Na primeira, o conceito de probabilidade envolve uma sequência de repetições para um determinado evento, tratado como um subconjunto de θ. A teoria tem como base a 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 inferência estatística Bayesiana tem por base a distribuição condicional do parâmetro $\hat{\theta}$, dado o vetor de dados $\boldsymbol{y}$, ou seja, a distribuição *a posteriori* do parâmetro dadas as observações fenotípicas. 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 θi 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, visto que conecta a distribuição *a priori* à distribuição *a posteriori*. De posse desses componentes, é utilizado o teorema de Bayes para obter-se a distribuição *a posteriori* que, em termos de densidade de probabilidade, tem a seguinte formulação para distribuição de uma variável aleatória contínua:

$$f(\theta|y) = \frac{f(y,\theta)}{f(y)} = \frac{f(y|\theta)\ f(\theta)}{\int_R f(y|\theta)\, f(\theta)\, d\theta},$$

em que: θ é o vetor de parâmetros; y é o vetor de dados; $f(\theta|y)$ representa a distribuição condicional de θ dado y, ou distribuição *a posteriori*; $f(y|\theta)$ função densidade de probabilidade (FDP) da distribuição condicional de uma observação y dado θ; $f(\theta)$ FDP da distribuição *a priori*, também conhecida como a densidade marginal de θ (Esta função denota o grau de conhecimento acumulado sobre θ); $f(y|\theta)\ f(\theta)$ representa a função densidade conjunta de y e θ; $f(y) = \int_R f(y|\theta)\, f(\theta)\, d\theta$ representa a distribuição marginal ou preditiva de y com respeito a θ, onde R é a amplitude da distribuição de θ; E_θ significa a esperança com respeito à distribuição θ.

Os resultados de interesse gerados pelas análises Bayesianas são, em geral, as distribuições marginais *a posteriori* do parâmetro genético. O problema básico da implementação da análise Bayesiana refere-se à integração numérica de distribuições marginais. Vários algoritmos para aproximar integrais foram implementados, porém, estão sendo substituídos vantajosamente por métodos computacionais intensivos, os quais são essencialmente estatísticos, tais como métodos de simulação estocástica como o Método de Monte Carlos Via Cadeia de Markcov (MCMC).

Dentro do MCMC, os dois algoritmos mais adotados para obtenção das estimativas são o amostrador de Gibbs e o Metropolis Hastings, que são utilizados para geração de amostras de distribuição candidata para gerar o MCMC, contornando assim o problema da obtenção das distribuições marginais por métodos analíticos, aprimorando assim a utilização da GWS.

2.4.2 Principais métodos de GWS

O sucesso da GWS está atrelado à escolha do método de predição dos efeitos dos marcadores. De acordo com Resende et al. (2012), um método ideal para GWS deve contemplar três atributos que são: acomodar a arquitetura genética do caráter em termos de genes de pequenos e grandes efeitos e suas distribuições; realizar a regularização do processo de estimação em presença de multicolinariedade e grande número de marcadores, usando para isso, estimadores do tipo *shrinkage* (trata o efeito dos marcadores como variáveis aleatórias estimando-os simultaneamente encurtando os coeficientes de regressão); e realizar a seleção de covaráveis (marcadores) que afetam a característica em análise.

A multicolinearidade é definida como a presença de uma forte relação linear entre as vaiáveis explicativas, levando a inferências errôneas. Para identificar a presença de multicolinariedade, é utilizado o índice de condição, dado como a razão entre o maior e o menor autovalor da matriz de covariâncias entre as covariáveis.

Os principais métodos para a GWS podem ser divididos em três grandes classes: regressão explícita, implícita e com redução dimensional. Os métodos explícitos são divididos em dois grupos: métodos de estimação penalizada (RR-BLUP, LASSO, EN, RR-BLUP-Het) e métodos de estimação Bayesiana (BayesA, BayesB, Fast Bayes B, Bayes Cπ, Bayes Dπ, Bayes R, Bayes RS, BLASSO, IBLASSO, dentre outros).

Na família dos métodos de estimação penalizada (regressão linear) se destaca o método RR-BLUP. Este método admite a pressuposição de que os efeitos de marcadores são considerados aleatórios, com distribuição normal e variância homogênea (mesma variância para todos os SNPs). Esta variância, bem como a variância residual são consideradas como desconhecidas e podem ser estimadas juntamente com os efeitos dos marcadores mediante resolução das equações de modelos mistos (RESENDE et al., 2008).

As abordagens Bayesianas permitem que as variâncias dos efeitos dos SNPs sejam diferentes entre si. De acordo com Meuwissen, Hayes e Goddard (2001), a regressão Bayesiana pode ser utilizada nas situações em que se tem mais marcadores (covariáveis) do que observações, atribuindo distribuição *a priori* aos coeficientes de regressão.

Os modelos Bayes A e Bayes B apresentam forte dependência das distribuições *a priori* da variância do marcador (GIANOLA et al., 2009). Por outro lado, em relação a esses métodos, o modelo Bayes Cπ é menos sensível à pressuposição *a priori* da variância do marcador, pois todos os SNPs têm variância comum e a proporção de SNPs sem efeito (π) tem distribuição *a priori* uniforme, que é estimada durante a análise (HABIER et al., 2011).

No modelo Bayes C, π é considerado como tendo valor fixo (FERNANDO; GARRICK, 2013), que proporciona detecção mais acurada de QTLs em relação ao Bayes Cπ, principalmente para características com herdabilidade de moderada a alta magnitude e quando há número suficiente de informações disponíveis (VAN DEN BERG; FRITZ; BOICHARD, 2013).

No método Bayes Dπ, as informações consideradas a *priori* para a variância de efeito de marcadores são as mesmas do método Bayes B, em que a variância dos marcadores com efeito maior do que zero tem distribuição qui-quadrado invertida e escalonada com probabilidade 1- π e π quando o efeito do marcador é zero (HABIER et al., 2011).

No método LASSO (*Least Absolute Shrinkage and Selection Operator*) Bayesiano, ao contrário de outros métodos bayesianos, não existe π e uma distribuição controlada por λ é declarada para toda coleção de variância dos locos marcadores. Esse método determina a parametrização das distribuições *a priori* para os efeitos dos marcadores. A parametrização leva à estimação das variâncias individuais para cada marcador (τ_j) condicionalmente ao parâmetro de regularização utilizado (λ).

A formulação Bayesiana do lasso (BLASSO) inclui um termo de variância comum para modelar ambos os termos, os resíduos e efeito genéticos dos marcadores (PARK; CASELLA, 2008). Legarra et al. (2011) propuseram o método BLASSO melhorado (IBLASSO), o qual usa dois termos de variância, um para modelar os resíduos e outro para modelar os efeitos genéticos dos marcadores.

A predição do valor genômico pode envolver múltiplos passos (*multi-step*), os quais utilizam como informações de genealogia a matriz **A** (matriz de parentesco baseada no pedigree) e a matriz **G** (matriz de parentesco com informações genômicas), que geralmente captura mais informação do que a primeira. Porém, além de maior demanda computacional, as etapas dos métodos *multi-step* dependem de muitos parâmetros e suposições, os quais são difíceis de verificar. Quando a matriz de parentesco está incompleta e nem todos os animais da população são genotipados, a estimação dos efeitos dos SNPs com a utilização do método *multi-step* pode não ser acurada, pois se torna necessária a utilização de projeções dos fenótipos de indivíduos aparentados com aqueles que foram genotipados, ou seja, pseudo-fenótipos (LEGARRA et al., 2014).

Em virtude dos problemas na execução dos métodos *multi-step*, surgiu a necessidade de um procedimento unificado para a predição genômica (LEGARRA et al., 2014). Dessa forma, Legarra, Aguilar e Misztal (2009), Aguilar et al. (2010) e Christensen e Lund (2010) desenvolveram a teoria básica para o método BLUP Genômico em Passo Único (ssGBLUP, *Single-step Genomic* BLUP), que combina a informação das matrizes **A** e **G** na matriz híbrida denominada **H** (matriz híbrida), que

permite gerar predições para indivíduos genotipados e não genotipados simultaneamente.

Recentemente, Tonussi et al. (2017) apresentaram valores genéticos 30% mais acurados usando ssGBLUP em relação ao BLUP tradicional, em uma população simulada de bovinos de corte com presença de reprodutores múltiplos. A informação extra de animais não genotipados, a capacidade de explicar a pré-seleção e a independência de pseudo-fenótipos são parcialmente responsáveis por ganhos em acurácia com o uso do método ssGBLUP, em relação a outros métodos genômicos (LEGARRA et al., 2014).

Os métodos acima sugerem a distribuição normal para yi, sendo que para as *prioris* dos efeitos sistemáticos é utilizada a distribuição normal com heterogeneidade (Bayes A e B) ou não variância (RR-BLUP), mistura de distribuições, no caso de Bayes Cπ e exponencial dupla (Lassos). Para o mérito genômico do animal, é usada a distribuição normal; para a *priori* dos componentes de variância, são adotadas as distribuições qui-quadrado invertida (χ^{-2}) e exponencial dupla, enquanto para *a posteriori* das variâncias são utilizadas as distribuições χ^{-2} e gama invertida. Em relação aos marcadores, são utilizadas as distribuições normal, t-Student e dupla exponencial, todas simétricas.

Em todos os modelos descritos acima, o fenótipo assume distribuição normal $N(\mu, \sigma 2)$. Esta pressuposição é assumida devido à 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 normal apresenta caudas leves e nem sempre se ajusta aos dados atípicos, como por exemplo, os *outliers* ou dados com assimetria. Com isso, podem ocorrer distorções dos resultados da inferência, visto que a distribuição não é robusta na presença de informações discrepantes, apesar da inferência englobar informações de todo o conjunto.

2.4.3 Distribuições de probabilidade

A teoria das probabilidades é essencial para o estudo da genética e para realizar inferências. A probabilidade está em uma porção intermediária entre estatística e

matemática. Há várias interpretações de probabilidade como clássica, frequentista, lógica e subjetiva. Algumas definições são relevantes como:

- Espaço de probabilidade: possui $(\Omega, A, P[.])$, onde Ω é o espaço amostral, A é uma coleção de eventos e $P[.]$ é uma função de probabilidade com domínio em A;

- Probabilidade condicional: sejam C e B eventos de A, para um dado espaço amostral de probabilidade $(\Omega, A, P[.])$, a probabilidade condicional de um evento C dado o evento B é indicada por P[C/B] definida por:

$$P[\mathrm{C}|\mathrm{B}] = \frac{P[C \cap B]}{P[B]} \; se \; P[B] > 0.$$

Um conjunto de dados contém uma ou mais variáveis aleatórias. Em se tratando de características zootécnicas, várias pesquisas têm aplicado métodos quantitativos e objetivos. Associa-se a essas características o conceito de variável aleatória (VA), definida como uma função X que associa a cada elemento do espaço amostral $\omega \in \Omega$ a um número real $X(\omega)$ (BUSSAB; MORETTIN, 2014).

Uma VA pode ser qualitativa ou quantitativa (discreta ou contínua). Uma variável aleatória 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, 1976).

Associada a uma variável aleatória contínua (VAC) X qualquer, tem-se o conceito de função de probabilidade (FP), no caso de varáveis aleatórias discretas, e função densidade de probabilidade (FDP), no caso de variáveis aleatórias contínuas.

As Distribuições de Bernoulli, Binomial, Poisson, Hipergeométrica, Binomial negativa (Pascal) e Multinomial são exemplos de distribuições discretas. Uma variável aleatória não possui uma função de probabilidade que associe probabilidades a cada ponto ou valores de seu domínio. Estas probabilidades são calculadas para intervalo de valores do domínio através de uma FDP, que é uma função contínua, não negativa e satisfazem as propriedades $f(x) \geq 0$ para todo $x \in X$ e $\int_{-\infty}^{\infty} f(x)dx = 1$. O conceito de densidade (usado na construção do histograma) explica-se pelo fato das probabilidades poderem ser obtidas a partir dos cálculos de áreas. Basta fazer o número de classe tender ao infinito e suas amplitudes a zero. A linha poligonal, que é obtida pela união dos pontos médios da classe, corresponde à curva da função. Uma FDP é igual à probabilidade de que uma variável aleatória X assuma um valor inferior ou igual a x. Distribuições dos tipos uniforme, normal, exponencial, Gama, Beta, Cauchy,

Lognormal, Exponencial Dupla, Weibull, Logística, Pareto, t, F e Qui-Quadrado são exemplos de FDP.

As distribuições mais empregadas para modelar o processo de frequência são: Gama, Pareto, Lognormal, Weibull, Normal e Distribuições de Valores Extremos: GEV (*Generalized Extreme Value*) e GPD (*Genaralized* Pareto *Distribution*).

Distribuição normal

Amplamente utilizada na estatística, a distribuição normal auxilia nas aproximações para os cálculos de outras distribuições. Uma VAC X com distribuição normal ou Gaussiana com parâmetros μ (média) e σ^2 (variância) é denotada por $X \sim N(\mu, \sigma^2)$, se sua FDP for dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

em que $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ e $\sigma > 0$, respectivamente. A distribuição normal tem propriedades desejáveis em vários aspectos, como por exemplo, é simétrica e o estimador $\hat{\mu} = \bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ é não viesado de variância mínima.

Distribuição Gama

A distribuição Gama, além de ser simples, é também bastante flexível, pois tem um parâmetro de correção da forma da curva, possibilitando um melhor ajuste aos dados. É muito utilizada para modelar valores de dados positivos que são assimétricos à direita e maiores que zero. Uma VA que assume apenas valores não negativos tem distribuição Gama se sua FDP é dada por:

$$\boldsymbol{f(x) = \frac{\beta^{\alpha}\ x^{\alpha-1}\ \exp(-\beta x)}{\Gamma(\alpha)}} \text{ para } x \geq 0$$

onde os parâmetros da distribuição Gama, que podem assumir qualquer valor positivo, são: $\boldsymbol{\alpha}$ (parâmetro de forma) e $\boldsymbol{\beta}$ (parâmetro de escala).

