UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS

FERNANDA SCHNEBERGER DOS SANTOS

Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos – Um estudo de simulação

Pirassununga
2022

FERNANDA SCHNEBERGER DOS SANTOS

Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos – Um estudo de simulação

Versão Corrigida

Dissertação apresentada à Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Ciências do programa de pós-gradução em Zootecnia.

Área de Concentração: Qualidade e Produtividade Animal

Orientador: Prof. Dr. Joanir Pereira Eler

Coorientador: Dr. Rafael Espigolan

Pirassununga

2022

Ficha catalográfica elaborada pelo
Serviço de Biblioteca e Informação, FZEA/USP,
com os dados fornecidos pelo(a) autor(a)

S237e

Schneberger dos Santos, Fernanda
Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos ? Um estudo de simulação / Fernanda Schneberger dos Santos ; orientador Joanir Pereira Eler ; coorientador Rafael Espigolan. -- Pirassununga, 2022.
61 f.

Dissertação (Mestrado - Programa de Pós-Graduação em Zootecnia) -- Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo.

1. Acurácia. 2. Bovinos. 3. Genômica. 4. Erros genotípicos. 5. Simulação. I. Pereira Eler, Joanir , orient. II. Espigolan, Rafael, coorient. III. Título.

Permitida a cópia total ou parcial deste documento, desde que citada a fonte - o autor

UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos
Comissão de Ética no Uso de Animais

DISPENSA DE ANÁLISE ÉTICA

Comunicamos que o projeto de pesquisa abaixo identificado **está dispensado da análise** ética pela *Comissão de Ética no Uso de Animais* (CEUA), da *Faculdade de Zootecnia e Engenharia de Alimentos* da USP, por utilizar animais oriundos de coleções biológicas formadas anteriormente ao ano de 2008, ano da promulgação da Lei nº 11.794/2008 a qual estabelece procedimentos para o uso científico de animais.

Ressaltamos que atividades realizadas na vigência da referida lei ou que resulte em incremento do acervo biológico, devem ser submetidos à análise desta CEUA conforme disposto pelo *Conselho Nacional de Controle de Experimentação Animal* (CONCEA).

Projeto: Efeitos de erros de imputação na estimação de valores genéticos genômicos utilizando dados simulados de bovinos de corte
Pesquisador responsável: Fernanda Schneberger dos Santos (Orientador: Prof. Dr. Joanir Pereira Eler)

Pirassununga, 26 de outubro de 2020.

Profa. Dra. CRISTIANE GONÇALVES TITTO
Presidente da Comissão de Ética no Uso de Animais

Profa. Dra. DANIELE DOS SANTOS MARTINS
Vice-Presidente da Comissão de Ética no Uso de Animais

Av. Duque de Caxias Norte, 225 - 13635-900 – Pirassununga, SP
Fone: (19) 3565-6759 -- E-mail: cepfzea@usp.br

FERNANDA SCHNEBERGER DOS SANTOS

Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos – Um estudo de simulação

Dissertação apresentada à Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Ciências do programa de pós-gradução em Zootecnia.

Data da aprovação: 03 de fevereiro de 2022

Banca examinadora:

Prof. Dr. Joanir Pereira Eler
Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo
Presidente da Banca Examinadora

Profa. Dra. Rachel Santos Bueno Carvalho
Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo

Dr. Fernando de Oliveira Bussiman
Faculdade de Medicina Veterinária e Zootecnia – Universidade de São Paulo

Prof. Dr. José Bento Sterman Ferraz
Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo

Prof. Dra. Bárbara da Conceição Abreu Silva
Universidade Federal de Sergipe

DEDICATÓRIA

À todos os meus professores, sem seus esforços e virtudes, eu não seria eu.

À minha mãe, por sua integridade e dedicação, e por seu amor.

Aos meus irmãos, com quem sempre posso contar.

Ao meu pai, por me ensinar sobre perdão.

Dedico.

AGRADECIMENTOS

Ao Professor Doutor Joanir Pereira Eler, pela oportunidade, a orientação e a paciência. É difícil expressar o tamanho da minha admiração. Como todos os seus alunos, fui inspirada a buscar o conhecimento que se converte em dedicação e trabalho árduo, assim como em alcançar a humildade que não torna meu orgulho maior do que minha própria sombra. Você foi o primeiro a acreditar que eu seria capaz. Obrigada!

Ao meu Coorientador Doutor Rafael Espigolan, pelo exemplo, paciência e amizade. Desde o início andei à sombra do teu conhecimento, quando alcancei meu lugar ao sol, percebi que tive mais do que merecia naquela sombra e nunca serei grata o suficiente. Também pude aprender que algumas pessoas dão boas aulas, mas poucos é que são bons professores.