A FDP da distribuição Gama pode apresentar uma grande variedade de formas, dependendo, portanto, do parâmetro de forma $\boldsymbol{\alpha}$. Para valores de $\boldsymbol{\alpha}$ muito altos, a distribuição Gama tende à Gaussiana. O parâmetro de escala $\boldsymbol{\beta}$ tem a função de esticar ou encolher (escalonar) a função de densidade gama para a direita ou esquerda,

dependendo das magnitudes gerais dos valores dos dados. Essa variação da curva pode ser mais bem entendida nas Figuras 1 e 2.

Figura 1 – Distribuição gama para os valores de $\boldsymbol{\alpha = 1, 2, 3, 4}$ **e** $\boldsymbol{\beta = 1}$

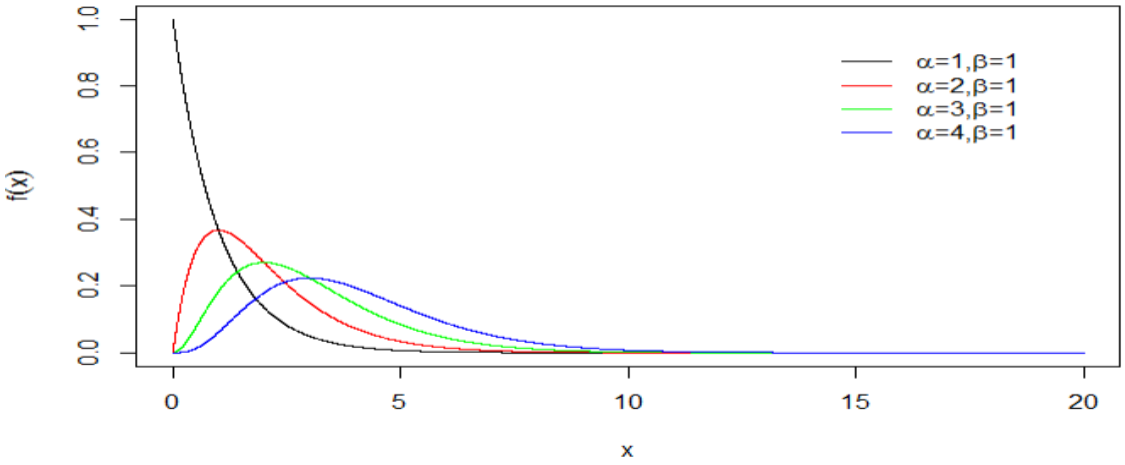
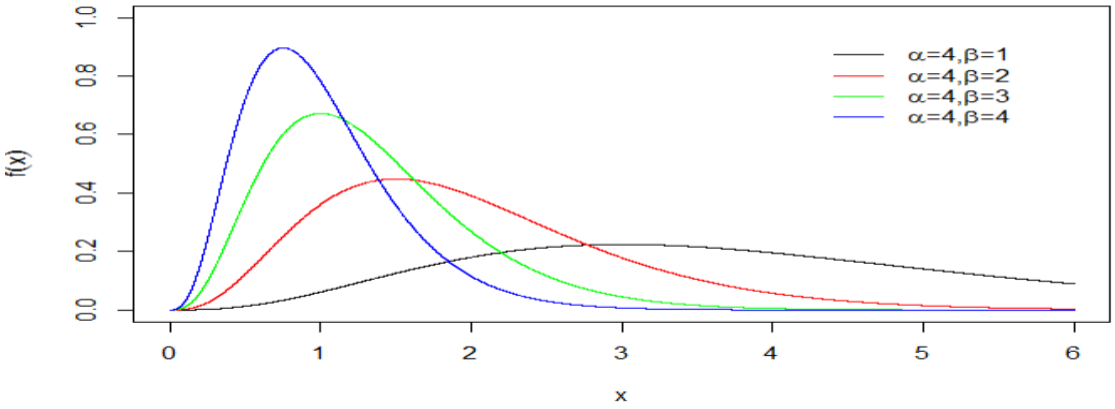


Figura 2 – Distribuição Gama para os valores de $\boldsymbol{\beta = 1, 2, 3, 4}$ **e** $\boldsymbol{\alpha = 4}$



No caso especial de $\boldsymbol{\alpha}$= $\boldsymbol{\beta}$ =1, tem-se a distribuição exponencial, pois Γ(1)=1, ficando-se com:

$$f(x) = \lambda e^{-\lambda x}$$

2.4.4 Ajuste de distribuição

Existem vários testes para verificar o ajuste de uma determinada distribuição aos dados. O teste de Kolmogorov ou Kolmogorov-Smirnov (K-S) é um dos mais utilizados. Este teste não se aplica a dados qualitativos nem a variáveis discretas, pois a tabela disponível para o teste só é exata caso a distribuição testada seja contínua.

O K-S consiste em encontrar a distância máxima entre as funções distribuição acumulada esperada e a observada. Para tal, é necessário obter uma distância máxima entre as duas. Posteriormente, essas distâncias são confrontadas com um valor teórico. Só assim, através do teste K-S é possível afirmar que a distribuição que está a ser testada se ajusta à amostra, com o nível de confiança requerido.

O teste de Kolmogorov-Smirnov destina-se a averiguar se uma amostra pode ser considerada como proveniente de uma população com uma determinada distribuição. As hipóteses em teste serão H_0: a população tem uma determinada distribuição D; e H_1: a população não tem a distribuição D. O teste funciona comparando-se, para cada número real x, duas percentagens: a percentagem de valores da amostra inferiores ou iguais a x; e a percentagem de valores da população inferiores ou iguais a x, admitindo que a população tem a distribuição D. Se o valor absoluto da maior das diferenças obtidas puder ser considerado suficientemente pequeno, então os dados levarão à aceitação da hipótese H_0. Além de conhecer à distribuição à qual os dados mais se aproximam, é necessário verificar qual modelo é mais apropriado.

2.4.5 Seleção e ajuste de modelos

As medidas mais comuns empregadas para avaliar o quão bem o modelo se ajusta aos dados em uma amostra são o LOG L e o coeficiente de regressão R^2. A soma dos quadrados dos resíduos é usada como medida de falta de ajuste e a super-parametrização ocorre quando há falhas ao fazer as predições em uma diferente amostra teste, mesmo que o modelo se ajuste bem aos dados. Em inferência Bayesiana, o modelo mais comum de avaliar o ajuste ou selecionar um modelo de melhor ajuste de diferentes metodologias estatístico estimado é a generalização do critério de informação de Akaike Bayesiano (AIC). O AIC admite a existência de um modelo "real" desconhecido que descreve os dados e tenta escolher, dentre um grupo de modelos avaliados, aquele que minimiza a divergência de Kullback-Leibler (K-L).

O critério DIC (*Deviance Information Criterion*) é uma generalização do AIC e é empregado em problemas de seleção de modelos bayesianos, onde as distribuições posteriores dos modelos são obtidas pela simulação de 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 próximo de 0 são os resíduos (GIANOLA; FERNANDO, 1986).

2.4.6 Simulação de dados

A simulação de dados é uma ferramenta que permite avaliar e validar métodos e modelos propostos, predizer futuras mudanças nos parâmetros genéticos, inclusive em áreas de pesquisa de mapeamento de QTLs e seleção genômica (SARGOLZAEI; SCHENKEL, 2009), e em diversas outras áreas. Essa ferramenta também permite auxiliar no planejamento ou escolha de estratégias que resultem em maiores ganhos genéticos, além de ser capaz de predizer os efeitos da seleção em longo prazo.

Com a dinâmica da simulação de dados, é possível criar vários cenários semelhantes aos encontrados em situações reais e, consequentemente, elaborar estratégias para sanar problemas, mensurar efeitos de interesse e entender mecanismos biológicos com menores custos de implementação e tempo. Existem vários programas de simulação de dados genômicos, entre eles se destaca o QMsim.

O programa QMsim foi elaborado para simular dados de genotipagem em larga escala. Resumidamente, a simulação neste programa é feita em duas etapas produzidas através da confecção de um *card*. A primeira, uma população histórica é simulada para estabelecer o equilíbrio mutação-deriva e, na segunda etapa, são geradas estruturas populacionais recentes, de acordo com a particularidade de cada estudo. O QMsim é

satisfatório quando se fala em demanda computacional, porém necessita de bons computadores e memórias.

Habier et al. (2013) estudaram as contribuições do LD, co-segregação e parentesco genético aditivo sobre a acurácia dos GEBVs dependendo da densidade de painéis de SNPs, tamanho da população de treinamento e extensão do LD. Tribout, Larzul e Phocas (2012) compararam em termos de tendência genética, acurácia e endogamia, a eficiência de um esquema atual de melhoramento de uma linhagem de machos suínos baseada na combinação de fenótipos de candidatos e de seus parentes com esquemas alternativos baseados na seleção genômica. Lillehammer, Meuwissen e Sonesson (2011) compararam esquemas alternativos para a implementação da seleção genômica para melhorar características maternas em suínos com esquemas de melhoramento convencional e de teste de progênie. Vários trabalhos foram realizados com o objetivo de comparar as acurácias de predição dos valores genéticos genômicos de diferentes métodos estatísticos (HABIER et al., 2007; MUIR, 2007; ZHANG et al., 2010; JIA e JANNINK, 2012; HOWARD et al., 2014).

3 REFERÊNCIAS

AGUILAR, I. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, n. 2, p. 743–752, 2010.

ALMEIDA, Í. F.; CRUZ, C. D.; RESENDE, M. D. V. Validação e correção de fenótipos na seleção genômica ampla. **Pesquisa Agropecuária Brasileira**, v. 51, p. 1973–1982, 2016. https://doi.org/10.1590/S0100-204X2016001200008.

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

ALVES, L. G. C. et al. 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.; OMEY, E.; AQUINO, C. T. M. Bayesian robustness modelling using the floor distribution. **REVSTAT Statistical Journal**, v. 16, p. 1-17, 2017.

ASSOCIAÇÃO BRASILEIRA DE CRIADORES DE OVINOS – ARCO, 2018. Padrões raciais. Disponível em: http://www.arcoovinos.com.br/index.php/mn-srgo/mn-padroesraciais/40-santa-ines/. Acesso em: 22 mai. 2019.

BENNETT, S. Solexa Ltda. **Pharmacogenomics**, v.5, n.4, p.433-8, 2004.

BIAGIOTTI, D. **Associação e seleção genômica ampla em ovinos Santa Inês para características relacionadas a resistência à endoparasitas**. 2016. 73f. Tese (Doutorado em Ciência Animal) – Programa de Pós-graduação em Ciência Animal, Universidade Federal do Piauí, 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.

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

CAVALCANTI, J. J. V. et al. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, v. 34, n. 3, p. 840-846, 2012.

CHRISTENSEN, O. F.; LUND, M. S. Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**, v. 42, n. 2, p. 1-8, 2010.

CONSTANTINO, C. et al. Desempenho, qualidade da carcaça e carne de ovinos de descarte de diferentes idades e gêneros. **PUBVET,** v.12, n.2, p.1-9, 2018.

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

FAOSTAT. Food and Agriculture Organization of the United Nations**. Production sheep by country**. Disponível em: www.fao.org/faostat/en/#data/QA/visualize. Acesso em: 29 ago.2017.

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.

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

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. **Genetics,** v. 177, n. 4, p. 2389–2397, 2007.

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

HABIER, D.; FERNANDO, R. L.; GARRICK, D. J. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. **Genetics**, v. 194, p. 597-607, 2013.

HAYES, S. C., LUOMA, J. B., BOND, F. W., MASUDA, A.; LILLIS, J. Acceptance and commitment therapy: Model, processes and outcomes. **Behaviour Research and Therapy**, v.44, p.1-25, 2006.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **Genes Genomes Genetics**, v. 4, n. 6, p. 1027–1046, 2014.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo agropecuário : resultados preliminares**. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3093/agro_2017_resultados_preliminares.pdf. Acesso em: 23 ago. 2019.

IBTISHAM, F.et al Progress and future prospect of *in vitro* spermatogenesis. **Oncotarget,** v. 8**,** n. 39, p.66709 - 66727 , 2017. https://doi.org/10.18632/oncotarget.19640

JIA, Y.; JANNINK, J. L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. **Genetics**, v. 192, n. 4, p. 1513–1522, 2012. https://doi.org/10.1534/genetics.112.144246

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics,** v. 124, n. 3, p. 743-756, 1990. https://doi.org/10.1093/genetics/124.3.743

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, v. 92, n.9, p. 4656-4663, 2009.

LEGARRA, A. et al. Improved Lasso for genomic selection. **Genetics Research**, v.93, n.1, p.77-87, 2011. https://doi.org/ 10.1017/S0016672310000534

LEGARRA, A. et al. Single step, a general approach for genomic selection. **Livestock Science**, v. 166, p. 54-65, 2014. https://doi.org/10.1016/j.livsci.2014.04.029

LI, H. et al. An efficient unified model for genome-wide association studies and genomic selection. **Genetics Selection Evolution**, v. 49, n. 64, p. 1-8, 2017. https://doi.org/10.1186/s12711-017-0338-x

LILLEHAMMER, M.; MEUWISSEN, T. H. E.; SONESSON, A. K. Genomic selection for maternal traits in pigs**. Journal of Animal Science**, v. 89, n. 12, p. 3908-3916, 2011. https://doi.org/10.2527/jas.2011-4044

KASS, R. E.; RAFTERY, A. E. Bayes Factors. **Journal of the American Statistical Association,** v. 90, n. 430, p. 773-795, 1995. https://doi.org/10.2307/2291091

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors**. Nature,** v.437, n.7057, p.376-380, 2005.

MCMANUS, C. et al. The challenge of sheep farming in the tropics: aspects related to heat tolerance. **Revista Brasileira de Zootecnia,** v.40, p.107-120, 2011.

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. https://doi.org/10.1590/S1516-35982010001300026

MEIRA, A. 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.13, n. 3, p. 460-468, 2019. https://doi.org/10.1017/S1751731118001362

MEUWISSEN, T.; HAYES, B.J.; GODDARD M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**. v. 157, n. 4, p. 1819-1829, 2001. https://doi.org/10.1093/genetics/157.4.1819

MEYER, P. L. **Probabilidade aplicações à estatística**. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 342–355, 2007. https://doi.org/ 10.1111/j.1439-0388.2007.00700.x

OECD/FAO - ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT / FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **Agricultural Outlook 2018-2027**, OECD Publishing, Paris/FAO, Rome, 2018. Disponível em: https://doi.org/10.1787/agr_outlook-2018-en. Acesso em: 26 abril 2019.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, 2008.https://doi.org/10.1198/016214508000000337

RESENDE, M. D. V. et al. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, v. 56, p. 63-78, 2008.

RESENDE, M.D.V. et al. **Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial**. Viçosa: Ed. UFV, v.1. p. 291, 2012.

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, p.1-14, 2018. https://doi.org/10.1186/s12864-018-4777-8

SANTOS, G. V. **Estudo genômico aplicado ao melhoramento genético de ovinos tropicais para resistência à endoparasitas**. 2018. 90f. Tese (Doutorado em Ciência Animal) – Programa de Pós-graduação em Ciência Animal, Universidade Federal do Piauí,, 2018.

SARGOLZAEI, M.; SCHENKEL, F. S.. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, v. 25, n.5, p. 680-681, 2009. https://doi.org/10.1093/bioinformatics/btp045

SARMENTO, J. L. R.; et al. Estimação de parâmetros genéticos para características de crescimento de ovinos Santa Inês utilizando modelos uni e multicaracterísticas. **Arquivo Brasileiro Medicina Veterinária e Zootecnia**, v. 58, n. 4, p. 581-589, 2006. https://doi.org/10.1590/S0102-09352006000400021

SENA, L.S. S. et al. Genetic evaluation of tropical climate-adapted sheep for carcass traits including genomic information. **Small Ruminant Research**. v. 188, 106120, 2020. https://doi.org/10.1016/j.smallrumres.2020.106120

SENA, L.S. et al. Genetic parameters for carcass traits and body size of meat sheep. **Semina: Ciências Agrárias**, v.37, p.2477-2486, 2016. https://doi.org/10.5433/1679-0359.2016v37n4Supl1p2477

SIMIANER, H. Genomic and other revolutions – why some technologies are quickly adopted and others are not. **Animal Frontiers**, v. 6, n. 1, p. 53-58, 2016. https://doi.org/10.2527/af.2016-0008

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. https://doi.org/10.1590/S1516-35981999000600012

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.

TONUSSI, R. L. et al. Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. **PLoS One**, v. 12, n. 9, e0181752, 2017. https://doi.org/10.1371/journal.pone.0181752

TRIBOUT, T.; LARZUL, C.; PHOCAS, F. Efficiency of genomic selection in a purebred pig male line. **Journal of Animal Science**, v. 90, n. 12, p. 4164-4176, 2012. https://doi.org/10.2527/jas.2012-5107

VALOUEV, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. **Genome Research**, v.18, n.7, p.1051-1063, 2008. https://doi.org/ 10.1101/gr.076463.108

VAN DEN BERG, I.; FRITZ, S.; BOICHARD, D. QTL fine mapping with Bayes C(π): a simulation study. **Genetics Selection Evolution**, v. 45, n.1, 2013. https://doi.org/ 10.1186/1297-9686-45-19

VIANA, J. G. A. Panorama Geral da Ovinocultura no Mundo e no Brasil. **Revista Ovinos**, v. 4, n.12, p. 1-9, 2008.

ZHANG, Z. et al. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. **Plos One**, v. 5, n. 9, e126482010. https://doi.org/10.1371/journal.pone.0012648

CAPÍTULO I

Sensibilidade dos métodos RR-BLUP e BLASSO para dados com distribuição assimétrica Gama com o uso de diferentes tamanhos amostrais e níveis de variabilidade fenotípica

Destaques

- A simulação de dados possibilita avaliar a sensibilidade de métodos genômicos a distribuições assimétricas.
- Os métodos RR-BLUP e BLASSO são menos acurados quando aplicados a dados com distribuição gama analisados assumindo distribuição normal.
- O método RR-BLUP foi mais acurado que o BLASSO nas condições deste estudo.

Sensibilidade dos métodos RR-BLUP e BLASSO para dados com distribuição assimétrica gama com o uso de diferentes tamanhos amostrais e níveis de variabilidade fenotípica

E. H. U. Pereira[a,b], B. L. Barbosa[b], L. S. Sena[b], M. B. Oliveira[c], J. L. R. Sarmento[d]

[a]Colégio Técnico de Teresina, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[b]Programa de Pós-graduação em Ciência Animal, Centro de Ciências Agrárias, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[c]Departamento de Estatística, Centro de Ciências da Natureza, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[d]Departamento de Ciência Animal, Centro de Ciências Agrárias, Universidade Federal do Piauí, Teresina, Piauí, Brasil

Autor para correspondência (contato) E. H. U. Pereira

e-mail: expedito.ulisses@ufpi.edu.br

Colégio Técnico de Teresina, Universidade Federal do Piauí, Campus Universitário Ministro Petrônio Portella, Teresina, PI 64049-550, Brasil

RESUMO

Este estudo objetivou verificar a sensibilidade dos métodos genômicos RR-BLUP e BLASSO para dados com distribuição assimétrica gama, assumindo diferentes tamanhos populacionais e níveis de variabilidade fenotípica. Dados genômicos foram simulados no software QMSIM. Marcadores bialélicos foram distribuídos de acordo com o número de QTLs já conhecidos na literatura que tenham efeitos sobre a característica área de olho de lombo (AOL) e supondo no mínimo um QTL por cromossomo, caso não tenha relação conhecida na literatura com essa característica. As estimativas de herdabilidade da característica simulada e do QTL foram 0,30 e 0,18, respectivamente. Foram considerados três níveis de variância fenotípica (5, 10 e 15) e três tamanhos populacionais (400, 1.000 e 5.000 animais). Para comparar os métodos avaliados quando se alterou a distribuição da variável reposta de normal para gama, utilizou-se a análise de resíduos, critério de informação da deviance e acurácia de predição. Os desvios apresentaram variação de 20 vezes ou mais quando se alterou a distribuição dos dados de normal para gama. Ao comparar as acurácias nos diferentes cenários, foi verificada uma leve superioridade do método RR-BLUP em relação ao BLASSO e os valores se elevaram com o aumento da população. No entanto, quando os dados seguiram distribuição normal, a acurácia foi superior em comparação aos valores obtidos com uso de dados apresentando distribuição gama. Portanto, conclui-se que ao assumir distribuição normal para variável resposta quando esta apresenta distribuição gama, as capacidades preditivas dos métodos RRBLUP e BLASSO são reduzidas.

Palavras-chave: acurácia; assimetria; métodos genômicos; simulação de dados; SNP

Sensitivity of RR-BLUP and BLASSO methods for data with asymmetric distribution gamma using different sample sizes and levels of phenotypic variability

ABSTRACT

This study aimed to verify the sensitivity of the genomic methods RR-BLUP and BLASSO for data with an asymmetric gamma distribution, considering different sample sizes and phenotypic variances. Genomic data were simulated using the QMSIM software. Biallelic markers were distributed according to the number of QTLs known to affect the loin eye area (LEA), and assuming at least one QTL per chromosome if there was no relationship with LEA described in the literature. The heritability estimates of the simulated trait and QTL were 0.30 and 0.18, respectively. Three levels of phenotypic variance (5, 10, and 15) and three population sizes (400, 1,000, and 5,000) were considered. When the distributions were transformed, the methods were compared based on residual analysis, Deviance Information Criterion, and predictive accuracy. Deviations varied by 20 times or more when the data distribution was changed from normal to gamma. RR-BLUP was more accurate than BLASSO, and the values increased as the population size increased. However, when data showed a normal distribution, the accuracy was higher compared to the values obtained using gamma-distributed data. Therefore, when assuming a normal distribution for the response variable with gamma distribution, the predictive abilities of the RR-BLUP and BLASSO methods are lower.

Keywords: accuracy, asymmetry, data simulation; methods, SNP

1 Introdução

Os métodos usados na identificação de indivíduos geneticamente superiores têm evoluído continuamente, abrangendo desde uma simples avaliação visual até os mais avançados, que incluem informações genômicas e fenotípicas simultaneamente. A seleção genômica ampla (GWS), idealizada no ano de 2001[1], foi a primeira forma de predizer o fenótipo futuro de uma população com base em informações de marcadores moleculares do tipo *Single nucleotide polymorphism* (SNP) e informações fenotípicas.

Devido às grandes dificuldades estatísticas e computacionais na implementação da seleção genômica ampla, diferentes métodos e modelos têm sido propostos. De acordo com Resende et al.[2], um método ideal para GWS deve contemplar estes três atributos: acomodar a arquitetura genética da característica, em termos de genes de pequenos e grandes efeitos e suas distribuições; realizar a regularização do processo de estimação em presença de multicolinariedade e grande número de marcadores, usando para isso, estimadores do tipo *shrinkage* (trata o efeito dos marcadores como variáveis aleatórias estimando-os simultaneamente encurtando os coeficientes de regressão); e realizar a seleção de covaráveis (marcadores) que afetam a característica avaliada.

Dentre os métodos propostos para a GWS, o RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) estima simultaneamente os efeitos dos marcadores que seguem distribuição normal[1,3]. Com o uso deste método, os efeitos dos marcadores são considerados como aleatórios, com variância comum, isto é, assume que todos os marcadores tenham a mesma contribuição para a variação genética. Porém, assumir mesma variância pode não ser adequado para marcadores localizados em regiões não associadas à variância genética, enquanto outros marcadores estão localizados em regiões associadas aos *loci* de Características Quantitativas (QTL)[4]. Segundo Almeida

et al.[5], para contornar esta situação, muitos autores propuseram metodologias que utilizam efeito *shrinkange* específico para cada marcador. No contexto Bayesiano, isto pode ser implementado usando *prioris* para os efeitos dos marcadores, a exemplo do método LASSO (*Least angle shrinkage and selection operator*)[6].

O estimador LASSO Bayesiano (BLASSO), além de ajustar uma variância separada para cada marcador, possui a vantagem de forçar os est[imadores. Isto permite, efetivamente, que alguns efeitos sejam iguais a zero, realizando simultaneamente o procedimento de *shinkage* e seleção de variáveis[5].

A escolha da melhor metodologia a ser utilizada está relacionada com a escolha do método, suposições e tipo de distribuição, tanto para os efeitos genéticos como para a variável resposta. Santos et al.[7] relataram que a distribuição de muitas dessas características apresenta comportamento assimétrico. Neste caso, os métodos usuais de seleção genômica, com base em expectativas condicionais, além de impossibilitarem a previsão de todas as distribuições de valores fenotípicos, podem sub ou superestimar os efeitos dos marcadores. De acordo com Nascimento et al.[8], embora vários métodos estatísticos propostos tenham se concentrado em lidar com problemas no que diz respeito à multicolinearidade e dimensionalidade em seleção genômica (por exemplo, RR-BLUP, Alfabeto bayesiano e regressões baseadas em kernel), poucos estudos examinaram desafios estatísticos com base em distribuições fenotípicas não normais (por exemplo, distribuições assimétricas), que são muito comuns na produção animal.

Portanto, o objetivo com esta pesquisa foi verificar a sensibilidade dos métodos genômicos RR-BLUP e BLASSO com a utilização de dados fenotípicos que não apresentam distribuição normal, com diferentes tamanhos amostrais e diferentes níveis de variabilidade fenotípica.

2 Material e métodos

2.1. Dados simulados

Uma população semelhante a ovinos (*Ovis aries*) foi simulada com o auxílio do programa QMSim[9], com utilização de parâmetros baseados em estimativas obtidas por Brito et al.[10] e Pertile et al.[11]. A simulação foi implementada utilizando o método *forward-time*, pelo qual forças evolutivas e desequilíbrio de ligação (LD) são simulados da geração mais Calafell et al.[12]. A simulação foi realizada em duas etapas: na primeira, uma população histórica foi simulada para estabelecer o equilíbrio mutação-deriva; na segunda etapa, foram geradas estruturas populacionais recentes, que podem ser complexas. O programa QMSim permite que vários parâmetros sejam incorporados nos modelos de simulação para produzirem dados apropriados[9].

2.1.1. População histórica

A população histórica foi formada inicialmente por 2.000 indivíduos por geração, permanecendo assim até a geração 1.000, acompanhada de um decréscimo gradual do tamanho da população até a geração 1.020, que foi composta por 1.000 animais, a fim de criar o LD inicial e estabelecer o equilíbrio mutação-deriva em gerações históricas. O número de machos decaiu de 1.000 na primeira geração para 500 na última geração histórica, em iguais proporções entre os sexos. O sistema de acasalamento foi baseado na união aleatória dos gametas, sem que tenha havido migração ou sobreposição de gerações, e com amostragem aleatória. Os números de machos e fêmeas foram mantidos constantes em cada geração e as frequências alélicas foram fixadas em 0,5.