Aos professores Dra. Rachel Santos Bueno Carvalho, Dr. José Bento Sterman Ferraz e Dr. César Gonçalves de Lima, pelo apoio e as conversas. À Ms. Elisângela Chicaroni de Mattos Oliveira pela contribuição ao desenvolvimento de todos nós, sempre de forma sensata e acalentadora e por incontáveis momentos de alegria e amizade.

Aos professores Dr. Fernando Sebastián Baldi Rey, Dr. Júlio César de Carvalho Balieiro e Dr. Roberto Carvalheiro pelas sugestões e correções durante o exame de qualificação de mestrado. Conhecimento é acumulativo e sempre colaborativo.

Aos meus amigos do Laboratório de Melhoramento Animal e Biotecnologia “Dr. Gordon Dickerson”, Bárbara Abreu, Fernando Bussiman, Alisson Acero (Chiquita), Caroline Assis, Felipe Eguti, Hugo Borges e Fernando Cançado. Vocês são a minha sorte. Parece-me cedo para agradecer, mas sempre tarde para me desculpar. Somos jovens em uma meritocracia que não preza por solidariedade, com uma enorme facilidade para desprezar os menos afortunados, e que religiosamente recompensa “vencedores” ou castiga “perdedores”. A universidade precisa ensinar-nos que a força do progresso da ciência é sua capacidade de autocorreção, não importa o quanto você for inteligente, o erro faz parte da equação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos de mestrado.

À Universidade de São Paulo e a Faculdade de Zootecnia e Engenharia de Alimentos pela estrutura para alcançar este título.

Aos membros do Serviço de Pós-Graduação, Érica Cristina Mello e Maria Cecilia Albernaz.

Ao programa de Pós-Graduação em Zootecnia, coordenado pela Profa. Dra. Cristiane Gonçalves Titto.

À “Tia Vera” do GMA por seu zelo e incansável determinação. A limpeza é uma arte subestimada e apenas compreendida quando poeira e pó dominam cadeiras, mesas e até pensamentos.

“To be a rock and not to roll.”

Led Zeppelin

LISTA DE ILUSTRAÇÕES

Figura 1 – Coeficiente de regressão do TBV sobre o GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)……...…….....48

Figura 2 – Coeficiente de regressão do TBV sobre o GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)……48

Figura 3 – Coeficiente de correlação de Spearman entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)……...50

Figura 4 – Coeficiente de correlação de Spearman entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)……50

Figura 5 – Valor médio do GEBV para os animais genotipados (GE), em todos os cenários de taxas de erros genotípicos……..51

Figura 6 – Valor médio do GEBV para os 100 animais superiores genotipados (100GE), em todos os cenários de taxas de erros genotípicos……..52

Figura 7 – Coeficientes de correlação de Pearson entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)……...53

Figura 8 – Coeficientes de correlação de Pearson entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)……53

Figura 9 – Acurácia de predição BIF em todos os cenários de taxas de erros genotípicos……...54

Figura 10 – Variância do erro de predição, em todos os cenários de taxas de erros genotípicos.....55

Figura 11 – Regressão linear das taxas de erros genotípicos sobre acurácia de predição BIF.....55

Figura 12 – Desvio padrão (σ) dos elementos fora da diagonal da matriz genômica, em todos os cenários de taxas de erros genotípicos.....56

RESUMO

SANTOS, F. S. **Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos** – Um estudo de simulação. 2021. 60 f. Dissertação (Mestrado) – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2021.

A seleção genômica (SG) e a adoção de marcadores moleculares do tipo SNP (*single nucleotide polymorphism*) aumentaram a eficiência de muitos programas de melhoramento genético ao redor do mundo. Com a utilização dessas novas ferramentas, aumentou-se a acurácia de predição dos animais jovens e sem progênie. A aplicação da informação genômica requer inúmeros genótipos para formação de grupos de referência e avaliação de candidatos, aumentando os custos financeiros. Considerando os erros de genotipagem, suas origens são pouco claras, ou inevitáveis, com taxas muitas vezes desconhecidas e, com o crescimento dos bancos de dados genômicos e a grande diversidade de painéis comercializados, ocorre o agravamento desse problema. Em especial, a adoção da imputação dos genótipos para maior densidade consiste em um dos principais fatores de ocorrência de erros genotípicos e inconsistências mendelianas em bancos de dados. Portanto, os objetivos deste trabalho foram: simular taxas de erros genotípicos em dados simulados de bovinos de corte que mimetizem erros de imputação; avaliar o efeito da presença de erros genotípicos na acurácia de predição genômica; e identificar o viés na performance das avaliações genéticas em cenários de taxas crescentes de erros genotípicos. Os resultados mostraram a existência de efeitos devido aos erros genotípicos, com uma tendência que sugere desfavorecimento dos animais geneticamente superiores. A acurácia de predição, medida como correlação entre GEBV (*Genomic Estimated Breeding Value*) e TBV (*True Breeding Value*), diminuiu em resposta a taxas crescentes de erros genotípicos, sendo esse efeito mais intenso nos 100 animais superiores genotipados. O viés obtido por regressão do TBV sobre o GEBV não apresentou alteração na comparação com todos os animais, entretanto observou-se que o GEBV de animais superiores foram mais viesados ($b < 1$) quanto maior a taxa de erros. A utilização de genótipos com altas taxas de erros pode gerar ranqueamento ilegítimo tanto para animais superiores como também inferiores.