2.1.2. Expansão populacional

No total, 1.000 indivíduos (500 machos fundadores e 500 fêmeas fundadoras) foram selecionados aleatoriamente da última população histórica. A fim de ampliar a população, oito gerações foram simuladas atribuindo três progênies por fêmea, por geração; união aleatória de gametas; crescimento exponencial do número de fêmeas; e nenhuma seleção foi considerada nesta etapa.

2.1.3. População recente

Essa etapa foi dividida em duas partes. Na primeira parte, a população POP1 foi composta por 400 machos e 10.000 fêmeas oriundos da última geração (geração 8) da população de expansão, em que foram selecionados e acasalados de maneira aleatória. Foi utilizada uma taxa de substituição de 60% dos machos e 20% das fêmeas, com taxa de crescimento populacional nula, considerando dois descendentes por fêmea por geração no decorrer de oito gerações. Em cada geração, os animais com melhores valores fenotípicos foram mantidos e os animais considerados de piores valores fenotípicos foram descartados. Os valores genéticos dos animais foram gerados pela metodologia *Best linear unbiased predictor* (BLUP), utilizando a metodologia de modelos mistos[13] para um modelo animal, considerando a verdadeira variância genética aditiva.

Na segunda parte (definição da população recente), foram simuladas as populações denominadas POP2, que diferiram somente na quantidade de variância fenotípica. Em todos os cenários, os fundadores das populações POP2, foram obtidos a partir da última geração (geração 8) da população POP1. Esses animais constituíram rebanhos através de oito gerações, onde foram selecionados e descartados animais nos mesmos moldes da população POP1. Foi considerada uma taxa de substituição de 20%

de machos e 5% das fêmeas. No total, foram simulados 164.010 animais ao longo de oito gerações na POP2. Após selecionar apenas os animais das três ultimas gerações, restaram 57.750 animais, a partir dos quais as diferentes análises foram realizadas. Foram simulados cenários cujos coeficientes de herdabilidade do QTL e da característica foram 0,18 e 0,30, respectivamente.

2.1.4. Genoma

O genoma foi simulado com 26 pares de cromossomos autossômicos semelhantes aos da espécie *Ovis aries*, com tamanho idêntico ao genoma real de ovinos com base na plataforma AnimalQTLdB (https://www.animalgenome.org/cgi-bin/QTLdb/index), totalizando 2657 cM. A vantagem de simular o número real de cromossomos com comprimento idêntico ao genoma ovino é criar um cenário mais realista com relação ao número de *loci* de marcadores e QTLs fisicamente desvinculados[10].

Foram distribuídos 12 mil marcadores bialélicos de forma equidistante, considerando a presença de 2014 QTLs bialélicos distribuídos através de pesos, de acordo com o número de QTLs já conhecidos na literatura que tenham efeitos sobre a característica área de olho de lombo (AOL) e supondo no mínimo um QTL por cromossomo que não tenha relação conhecida na literatura com essa característica. Os efeitos alélicos aditivos desses QTLs foram simulados a partir de uma distribuição gama, com parâmetro forma $\alpha = 0{,}4$. Os marcadores genéticos e QTLs foram obtidos levando em conta uma MAF (*minor allele frequency*) de 0,05. Os locos segregaram com dois, três ou quatro alelos cujas posições foram distribuídas aleatoriamente. A taxa de erro de genotipagem de marcadores foi de 0,005 e a taxa de genótipo de marcadores em falta foi de 0,01. Considerou-se uma taxa de mutação recorrente de 10^{-5} para marcadores e QTLs.

196

197 *2.2. Análises genético-quantitativas*

198

199 O efeito fixo do sexo com dois níveis foi considerado para a característica
200 simulada com informações semelhantes a AOL. O modelo geral adotado, que utiliza os
201 modelos mistos como princípio, pode ser dado a partir do seguinte modelo animal:

202

$$y = \mu + Xb + Wm + e$$

203

204 em que: y é o vetor das observações da característica estudada; μ é a média das
205 observações; b é o vetor de parâmetro associado ao efeito fixo; X e W são matrizes de
206 incidência dos efeitos fixos e aleatórios dos SNPs, respectivamente; m é o vetor de
207 efeitos aleatórios associado aos marcadores, com $m \sim$ N (0, $\mathbf{G}\sigma^2$), onde **G** é a matriz de
208 parentesco genômico[14]; e e é o vetor de erros aleatórios, assumindo que se tem
209 distribuição normal, com média zero e variância igual a σ_e^2: $e \sim$ N (0, $I\sigma_e^2$). As equações
210 de modelo misto genômico para predição de m equivalem a:

211

212 $$\begin{bmatrix} \mathrm{X'X} & \mathrm{X'W} \\ \mathrm{W'X} & \mathrm{W'W} + k.\,\mathrm{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathrm{b}} \\ \hat{\mathrm{m}} \end{bmatrix} = \begin{bmatrix} \mathrm{X'y} \\ \mathrm{W'y} \end{bmatrix}, \text{ em que } \mathrm{k} = \frac{\sigma_e^2}{\sigma_{g/n}^2}$$

213

214 O valor genético genômico do indivíduo j é dado por $VGG = y_i = \sum_i w_{ij}\, m_i$, em
215 que o componente w_{ij} refere-se ao elemento i da linha j da matriz W, referente ao
216 indivíduo j; e m representa as estimativas dos efeitos genéticos aditivos dos
217 marcadores. O método utiliza distribuição *a priori* normal para os efeitos sistemáticos e

para os efeitos dos marcadores e distribuição qui-quadrado invertida para os componentes de variância. Nesse método admite-se a pressuposição de que os efeitos de marcadores são considerados aleatórios, com distribuição normal e variância homogênea (mesma variância para todos os SNPs). Esta variância e a variância residual são consideradas como desconhecidas e podem ser estimadas juntamente com os efeitos dos marcadores mediante a resolução das equações de modelos mistos. No modelo acima, o fenótipo assume distribuição normal, assim como a maioria dos modelos de estimação.

As análises e ajustes aos dados foram realizados utilizando o *software* R versão 3.6.2[15]. Foram criados seis cenários com o mesmo coeficiente de herdabilidade para o QTL (0,18) e para a característica (0,30), como está representado na Tabela 1. Três níveis de variância fenotípica (5, 10 e 15) com três tamanhos de população (400, 1.000 e 5.000 animais) foram considerados. Os valores acima definidos foram utilizados para avaliar o desempenho dos métodos quando seus dados fenotípicos não apresentam distribuição normal, sob diferentes tamanhos amostrais e diferentes níveis de variabilidade fenotípica.

O controle de qualidade dos dados genômicos foi realizado com uso do software PREGSF90[16]. Foram removidos marcadores com MAF menor que 0,05, *Call Rate* menor que 0,95 e desvio extremo do equilíbrio de Hardy-Weinberg. Após o controle de qualidade, as análises foram divididas em duas categorias, de modo que em uma a variável resposta seguiu distribuição normal e na outra a variável seguiu distribuição gama.

Para avaliar a sensibilidade dos modelos sob a pressuposição de normalidade para a variável resposta, aplicou-se transformação de distribuição da variável resposta

simulada seguindo distribuição normal para distribuição gama, de acordo com a fórmula abaixo:

$x_{g=(x_n)^2}$, com os parâmetros $\alpha = \beta \,.\mu_g$ e $\beta = \frac{\mu_g}{(\sigma_g)^2}$,

onde x_n são os dados fenotípicos simulados a partir de uma distribuição normal; x_g são os dados fenotípicos transformados para distribuição gama; μ_g é a média dos dados; σ_g é o desvio padrão dos dados; e $\alpha \; e \; \beta$ são os parâmetros da distribuição gama.

Para verificação da efetividade da transformação dos dados, foi aplicado o teste de Kolmogorov-Smirnov no software R, utilizando a função *ks.test* com os dados e parâmetros mostrados acima. Além disso, o comportamento dos dados foi observado através de histograma, em todos os cenários.

Para cada cenário mostrado na Tabela 1 aplicou-se os dois modelos: um com dados fenotípicos sob normalidade; e outro transformado para distribuição gama. Isso gerou uma combinação de nove cenários com quatro modelos, dos quais dois seguiram distribuição normal e dois seguiram distribuição gama, totalizando 36 análises diferentes.

O pacote BGLR[17], contido no R, foi utilizado para aplicar os métodos RRBLUP e BLASSO, que já são consolidados e estão presentes na rotina do BGLR. A função com o mesmo nome do pacote mencionado acima estimou o efeito do sexo como fixo e os efeitos dos marcadores como aleatórios, todos com o componente aleatório σ_m^2.

Foram realizadas análises com cadeias de 1.000.000 de ciclos com *burn-in* de 100.000 amostras. Para evitar redundância nas informações devido às correlações seriais entre os valores gerados subsequentemente, foram tomadas amostras a cada 100 ciclos (intervalo de amostragem) para obter a distribuição *a posteriori*, que contou com 9.000 amostras.

Os modelos foram comparados por análise de resíduos, observando a média, desvio padrão, critério de informação da deviance (DIC) e acurácia. Como os dados foram obtidos por simulação, os valores genômicos verdadeiros dos animais foram fornecidos. Neste caso, a acurácia foi obtida em função da correlação de Pearson entre o valor genético verdadeiro e o valor genético predito. Quanto mais próximo de 1 essa correlação, melhor e mais acurado é o modelo.

3 Resultados e discussão

As mudanças nas distribuições dos dados foram verificadas utilizando o teste de Kolmogorov-Smirnov, bem como a verificação de seu comportamento através dos histogramas, como pode ser observado na Figura 1. Na parte superior da Figura 1, verifica-se a simetria na distribuição dos dados, enquanto na parte inferior a assimetria é positiva, o que pode caracterizar as distribuições normal e gama, respectivamente.

Na análise de resíduos, verificou-se que as médias ficaram próximas de zero, independente da variância fenotípica, número de animais ou método utilizado (Tabela1). Em relação à distribuição normal, o método RR-BLUP foi ligeiramente melhor, pois seus desvios foram menores quando consideradas as três variâncias. Somente no cenário BLASSO Gama com variância 5, este método foi superior ao RR-BLUP, de acordo os desvio dos resíduos. Verificou-se que, ao aumentar a variância fenotípica, o RR-BLUP apresentou menores desvios quando comparado ao BLASSO. Quando os dados seguem distribuição gama, seus desvios chegam a ser 20 vezes maiores, quando comparados aos que assumem distribuição normal, independente do método.

A discrepância nos desvios dos resíduos pode ser melhor observada na Figura 2, notando a amplitude do eixo y em cada uma das situações. Tal resultado era esperado, pois quando se quebra a pressuposição da normalidade da variável resposta, o modelo tende a ser menos acurado.

Desta forma, através da análise residual, os métodos RRBLUP e BLASSO, quando aplicados com dados sob distribuição normal, mostraram-se melhores, mais concentrados em torno de 0 e menos dispersos, em comparação ao cenário em que os métodos foram aplicados a dados com distribuição gama para a variável resposta.

Ao manter o mesmo tamanho amostral, dentro de cada método com sua distribuição específica, o valor do DIC aumentou à medida que a variância fenotípica aumentou (Tabela 3). Nota-se também que, ao fixar a variância fenotípica, o valor de DIC diminuiu de acordo com a redução do tamanho amostral. Isso faz sentido, visto que conforme a informação amostral diminuiu, a verossimilhança diminuiu, mas a quantidade de parâmetros continuou a mesma. Quando a distribuição da variável resposta é gama, o valor do DIC é sempre maior, quando comparado ao valor do DIC obtido com o método assumindo a distribuição normal. Isto indica melhor ajuste quando a distribuição normal foi assumida.

Ao comparar apenas a distribuição assumida para os dados dentro dos métodos, as maiores variações de acurácia foram observadas no cenário que possui variância fenotípica 15 e tamanho amostral 400, com o uso do método BLASSO (variação chegou a 12,3%), e no cenário com variância 5 e tamanho amostral 1.000, com uso do método RR-BLUP (variação de 7,1) (Tabela 3). Vale ressaltar que se observou o aumento da acurácia com o aumento do tamanho amostral dentro de cada distribuição com uso de diferentes métodos. De acordo com Hayes et al.[18], mais informações fenotípicas e

313 genotípicas resultarão em mais observações por alelo de SNP e, consequentemente,
314 maiores correlações de predição genômica serão obtidas.

315 Ao comparar as acurácias dos métodos nos diferentes cenários, foi verificada leve
316 superioridade do RR-BLUP em relação ao BLASSO. Almeida et al.[5] encontraram
317 resultados semelhantes ao compararem essas mesmas metodologias e concluíram que,
318 de maneira geral, o método RR-BLUP é mais acurado, enquanto o BLASSO é menos
319 viesado. Alguns autores apontam para a superioridade de métodos Bayesianos quando
320 aplicados na seleção genômica, em relação ao método BLUP[1,19,20]. Entretanto, há relatos
321 de inversão nesse comportamento[21,22]. Zhong et al.[22], por exemplo, obtiveram valores de
322 acurácia de 0,62 e 0,61, com uso dos métodos RR-BLUP e Bayes B, respectivamente.
323 De acordo com os autores supracitados, uma distribuição mais complexa de efeitos
324 aleatórios, como aquela utilizada em métodos Bayesianos, somente é útil quando os
325 marcadores estão fortemente associados com o QTL.

326 No presente estudo, ao mudar a distribuição dos dados para a distribuição gama,
327 estes apresentaram distribuição com comportamento assimétrico. Santos et al.[7]
328 relataram que, neste caso, a mediana ou outro quartil é possivelmente mais adequado
329 para explicar a relação funcional entre os marcadores e as variáveis avaliadas, uma vez
330 que a média não é a melhor medida para representar distribuições.

331 Uma alternativa para contornar o problema com a assimetria foi proposta por
332 Nascimento et al.[8], que utilizou Regressão Quantílica Regularizada (RQR) na seleção
333 genômica. De acordo com estes autores, o método RQR, além de contornar os
334 problemas de multicolinearidade e dimensionalidade, também leva em conta a
335 possibilidade de assimetria na distribuição do fenótipo avaliado. Além disso, este

método aumenta a precisão dos efeitos do marcador e, consequentemente, os valores genômicos estimados para características cujas distribuições mostrem assimetria.

4. Conclusão

A predição de valores genéticos com dados que assumem distribuição gama reduz a capacidade preditiva dos métodos RR-BLUP e BLASSO em até 7,1 e 12,3%, respectivamente. O método RR-BLUP apresentou leve superioridade em relação ao método BLASSO neste estudo. O desenvolvimento de novos métodos ou o aperfeiçoamento daqueles já existentes para possibilitar o uso de distribuições assimétricas poderá permitir ganho adicional em acurácia, por conta do melhor ajuste do método aos dados.

Declaração de competição de interesses

Os autores declaram que não têm competição de interesses financeiros ou relações pessoais que possam influenciar este trabalho.

Agradecimentos

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES – Registro número: 001).

359 **Referências**

360

361 1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using
362 genome-wide dense marker maps. *Genetics*. 2001;157(4):1819-1829.
363 doi:10.1093/genetics/157.4.1819
364 2. Resende MDV, Silva FF, Lopes PS, Azevedo CF. *Seleção Genômica Ampla (GWS)*
365 *via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão*
366 *Aleatória Multivariada (RRM) e Estatística Espacial*. Viçosa: Universidade Federal
367 de Viçosa; 2012.
368 3. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge
369 regression. *Genet Res*. 2000;75(2):249-252. doi:10.1017/s0016672399004462
370 4. Goddard ME, Hayes BJ. Genomic selection. *J Anim Breeding Genet*. 2007;124(6):
371 323-330. https://doi.org/10.1111/j.1439-0388.2007.00702.x
372 5. Almeida IF, Cruz CD, Resende MDV. Validação e correção de fenótipos na seleção
373 genômica ampla. *Pesq Agropec Bras.* 2016; 51 (12); 1973-1982.
374 https://doi.org/10.1590/S0100-204X2016001200008
375 6. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Statist Soc B*.
376 1996; 58(1):267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
377 7. Santos PM, Nascimento ACC, Nascimento M, Silva FF, Azevedo CF, Mota RR,
378 Guimarães SEF, Lopes PS. Use of regularized quantile regression to predict the
379 genetic merit of pigs for asymmetric carcass traits. *Pesq Agropec Bras*. 2018;53(9):
380 1011–1017. https://doi.org/10.1590/S0100-204X2018000900004
381 8. Nascimento M, Silva FF, Resende MDV, Cruz CD, Nascimento ACC, Viana JMS,
382 Azevedo CF, Barroso LMA. Regularized quantile regression applied to genome-

383 enabled prediction of quantitative traits. *Genet Mol Res.* 2017;16(1):1-12.
384 https://doi.org/10.4238/gmr16019538
385 9. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock.
386 *Bioinformatics*. 2009;25(5):680-681. https://doi.org/10.1093/bioinformatics/btp045
387 10. Brito FV, Braccini Neto J, Sargolzaei M, Cobuci JA, Schenkel FS. Accuracy of
388 genomic selection in simulated populations mimicking the extent of linkage
389 disequilibrium in beef cattle. *BMC Genetics.* 2011;12(80): 1-10.
390 https://doi.org/10.1186/1471-2156-12-80
391 11. Pértile SFN, Silva FF, Salvian M, Mourão GB. Seleção e associação genômica
392 ampla para o melhoramento genético animal com uso do método ssGBLUP. *Pesq*
393 *Agropec Bras*. 2016;51(10):1729-1736. https://doi.org/10.1590/S0100-
394 204X2016001000004
395 12. Calafell F, Grigorenko EL, Chikanian AA, Kidd KK. Haplotype evolution and
396 linkage disequilibrium: A simulation study. *Hum Hered*. 2001;51(1-2):85-96.
397 https://doi.org/10.1159/000022963
398 13. Henderson CR. *Applications of Linear Models in Animal Breeding*. Guelph:
399 University of Guelph; 1984.
400 14. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*.
401 2008;91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980
402 15. R Core Team R. *A language and environment for statistical computing*. R
403 Foundation for Statistical Computing, Vienna, Austria, 2018.
404 16. Misztal I, Tsuruta S, Lourenco DAL, Masuda Y, Aguilar I, Legarra A, Vitezica Z.
405 Manual for BLUPF90 family of programs. Available from
406 http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf. 2018.
407 Accessed May 20, 2019.

17. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198(2):483-495. https://doi.org/10.1534/genetics.114.164442

18. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92(2):433-443. doi:10.3168/jds.2008-1646

19. Usai MG, Goddard ME, Hayes BJ. LASSO with cross-validation for genomic selection. Genet Res (Camb). 2009;91(6):427-436. doi:10.1017/S0016672309990334

20. Crossa J, De los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*.2010;186(2):713-724. https://doi.org/10.1534/genetics.110.118521

21. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389-2397. doi:10.1534/genetics.107.081190

22. Zhong S, Dekkers JC, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics*. 2009;182(1):355-364. doi:10.1534/genetics.108.098277

Tabela 1. Cenários simulados no software QMSim com painel de densidade 12k e herdabilidades do QTL e da característica respectivamente iguais a 0,18 e 0,30, com diferentes variâncias fenotípicas e tamanhos amostrais.

Variância fenotípica	Tamanho populacional
5	400
	1000
	5000
10	400
	1000
	5000
15	400
	1000
	5000

QTL: loci de característica quantitativa.

Tabela 2. Análise de resíduos dos métodos para diferentes distribuições, tamanhos amostrais e variâncias fenotípicas.

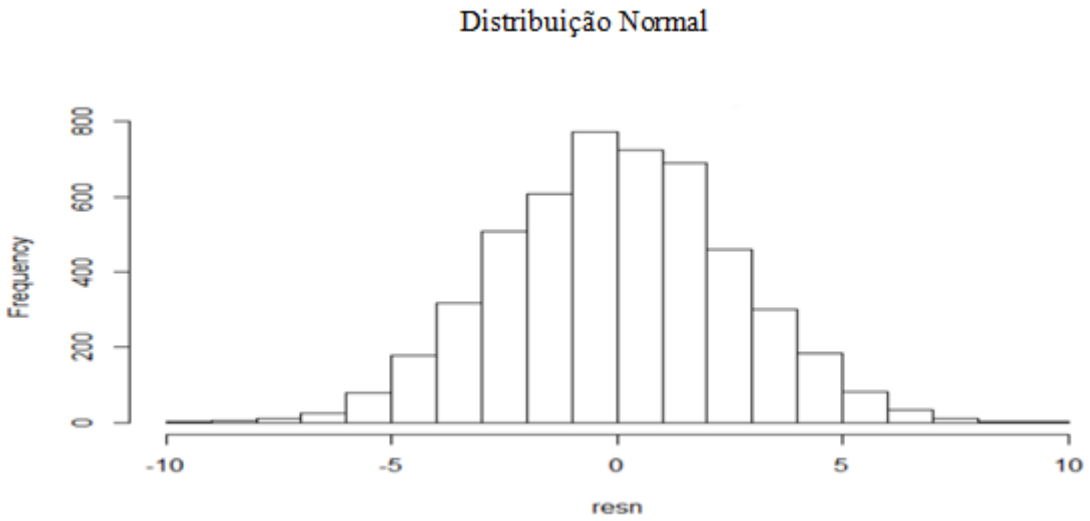
Método	Distribuição	N	Média dos resíduos			Desvio dos resíduos		
			VF = 5	VF = 10	VF = 15	VF = 5	VF = 10	VF = 15
RR-BLUP	Gama	400	0,0021	0,0006	-0,0578	21,28	41,44	51,60
		1000	0,0016	0,0086	-0,0077	20,66	46,74	64,47
		5000	0,0027	0,0019	-0,0038	23,24	47,06	67,47
RR-BLUP	Normal	400	0,0001	0,0025	0,0008	1,686	2,291	2,578
		1000	-0,0004	0,0004	0,0002	1,709	2,540	3,077
		5000	-0,0006	0,0003	0,0004	1,812	2,569	3,108
BLASSO	Gama	400	0,2591	0,0049	-0,0321	14,33	43,39	79,00
		1000	0,0099	-0,0493	0,0118	20,48	54,43	69,11
		5000	-0,0076	-0,0015	-0,0061	22,97	48,11	66,16
BLASSO	Normal	400	0,0026	0,0029	0,0014	2,045	2,889	3,221
		1000	-0,0010	0,0014	0,0021	1,917	2,842	3,448
		5000	-0,0003	-0,0002	0,00008	1,815	2,585	3,123

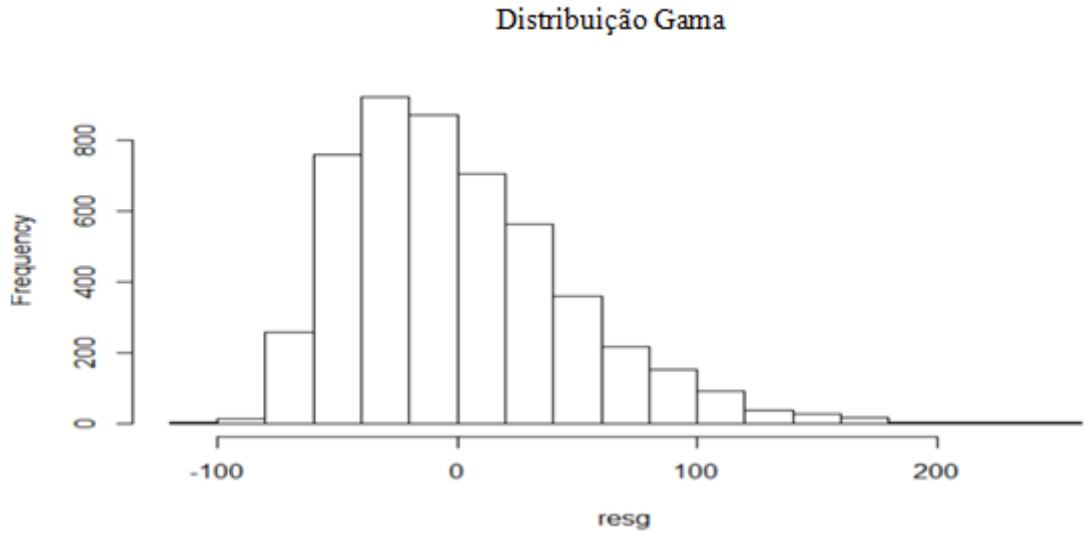
RR-BLUP: *Ridge Regression Best Linear Unbiased Prediction*; BLASSO: *Bayesian Least angle shrinkage and selection operator*; N: Tamanho amostral; VF: variância fenotípica.

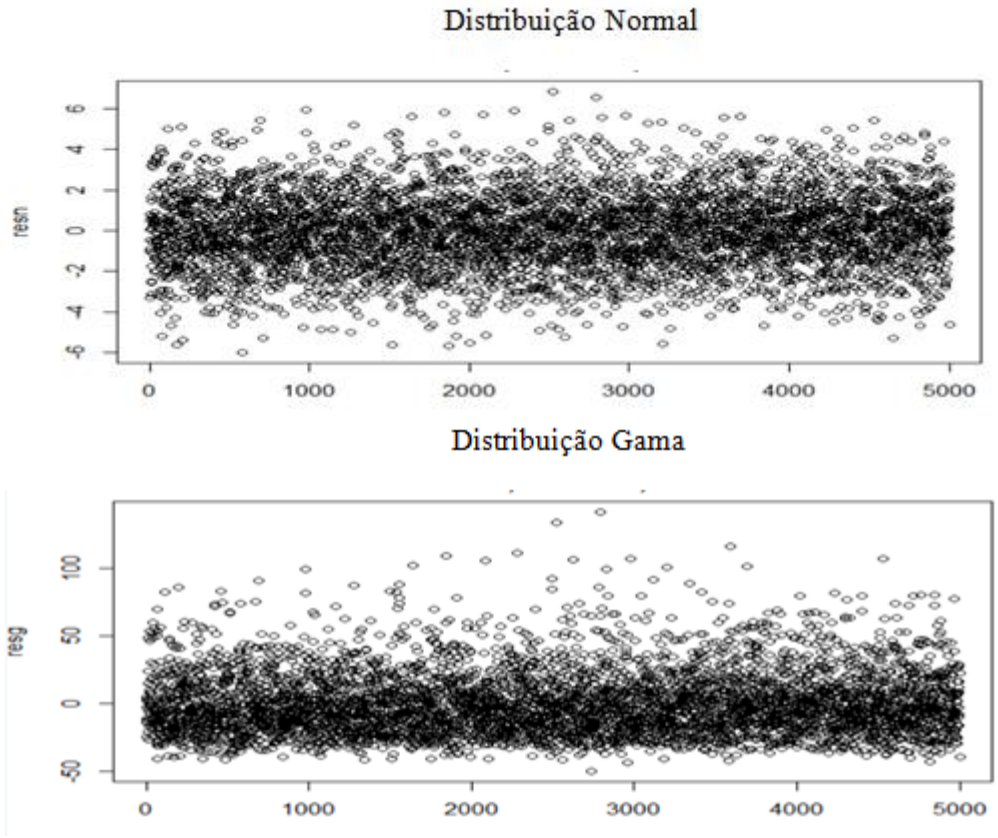
Tabela 3. Critério de Informação da Deviance (DIC) e acurácia dos métodos para diferentes distribuições, tamanhos amostrais e variâncias fenotípicas.

Método	Distribuição	Tamanho amostral	DIC			Acurácia		
			VF = 5	VF = 10	VF = 15	VF = 5	VF = 10	VF = 15
RR-BLUP	Gama	400	3806,73	4360,29	4349,96	0,4995	0,5315	0,5684
		1000	9350,84	10931,7	11712,5	0,5152	0,5442	0,6049
		5000	47220,1	52691,6	58165,6	0,6001	0,5920	0,6060
RR-BLUP	Normal	400	1776,55	2049,77	2186,32	0,5113	0,5419	0,5978
		1000	4366,16	5139,79	5562,37	0,5518	0,5555	0,6203
		5000	21860,0	25407,7	27412,5	0,6205	0,6056	0,6219
BLASSO	Gama	400	3750,19	4371,48	4724,45	0,4881	0,5222	0,5263
		1000	9346,23	10963,1	11735,6	0,5037	0,5252	0,5818
		5000	47140,1	54362,5	58086,1	0,5797	0,5643	0,5863
BLASSO	Normal	400	1795,59	2074,88	2220,40	0,5063	0,5356	0,5912
		1000	4383,73	5159,23	5484,04	0,5477	0,5508	0,6161
		5000	21855,3	25416,8	26420,3	0,6208	0,6053	0,6204

RR-BLUP: *Ridge Regression Best Linear Unbiased Prediction*; BLASSO: *Bayesian Least Angle Shrinkage and Selection Operator*; VF: variância fenotípica.


Distribuição Normal
Frequency
0
200
400
600
800
-10
-5
0
5
10
resn


Distribuição Gama
Frequency
0
200
400
600
800
-100
0
100
200
resg



resn: resíduos da distribuição normal; resg: resíduos da distribuição gama.

Legendas das figuras

Figura 1. Histograma dos dados simulados e transformados com variância fenotípica 10 e população com 5.000 animais.

Figura 2. Desvios de dados simulados e transformados com variância fenotípica 5 e população com 5.000 animais.

CAPÍTULO 2

Método ssGBLUP aplicado a dados simulados em *Ovis aries* com distribuição Gama

Destaques

- A simulação de dados possibilita avaliar a sensibilidade e capacidade preditiva de métodos genômicos a distribuições assimétricas dos dados.
- Os componentes de variância aumentaram em até 100 vezes com uso de dados normais assumindo distribuição gama.
- Quando a pressuposição de normalidade para a variável resposta não é atendida, a capacidade preditiva do método ssGBLUP diminui.

Método ssGBLUP aplicado a dados simulados em *Ovis aries* com distribuição gama

E. H. U. Pereira[a,b], B. L. Barbosa[b], L. S. Sena[b], M. B. Oliveira[c], J. L. R. Sarmento[d]

[a]Colégio Técnico de Teresina, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[b]Programa de Pós-graduação em Ciência Animal, Centro de Ciências Agrárias, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[c]Departamento de Estatística, Centro de Ciências da Natureza, Universidade Federal do Piauí, Teresina, Piauí, Brasil;

[d]Departamento de Ciência Animal, Centro de Ciências Agrárias, Universidade Federal do Piauí, Teresina, Piauí, Brasil

Autor para correspondência (contato) E. H. U. Pereira

e-mail: expedito.ulisses@ufpi.edu.br

Colégio Técnico de Teresina, Universidade Federal do Piauí, Campus Universitário Ministro Petrônio Portella, Teresina, PI 64049-550, Brasil

RESUMO

Objetivou-se com esta pesquisa verificar a sensibilidade e a capacidade preditiva do método ssGBLUP com uso de dados com distribuição assimétrica gama, considerando diferentes estruturas de variância fenotípica e número de animais genotipados. Os dados genômicos foram simulados no software QMSIM. Marcadores bialélicos foram distribuídos de forma de acordo com o número de QTLs já conhecidos na literatura que tenham efeitos sobre a característica área de olho de lombo (AOL). As estimativas de herdabilidade (h^2) para a característica simulada e pra QTLs foram assumidas com os valores de 0,30 e 0,18 respectivamente. Presumiu-se três níveis de variância fenotípica (5, 10 e 15) com dois tamanhos populacionais (1.000 e 5.000 animais). Para cada tamanho populacional foram genotipados 100, 50, 20 e 0% dos animais que foram avaliados com dados fenotípicos simulados seguindo distribuição normal e dados fenotípicos simulados transformados para distribuição gama. Para mensurar a capacidade preditiva do método ao se alterar distribuição normal da variável reposta para a gama, foram analisados os resíduos, os componentes de variância e a acurácia de predição. As estimativas de h^2 apresentaram variações diferentes em função da variância fenotípica assumida. Verificou-se que a variância genética aditiva e a variância residual aumentaram até 100 vezes quando os dados foram transformados para distribuição gama. As médias dos resíduos dos valores genéticos preditos aumentaram de 10 até 23 vezes quando se alterou a distribuição dos dados fenotípicos para gama nas variâncias 5 e 15, respectivamente. Quando a pressuposição de normalidade para a variável resposta no método ssGBLUP não é atendida, a capacidade preditiva diminui.

Palavra-chave: acurácia, distribuição assimétrica, ovinos, seleção genômica, SNP.

ABSTRACT

This study aimed to verify the sensitivity and predictive capacity of the ssGBLUP method using data with asymmetric gamma distribution, considering different structures of phenotypic variance and number of genotyped animals. Genomic data were simulated using the QMSIM software. Biallelic markers were distributed according to the number of QTLs known to affect the loin eye area (LEA). The heritability estimates (h^2) of the simulated trait and QTL were 0.30 and 0.18, respectively. Three levels of phenotypic variance (5, 10 and 15) and two population sizes (1,000 and 5,000) were considered. For each population size, 100, 50, 20 and 0% of the animals were genotyped and evaluated with normal- or gamma-distributed data. Residuals, variance components and predictive accuracy were analyzed to measure the predictive capacity of the method when changing the normal distribution of the response variable to gamma. The h^2 estimates showed different variations depending on the phenotypic variance. The additive genetic variance and the residual variance increased up to 100 times when data were gamma-transformed. The mean residuals of predicted breeding values increased from 10 to 23 times when gamma distribution was used considering variances 5 and 15, respectively. When the assumption of normality for the response variable in the ssGBLUP method is not met, the predictive capacity decreases.

Keywords: accuracy, asymmetric distribution, sheep, genomic selection, SNP.

1 Introdução

Por meio da integração da tecnologia genômica com ferramentas da genética quantitativa, a seleção genômica ampla (GWS) possibilitou o aperfeiçoamento de programas de melhoramento genético, propiciando progresso nos sistemas de avaliação genética animal, em diferentes espécies de interesse pecuário. Esta perspectiva promoveu melhoria na capacidade preditiva e ganhos que estão associados a intervalos de geração mais curtos, maior intensidade de seleção e maiores acurácias de predição do mérito genético de animais[1-3].

Antes da implementação da GWS os valores genéticos dos animais eram preditos apenas por meio de dados fenotípicos e de pedigree, utilizando a matriz de numeradores dos coeficientes de parentesco (**A)**. Entretanto, com a introdução de informações de marcadores genéticos de alta densidade, a matriz **A** passou a ser substituída pela matriz de parentesco genômico (**G**), que é produzida somente com as informações de marcadores moleculares, ou seja, formada apenas por animais genotipados.

Os métodos que utilizam somente a matriz **G** são conhecidos como *multi-step*, pois as informações genômicas são utilizadas separadamente nas análises, já que, normalmente, nem todos os animais com informações fenotípicas são genotipados. Neste sentido, uma alternativa para utilizar informações simultâneas de indivíduos genotipados e não genotipados foi proposta por Legarra, Aguilar e Misztal[4], Aguilar et al.[3] e Christensen e Lund[5], que desenvolveram a teoria básica para o método BLUP Genômico em Passo Único (*Single-step Genomic* BLUP – ssGBLUP).

Na abordagem do método *single-step*, pinformações de pedigree, fenótipos e genótipos são utilizadas em conjunto, combinando a matriz **A** com a matriz **G** em uma

matriz híbrida (**H**), para predizer os valores genéticos genômicos (GEBVs). Vários estudos já mostraram que o método ssGBLUP é computacionalmente eficiente e acurado para fins de avaliação genômica[3,6,7]. O método mencionado anteriormente é o mais apropriado para a obtenção de GEBVs confiáveis e não viesados, quando há falta de informações de pedigree para animais genotipados[8]. Apesar da evolução de diferentes métodos para buscar contornar problemas na predição de valores genômicos, ainda são raros estudos que levam em consideração a assimetria na distribuição dos dados.

Alguns métodos estatísticos propostos têm se concentrado em lidar com problemas no que diz respeito à multicolinearidade e dimensionalidade em seleção genômica (RR-BLUP, Alfabeto bayesiano, PLS, regressões baseadas em Kernel, etc). Entretanto, de acordo Nascimento et al.[9], poucos estudos examinaram os desafios estatísticos com base em distribuições fenotípicas não normais, por exemplo, distribuições assimétricas. Desta forma, novos estudos são necessários para desenvolver maneiras que contornem problemas como a presença de distribuições assimétricas nos dados.

Para avaliar novas metodologias com baixo custo de implementação e tempo, a simulação de dados é uma importante ferramenta que permite avaliar e criar vários cenários semelhantes ao real, e assim mensurar efeitos de interesse e estimar parâmetros genéticos para características de interesse. Portanto, o objetivo desta pesquisa foi verificar a sensibilidade do método ssGBLUP quando aplicado a dados com distribuição gama, com diferentes números de animais genotipados e com diferentes níveis de variância.

2 Material e métodos

2.1 Dados simulados

Uma população semelhante a ovinos (*Ovis aries*) foi simulada com o auxílio do programa QMSim[10], com utilização de parâmetros baseados em estimativas obtidas por Brito et al.[11] e Pertile et al.[12]. A simulação foi implementada utilizando o método *forward-time*, pelo qual forças evolutivas e desequilíbrio de ligação (LD) são simulados da geração mais antiga até a mais recente[13]. A simulação foi realizada em duas etapas: na primeira, uma população histórica foi simulada para estabelecer o equilíbrio mutação-deriva; e, na segunda etapa, foram geradas estruturas populacionais recentes, que podem ser complexas. O programa QMSim permite que vários parâmetros sejam incorporada nos modelos de simulação para produzirem dados apropriados[10].

2.1.2 População histórica

A população histórica foi formada inicialmente por 2.000 indivíduos por geração, permanecendo assim até a geração 1.000, acompanhada de um decréscimo gradual do tamanho da população até a geração 1.020, foi composta por 1.000 animais, a fim de criar o LD inicial e estabelecer o equilíbrio mutação-deriva em gerações históricas. O número de machos decaiu de 1.000 na primeira geração, para 500, na última geração histórica em iguais proporções entre os sexos. O sistema de acasalamento foi baseado na união aleatória dos gametas, sem que tenha havido migração ou sobreposição de gerações e amostrados aleatoriamente. O número de macho e fêmeas foi mantido constante em cada geração e as frequências alélicas foram fixadas em 0,5.

2.1.3 Expansão populacional

Foram selecionados aleatoriamente 1.000 indivíduos (500 machos fundadores e 500 fêmeas fundadoras) da última população histórica. A fim de ampliar a população, oito gerações foram simuladas atribuindo três progênies por fêmea, por geração; união aleatória de gametas; crescimento exponencial do número de fêmeas; e nenhuma seleção foi considerada nesta etapa.

2.1.4 População recente

Essa etapa foi dividida em duas partes. Na primeira parte, a população POP1 foi composta por 400 machos e 10.000 fêmeas oriundos da última geração (geração 8) da população de expansão onde foram selecionados e acasalados de maneira aleatória. Foi utilizada uma taxa de substituição de 60% dos machos e 20% das fêmeas com taxa de crescimento populacional nula considerando dois descendentes por fêmea por geração no decorrer de oito gerações. Em cada geração, os animais com melhores valores fenotípicos foram mantidos e descartados os animais considerados de piores valores fenotípicos. Os valores genéticos dos animais foram gerados pela metodologia *Best linear unbiased predictor* (BLUP), utilizando a metodologia de modelos mistos[14] para um modelo animal, considerando a verdadeira variância genética aditiva.

Na segunda parte, a definição da população recente, foram simuladas as populações denominadas POP2, que diferiram somente na quantidade de variância fenotípica. Em todos os cenários, os fundadores das populações POP2, foram obtidos a partir da última geração (geração 8) da população POP1. Esses animais constituíram rebanhos através de oito gerações, onde foram selecionados e descartados animais nos mesmos moldes da população POP1. Foi considerada uma taxa de substituição de 20% de machos e 5% das fêmeas. No total, foram simulados um total de 164.010 animais ao

longo de oito gerações na POP2. Após selecionar apenas os animais das três ultimas gerações, restaram 57.750 animais, a partir dos quais as diferentes análises foram realizadas. Foram simulados cenários cujos coeficientes de herdabilidade do QTL e da característica foram, respectivamente, 0,18 e 0,30.

2.1.5 Genoma

O genoma foi simulado com 26 pares de cromossomos autossômicos semelhantes aos da espécie *Ovis aires*, com tamanho idêntico ao genoma real de ovinos com base na plataforma AnimalQTLdB (https://www.animalgenome.org/cgi-bin/QTLdb/index) totalizando 2657 cM. A vantagem de simular o número real de cromossomos com comprimento idêntico ao genoma ovino é criar um cenário mais realista com relação ao número de *loci* de marcadores e QTLs fisicamente desvinculados[11].

Foram distribuídos 12 mil marcadores bialélicos de forma equidistante considerando a presença de 2014 QTLs bialélicos distribuídos através de pesos de acordo com o número de QTLs já conhecidos na literatura que tenham efeitos sobre a característica área de olho de lombo (AOL) e supondo no mínimo um QTL por cromossomo que não tenha relação conhecida na literatura com essa característica. Os efeitos alélicos aditivos destes QTLs foram simulados a partir de uma distribuição gama, com parâmetro forma $\alpha = 0{,}4$. Os marcadores genéticos e QTLs foram obtidos levando em conta uma MAF (*minor allele frequency*) de 0,05. Os locos segregam com dois, três ou quatro alelos cujas posições foram distribuídas aleatoriamente. A taxa de erro de genotipagem de marcadores foi de 0,005 e a taxa de genótipo de marcadores em falta foi de 0,01. Considerou-se uma taxa de mutação recorrente de 10^{-5} para marcadores e QTLs.

198

199 *2.2 Análises genético-quantitativas*

200 O modelo animal geral utilizado pode ser descrito como:

201 $$\boldsymbol{y}=\mathbf{X}\beta+\mathbf{Z}\alpha+\mathbf{e}$$

202 em que **y** é o vetor de observações para característica; β é o vetor de efeitos fixos (sexo
203 e geração); α é o vetor de efeitos genéticos aditivos diretos, em que $\alpha \sim N\ (0,\mathbf{G})$, com
204 $\mathbf{G}= \mathbf{A}\sigma_a^2$, na abordagem baseada no pedigree, e $\alpha \sim N\ (0,\mathbf{H}\sigma_a^2)$, em que σ_a^2 é a variância
205 genética aditiva, **A** é a matriz de numeradores dos coeficientes de parentesco e **H** é a
206 matriz de parentesco que combina informações de parentesco baseadas no pedigree e
207 em dados genômicos; **X** é a matriz de incidência dos efeitos fixos; **Z** é a matriz de
208 incidência dos efeitos genéticos aditivos; e $\boldsymbol{e}$ é o vetor de efeitos residuais, em que $\boldsymbol{e} \sim N$
209 $(0,\mathbf{R})$, com $\mathbf{R}= \mathbf{I}\sigma_e^2$.

210 Para a inclusão de informação genômica nos modelos com utilização do método
211 ssGBLUP, a inversa da matriz **A** ($\boldsymbol{A^{-1}}$) foi substituída pela inversa da matriz **H** (H^{-1}),
212 de modo que $\mathbf{H^{-1}}$ foi calculada de acordo com Aguilar et al.[3], como:

213

214 $$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

215 em que $\boldsymbol{A^{-1}}$ é a inversa da matriz de parentesco baseada na informação de *pedigree*,
216 $\boldsymbol{G^{-1}}$ é a inversa da matriz de parentesco genômico e $\mathbf{A}_{22}^{-1}$ é a inversa da matriz de
217 parentesco baseada no *pedigree* apenas para animais genotipados.

218 Os componentes de variância e valores genômicos foram obtidos com o auxílio de
219 programas da família BLUPF90[15]. As estimativas de componentes de variância obtidas
220 com uso do programa AIREMLF90 foram utilizadas como valores para predição dos

valores genômicos com uso do programa BLUPF90. Foram selecionados 57.770 animais das últimas três gerações, em um total de 164.010 animais de oito gerações simuladas. Destes animais selecionados foram retirados 5.000 indivíduos e, posteriormente, 1.000 animais destes para a formação das populações analisadas. Todas as seleções de animais foram feitas aleatoriamente.

As análises e ajustes aos dados foram feitas utilizando o *software* R versão 3.6.2[16]. Presumiu-se três níveis de variância fenotípica (5, 10 e 15), com dois tamanhos populacionais (1.000 e 5.000 animais). Para cada tamanho populacional, foram utilizadas informações de genótipo de 100, 50, 20 e 0% (este último corresponde ao BLUP tradicional) dos animais que foram avaliados com dados fenotípicos simulados seguindo distribuição normal e dados fenotípicos transformados para distribuição gama e para avaliar o comportamento do método ssGBLUP quando a variável resposta apresentou distribuição assimétrica. Combinando todos os cenários, 48 análises diferentes foram realizadas.

A estimativa de herdabilidade sugerida para a simulação da característica foi de 0,30 e do QTL foi de 0,18.

O controle de qualidade dos dados genômicos foi realizado com uso do software PREGSF90[15]. Foram removidos marcadores com MAF menor que 0,05, *Call Rate* menor que 0,95 e desvio extremo do equilíbrio de Hardy-Weinberg.

Para avaliar a sensibilidade do método a dados com distribuição assimétrica, aplicou-se transformação de distribuição da variável resposta simulada seguindo distribuição normal para distribuição gama, de acordo com a fórmula abaixo:

$x_{g=(x_n)^2}$, com os parâmetros $\alpha = \beta \,.\, \mu_g$ e $\beta = \frac{\mu_g}{(\sigma_g)^2}$,

onde x_n representa os dados fenotípicos simulados a partir de uma distribuição normal; x_g são os dados fenotípicos transformados para distribuição gama; μ_g é a média dos dados; σ_g é o desvio padrão dos dados; e $\alpha\ e\ \beta$ são os parâmetros da distribuição gama.

Para verificação da transformação dos dados, foi aplicado o teste de Kolmogorov-Smirnov no software R, utilizando a função *ks.test* com os dados e parâmetros mostrados na transformação acima. Além disso, foi observado o comportamento dos dados através de histogramas em todos os cenários.

A acurácia preditiva foi mensurada através da correlação de Pearson entre os valores genéticos estimados e os valores genéticos verdadeiros (TBVs) que foram obtidos com a soma dos efeitos poligênicos e do QTLs. Quanto mais próximo de 1o valor dessa correlação, melhor e mais acurado é o método.

3 Resultados e discussão

3.1 Estimativas de herdabilidade

As estimativas médias de herdabilidade (h^2) foram 0,22; 0,19 e 0,27, respectivamente, para os cenários com variâncias fenotípicas de 5, 10 e 15. No geral, as herdabilidades tenderam a crescer quando a os dados seguem distribuição normal dentro de cada variância (Tabela 1). No entanto, ao diminuir o tamanho amostral de 5.000 para 1.000, as herdabilidades sofreram maiores variações com tendência a serem menores na distribuição normal.

Em relação ao número de animais genotipados, não foram observadas tendências de grandes variações para h^2. Provavelmente, as variações não foram notadas, porque todos os componentes de variância que formam a herdabilidade são afetados da mesma forma e intensidade com mudanças na distribuição dos dados. Os valores de σ_a^2 e σ_e^2, ao

fixar o tamanho amostral, dentro da distribuição normal, não sofreram grandes variações e tenderam a aumentar com o aumento da variância fenotípica, como pode ser observado na Tabela 1. Porém, dentro da distribuição gama, além das estimativas para σ_a^2 e σ_e^2 serem instáveis, foram discrepantes, quando comparadas àquelas obtidas quando os dados apresentaram distribuição normal, com aumento em mais de 500%, dependendo do cenário.

Verificou-se que, com o aumento do número de animais genotipados, os desvios de σ_a^2 e σ_e^2 diminuíram, sendo que os cenários de maiores desvio foram aqueles que corresponderam a 0% de animais genotipados (análise correspondente ao BLUP tradicional). É importante lembrar que, desvio padrão maior significa menor confiança na estimativa. De fato, tal resultado era esperado, pois as informações genômicas avaliam diretamente cada segregação em nível individual e não médio como ocorre com o BLUP tradicional[17]. Sendo assim, os métodos genômicos são geralmente mais acurados.

Na literatura, existem variações entre os valores de herdabilidade para a característica a área de olho de lombo (AOL), diferenciado-se pelos métodos, raças e modelos. Os valores de herdabilidade obtidos nesta pesquisa estão, em sua maioria, entre os valores de 0,20 e 0,31, relatados por Sena et al.[18] e Figueiredo Filho et al.[19] respectivamente, em ovinos Santa Inês. Em raças ovinas, valores de herdabilidade de 0,19 e 0,07 para AOL, foram relatados por Ciappesoni et al.[20] e Kiya et al.[21], respectivamente. As estimativas de h² podem variar devido a propriedades específicas das populações e estrutura dos dados. Além disso, seus componentes são alterados quando se quebra a pressuposição dos modelos.

3.2 Valores genéticos e acurácia preditiva

As médias dos (GEBVs), mantiveram-se em torno de zero independente da distribuição, tamanho amostral e número de animais genotipados. Entretanto, os seus erros padrão (EP) aumentaram de 10 até 23 vezes quando se alterou a distribuição dos dados fenotípicos para gama nos cenários com variâncias fenotípicas 5 e 15, respectivamente (Tabela 2).

Vale ressaltar, que foi observada tendência de aumento do EP quando se diminuiu o número de animais genotipados na amostra. Dessa forma, os GEBVs preditos a partir de dados com distribuição gama são de menor confiança.

De modo geral, a acurácia entre os valores genéticos estimados e verdadeiros (simulados), aumentou quando se elevou o tamanho amostral e o número de animais genotipados, independente da distribuição dos dados. Entretanto, as acurácias dos dados que seguem distribuição normal, foram sempre maiores ao serem comparadas nos mesmos cenários (tamanho amostral e número de animais genotipados) das obtidas quando os dados seguiam distribuição gama e foram analisados como normais.

Os desvios entre os TBVs e o GEBVs variaram em até 0,977, quando se fixou a distribuição normal, com tendência de aumento quando se diminuiu o tamanho amostral e o número de animais genotipados (Tabela 3). Destaca-se também que, quando se aumentou a variância fenotípica esses desvios aumentaram. Porém, as maiores discrepâncias nos valores desses desvios ocorreram quando se alterou a distribuição dos dados de normal para gama, aumentando em cerca de até 13 vezes.

De acordo com os dados apresentados na Tabela 3, observou-se que fixando o tamanho amostral em 5.000 e a distribuição normal, ocorreram variações da acurácia entre 1 e 4% considerando as três variâncias fenotípicas. Alterando apenas o tamanho amostral para 1.000, essa variação ficou entre 1 e 9,2%. Tal resultado já era esperado,

visto que quanto maior o tamanho amostral e o número de animais genotipados maior será o número de informações incluído nas análises e menor será a disparidade dos valores genéticos estimados e verdadeiros.

Modificando a distribuição dos dados fenotípicos de normal para gama, dentro de um mesmo tamanho amostral de 5.000 animais, percorrendo todas as situações de número de animais genotipados, a perda de acurácia chegou a até 8,4%, entre o cenário de maior e menor número de animais genotipados, e até 14,7%, quando o tamanho amostral teve 1.000 animais. As maiores variações ocorreram com o uso variância fenotípica 5. Ao considerar as variâncias fenotípicas (10 e 15), as perdas de acurácia ficaram em torno de 1,3% até 7,3%, quando o tamanho amostral foi de 5.000 animais e de 1,4% até 9,4%, quando o tamanho amostral foi de 1.000 animais, considerando todas as situações de número de animais genotipados.

O comportamento dos resultados encontrados neste estudo está de acordo com a literatura. Por exemplo, com Hayes et al.[2], afirmaram que o uso de maior quantidade de informações fenotípicas e genotípicas resultará em mais observações por alelo de SNP e, consequentemente, maiores acurácias na predição genômica. Os ganhos com a predição genômica estão dentro dos relatados por Daetwyler et al.[22], que ao avaliarem características de carcaça em ovinos, observaram ganho médio no valor da acurácia real de 0,20 (0,07 a 0,31), na predição de valor genético com uso de informação genômica, e acurácia de predição média de 0,09, com uso do BLUP baseado apenas no pedigree.

Em relação à quebra de pressuposição de normalidade dos dados, ao mudar a distribuição dos dados para gama, estes apresentaram distribuição com comportamento assimétrico. Comparando cenários mais específicos considerando mesmo tamanho amostral e número de animais genotipados, variando somente a distribuição, em todas

as situações as melhores acurácias foram obtidas quando a os dados fenotípicos seguiram distribuição normal.

4 Conclusão

Ao quebrar a pressuposição de normalidade, a capacidade preditiva do método ssGBLUP diminuiu em até 8,4 e 14,7%, principalmente em pequenas populações. Os desvios entre os valores genéticos verdadeiros e preditos variaram aumentando em até 13 vezes. Quando comparados os cenários de maior e menor número de animais genotipados, foi verificado o ganho de acurácia de até 9,2%. Quanto maior a variância fenotípica, maiores foram os erros padrão dos GEBVs. A distância entre os valores genéticos verdadeiros e preditos aumentaram ao combinar pequenos tamanhos amostrais, baixo número de animais genotipados e o não atendimento da normalidade na distribuição dos dados fenotípicos.

Declaração de competição de interesses

Os autores declaram que não têm competição de interesses financeiros ou relações pessoais que possam influenciar este trabalho.

Agradecimentos

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES – Registro número: 001).

Referências

1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819-1829. doi:10.1093/genetics/157.4.1819
2. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*. 2009;92(2):433-443. doi:10.3168/jds.2008-1646
3. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci*. 2010;93(2):743-752. https://doi.org/10.3168/jds.2009-2730
4. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*. 2009;92(9):4656-4663. https://doi.org/10.3168/jds.2009-2061
5. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. *Genet Sel Evol*. 2010;42(1):2. https://doi.org/10.1186/1297-9686-42-2
6. Tsuruta S, Misztal I, Aguilar I, Lawlor TJ. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J Dairy Sci. 2011;94(8):4198-4204. https://doi.org/10.3168/jds.2011-4256
7. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. *Animal*. 2012;6(10):1565-1571. doi:10.1017/S1751731112000742
8. Tonussi RL, Silva O, Magalhães AFB, Espigolan R, Peripolli E, Olivieri BF, Feitosa FL, Antunes Lemos MV, Berton MP, Justino Chiaia HL, Cravo Pereira AS, Lôbo R

B, Framartino Bezerra LA, Magnabosco U, Lino Lourenço DA, Aguilar I, Baldi F. A pplication of single step genomic BLUP under different uncertain paternity scenarios using simulated data. *PLoS ONE*. 2017;12(9): e0181752. https://doi.org/10.1371/jou rnal.pone.0181752

9. Nascimento M, Silva FF, Resende MDV, Cruz CD, Nascimento ACC, Viana JMS, Azevedo CF, Barroso LMA. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. *Genet Mol Res.* 2017;16(1):1-12. https://doi.org/10.4238/gmr16019538
10. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*. 2009;25(5):680-681. https://doi.org/10.1093/bioinformatics/btp045
11. Brito FV, Braccini Neto J, Sargolzaei M, Cobuci JA, Schenkel FS. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genetics.* 2011;12(80): 1-10. https://doi.org/10.1186/1471-2156-12-80
12. Pértile SFN, Silva FF, Salvian M, Mourão GB. Seleção e associação genômica ampla para o melhoramento genético animal com uso do método ssGBLUP. *Pesq Agropec Bras*. 2016;51(10):1729-1736. https://doi.org/10.1590/S0100-204X2016001000004
13. Calafell F, Grigorenko EL, Chikanian AA, Kidd KK. Haplotype evolution and linkage disequilibrium: A simulation study. *Hum Hered*. 2001;51(1-2):85-96. https://doi.org/10.1159/000022963
14. Henderson CR. *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph; 1984.
15. Misztal I, Tsuruta S, Lourenco DAL, Masuda Y, Aguilar I, Legarra A, Vitezica Z. *Manual for BLUPF90 family of programs*. Available from

http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf. 2018. Accessed May 20, 2019.

16. R Core Team R. *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
17. Resende MDV, Lopes PS, Silva RL, Pires IE. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesq Flor Bras*. 2008;56: 63-78.
18. Sena LS, Santos GV, Torres TS, Sousa Júnior A, Rego Neto AA, Sarmento JLR, Biagiotti D. Genetic parameters for carcass traits and body size of meat sheep. *Semina Ciênc Agrár*. 2016;37:2477-2486. http://dx.doi.org/10.5433/1679-0359.2016v37n4Supl1p2477
19. Figueiredo Filho LAS, do Ó AO, Sarmento JLR, Santos NPS, Torres TS. Genetic parameters for carcass traits and body size in sheep for meat production. *Trop Anim Health Prod*. 2016;48(1):215–218. https://doi.org/10.1007/s11250-015-0921-5
20. Ciappesoni G, San Julián R, Navajas EA, Gimeno D, Gutierrez-Zamit E, Goldberg V, Brito G. *Genetic evaluation of the Texel breed in Uruguay: I. Carcass quality traits*. In: 60th Internat. Cong Meat Sci Technol. Punta del Este, Uruguay. 2014. https://doi.org/10.13140/2.1.2762.6888
21. Kiya CK, Pedrosa VB, Muniz KFA, Gusmão AL, Pinto LFB. Estimates of the genetic parameters of a Dorper flock in Brazil. *Small Rumin Res*. 2019;171:57-62. https://doi.org/10.1016/j.smallrumres.2018.12.007
22. Daetwyler HD, Swan AA, van der Werf JH, Hayes BJ. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genet Sel Evol*. 2012;44(1):33. https://doi.org/10.1186/1297-9686-44-33

Tabela 1. Estimativas de componentes de variância e herdabilidade para diferentes distribuições, tamanhos amostrais e variâncias fenotípicas.

σ_p^2	Distribuição	N	Genotipados	σ_a^2 ± DP	σ_e^2 ± DP	h²
5	Normal	5.000	5.000	1,01 ± 0,11	3,92 ± 0,11	0,204
			2.500	1,07 ± 0,14	3,86 ± 0,14	0,217
			1.000	1,03 ± 0,16	3,89 ± 0,16	0,210
			0	1,09 ± 0,17	3,83 ± 0,17	0,221
	Gama	5.000	5.000	142,04 ± 16,83	636,66 ± 18,03	0,182
			2.500	141,73 ± 21,25	638,14 ± 22,17	0,181
			1.000	135,42 ± 23,15	643,19 ± 24,41	0,173
			0	140,72 ± 24,29	638,00 ± 25,41	0,181
	Normal	1.000	1.000	0,69 ± 0,36	4,05 ± 0,38	0,145
			500	0,70 ± 0,50	4,04 ± 0,51	0,148
			200	1,24 ± 0,61	3,50 ± 0,61	0,261
			0	1,27 ± 0,61	3,46 ± 0,61	0,269
	Gama	1.000	1.000	117,72 ± 53,51	579,21 ± 56,14	0,169
			500	163,19 ± 73,80	533,79 ± 74,02	0,231
			200	255,02 ± 89,39	440,74 ± 86,95	0,366
			0	261,29 ± 89,70	434,39 ± 87,07	0,375
10	Normal	5.000	5.000	2,23 ± 0,23	7,87 ± 0,23	0,221
			2.500	2,46 ± 0,31	7,65 ± 0,30	0,243
			1.000	2,67 ± 0,37	7,43 ± 0,35	0,264
			0	2,75 ± 0,39	7,35 ± 0,37	0,272
	Gama	5.000	5.000	677,48 ± 75,38	2616,5 ± 76,35	0,205
			2.500	688,74 ± 95,28	2606,6 ± 95,41	0,209
			1.000	755,13 ± 111,95	2539,6 ± 110,39	0,229
			0	761,58 ± 116,8	2533,5 ± 115,48	0,231
	Normal	1.000	1.000	1,74 ± 0,81	8,54 ± 0,84	0,170
			500	1,82 ± 1,11	8,46 ± 1,13	0,177
			200	1,38 ± 1,29	8,90 ± 1,33	0,134
			0	1,27 ± 1,29	9,01 ± 1,33	0,123
	Gama	1.000	1.000	511 ± 260,69	2840,0 ± 260,69	0,152
			500	528,97 ± 355,55	2822,9 ± 365,79	0,134
			200	580,07 ± 413,99	3071,3 ± 431,36	0,160
			0	550,29 ± 412,90	3101,1 ± 431,35	0,150
15	Normal	5.000	5.000	3,54 ± 0,35	11,58 ± 0,34	0,234
			2.500	4,49 ± 0,49	10,70 ± 0,44	0,295
			1.000	4,16 ± 0,54	11,01 ± 0,52	0,274
			0	4,37 ± 0,58	10,87 ± 0,55	0,287
	Gama	5.000	5.000	1553,0 ± 159,77	5501,9 ± 159,96	0,220
			2.500	1908,6 ± 218,64	5171,30 ± 205,15	0,269
			1.000	1792,1 ± 240,59	5283,6 ± 233,73	0,253
			0	1832,8 ± 253,77	5248,7 ± 243,45	0,258
	Normal	1.000	1.000	3,55 ± 1,26	12,06 ± 1,26	0,227
			500	3,64 ± 1,64	11,96 ± 1,64	0,233
			200	4,09 ± 1,93	10,694 ± 1,89	0,276
			0	4,26 ± 2,01	11,34 ± 2,00	0,273
	Gama	1.000	1.000	2029,4 ± 619,89	5429,1 ± 603,82	0,272
			500	1943,5 ± 806,96	5506,7 ± 799,69	0,260
			200	2857,5 ± 946,65	4578,5 ± 912,17	0,273
			0	2460 ± 990,07	4985,7 ± 969,23	0,3303

N: Tamanho amostral; σ_a^2: variância genética aditiva; σ_e^2: variância ambiental; σ_p^2: variância fenotípica; h²: herdabilidade; DP: desvio padrão.

Tabela 2. Acurácias, GEBVs e seus erros padrão para diferentes distribuições, tamanhos amostrais e variâncias fenotípicas.

σ_p^2	Distribuição	N	Genotipados	Acurácia	Média GEBV	EP GEBV
5	Normal	5.000	5000	0,606	-0,004	0,862
			2500	0,592	-0,004	0,914
			1000	0,583	0,011	0,915
			0	0,581	0,015	0,935
	Gama	5.000	5000	0,586	-0,060	10,368
			2500	0,571	0,095	10,688
			1000	0,557	0,155	10,691
			0	0,555	0,200	10,822
	Normal	1.000	1000	0,474	-0,001	0,781
			500	0,469	-0,000	0,795
			200	0,464	0,002	1,020
			0	0,430	0,003	1,025
	Gama	1.000	1000	0,440	0,200	10,115
			500	0,435	0,009	11,710
			200	0,430	0,050	14,060
			0	0,404	0,057	14,500
10	Normal	5.000	5000	0,594	-0,001	1,267
			2500	0,586	0,004	1,366
			1000	0,577	0,009	1,429
			0	0,573	0,007	1,450
	Gama	5.000	5000	0,589	-0,051	21,054
			2500	0,570	0,076	23,230
			1000	0,561	0,196	24,410
			0	0,557	0,163	24,580
	Normal	1.000	1000	0,509	-0,000	1,233
			500	0,496	0,001	1,266
			200	0,470	-0,000	1,126
			0	0,468	1,110	1,080
	Gama	1.000	1000	0,500	-0,009	21,520
			500	0,491	0,033	21,700
			200	0,462	0,005	16,290
			0	0,461	0,008	15,420
15	Normal	5.000	5000	0,609	-0,003	1,589
			2500	0,596	0,010	1,798
			1000	0,586	0,120	1,775
			0	0,583	0,012	1,813
	Gama	5.000	5000	0,594	-0,077	33,530
			2500	0,580	0,269	37,540
			1000	0,567	0,305	37,180
			0	0,564	0,272	37,650
	Normal	1.000	1000	0,558	0,000	1,717
			500	0,550	0,000	1,750
			200	0,545	-0,001	1,973
			0	0,540	-0,000	1,871
	Gama	1.000	1000	0,550	0,018	40,280
			500	0,539	0,028	40,000
			200	0,535	-0,013	46,290
			0	0,532	0,000	43,950

GEBV: valores genéticos genômicos; N: tamanho amostral; EP: erro padrão; σ_p^2: variância fenotípica.

Tabela 3. Variações das acurácias e desvios entre TBVs e GEBVs para diferentes distribuições, tamanhos amostrais e variâncias fenotípicas.

σ_p^2	Distribuição	N	Genotipados	Acurácia	Perda de acurácia	Desvio entre TBVs e GEBVs
5	Normal	5.000	5000	0,606	0,000	1,056
			2500	0,592	-0,023	1,073
			1000	0,583	-0,037	1,086
			0	0,581	-0,041	1,084
	Gama	5.000	5000	0,586	-0,033	6,440
			2500	0,571	-0,048	5,683
			1000	0,557	-0,080	5,669
			0	0,555	-0,084	5,365
	Normal	1.000	1000	0,474	0,000	1,214
			500	0,469	-0,010	1,219
			200	0,464	-0,021	1,195
			0	0,430	-0,092	1,186
	Gama	1.000	1000	0,440	-0,071	4,404
			500	0,435	-0,082	5,799
			200	0,430	-0,092	8,484
			0	0,404	-0,147	9,251
10	Normal	5.000	5000	0,594	0,000	1,494
			2500	0,586	-0,013	1,505
			1000	0,577	-0,028	1,512
			0	0,573	-0,035	1,515
	Gama	5.000	5000	0,589	-0,008	19,135
			2500	0,570	-0,040	13,637
			1000	0,561	-0,055	14,168
			0	0,557	-0,062	14,074
	Normal	1.000	1000	0,509	0,000	1,698
			500	0,496	-0,025	1,699
			200	0,470	-0,076	1,752
			0	0,468	-0,080	1,756
	Gama	1.000	1000	0,500	-0,017	8,746
			500	0,491	-0,035	8,545
			200	0,462	-0,092	4,084
			0	0,461	-0,094	3,832
15	Normal	5.000	5000	0,609	0,000	1,796
			2500	0,596	-0,021	1,811
			1000	0,586	-0,023	1,832
			0	0,583	-0,042	1,834
	Gama	5.000	5000	0,594	-0,024	24,098
			2500	0,580	-0,047	25,532
			1000	0,567	-0,068	23,335
			0	0,564	-0,073	23,406
	Normal	1.000	1000	0,558	0,000	2,020
			500	0,550	-0,014	2,033
			200	0,545	-0,023	1,989
			0	0,540	-0,032	2,005
	Gama	1.000	1000	0,550	-0,014	23,338
			500	0,539	-0,034	21,576
			200	0,535	-0,041	31,602
			0	0,532	-0,046	26,994

GEBV: valores genéticos estimados; TBV: valor genético verdadeiro; σ_p^2: variância fenotípica; N: Tamanho amostral.

CONSIDERAÇÕES FINAIS

Embora inúmeros métodos estatísticos lidem com vários desafios na implementação da seleção genômica, o problema da assimetria da distribuição de valores fenotípico não é comumente considerados. Os métodos avaliados no trabalho diminuem seu ganho genético quando a pressuposição de normalidade não é atendida.

A baixa densidade do painel de SNPs simulado neste estudo, devido às dificuldades computacionais exigidas, pode ter contribuído para a diminuição da capacidade preditiva dos métodos. No entanto, os resultados desta pesquisa mostraram-se satisfatórios e servirão para nortear e auxiliar novos estudos.

Novas pesquisas sobre o tema são recomendadas para o desenvolvimento de novos métodos ou aperfeiçoamento daqueles já existentes, para superar o problema abordado. Para isto, é importante o incremento na densidade dos painéis de SNPs e pesquisas com dados reais, contribuindo assim para o melhoramento genético de ovinos e outras espécies de interesse econômico.