Palavras-chave: Acurácia. Bovinos. Genômica. Erros genotípicos. ssGBLUP.

ABSTRACT

SANTOS, F. S. **Effects of genotyping errors on the estimation of genomic breeding values –** A simulation study. 2021. 60 f. M.Sc. Dissertation – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2021.

Genomic selection (GS) and the adoption of SNP (Single Nucleotide Polymorphism)-type molecular markers have increased the efficiency of many breeding programs around the world. With the method, the prediction accuracy of young animals and those without progeny was increased. The application of genomic information requires numerous genotypes to form reference groups and evaluate candidates, increasing financial costs. Considering genotyping errors, their origins are unclear, or unavoidable, with rates often unknown. The growth of data collections and the great diversity of panels aggravate this problem, in particular, the adoption of the imputation of genotypes for higher density, which are the main factors in the occurrence of genotypic errors and Mendelian inconsistencies in databases. The objectives of this work are: to simulate genotypic error rates in simulated beef cattle data that mimic imputation errors; evaluate the effect of the presence of genotypic errors on the accuracy of genomic prediction; and identify bias in the performance of genetic evaluations in scenarios of increasing genotypic error rates. The results showed the existence of effects due to genotypic errors, with a trend that suggests the disadvantage of genetically superior animals. The prediction accuracy, measured as the correlation between GEBV and TBV, showed a decrease in response to increasing rates of genotypic errors, with this effect being more intense in the 100 superior animals genotyped. The bias obtained by regression of TBV on GEBV did not change in comparison with all animals. However, it was observed that the GEBV of superior animals was more biased ($b < 1$) the higher the error rate. The use of genotypes with high error rates can lead to erroneous ranking for both superior and inferior animals.

Keywords: Accuracy. Cattle. Genomic. Genotypic erros. ssGBLUP.

SUMÁRIO

1 INTRODUÇÃO....13
Hipótese....15
Objetivos....15
1.1.1 Objetivos gerais....15
1.1.2 Objetivos específicos....166
Referências....16
2 REVISÃO DE LITERATURA....19
Genética quantitativa....19
Seleção tradicional....21
Seleção genômica....23
2.1.1 Marcadores moleculares....23
2.1.2 Erros de genotipagem e imputação....24
2.1.3 GBLUP (Genomic BLUP)....28
2.1.4 Matriz de parentesco genômico (Matriz “G”)....29
2.1.5 ssGBLUP (Single-Step Genomic BLUP)....30
2.1.6 Matriz “H”....31
Referências....32
3 MATERIAL E MÉTODOS....40
Simulação dos fenótipos....40
Simulação dos genótipos....41
Controle de qualidade dos genótipos....41
Simulação dos erros genotípicos....42
Análises estatísticas....42
4 RESULTADOS E DISCUSSÃO....46
5 CONCLUSÃO....61
Referências....62

1 INTRODUÇÃO

O Brasil detém a posição comercial de um dos maiores produtores e exportadores de alimentos do mundo (FAO, 2021). O rebanho efetivo de bovinos brasileiro contém cerca de 218,2 milhões de cabeças (IBGE, 2020), após crescimento de 1,5% em relação ao ano anterior. Os animais zebuínos compõem cerca de 80% do rebanho efetivo (FERRAZ; ELER, 2010), incluindo seus cruzamentos com raças taurinas. Apresentam grande desuniformidade em desempenho reprodutivo, crescimento e produção, e estão dispersos nas cinco regiões do país, sob sistemas de criação extensivos e intensivos.

Ao longo dos anos, a principal estratégia do sistema de produção pecuário deixou de ser o aumento das áreas de pastejo em prol de medidas que aumentam produtividade por área ou unidade animal (MALAFAIA et al., 2021). O modelo de produção mais recente prioriza tecnologias e biotecnologias que demandam mais investimento em capital, para fazer mais no mesmo espaço, ou seja, tecnologias para “economizar terra”, com maiores rentabilidades e desempenhos técnicos (MARTA JÚNIOR; ALVES; CONTINI, 2011).

Para ganhos ainda maiores seria necessário melhor alinhamento de toda a cadeia produtiva, de mercado interno e/ou externo, que permita definir claros objetivos, padrões e protocolos a serem exigidos na produção, quanto à qualidade da carne, bem-estar animal, sustentabilidade e conservação ambiental, e melhoramento genético (KANASHIRO; FRAISSE, 2015).

Animais geneticamente superiores para ganho de peso e precocidade sexual, por exemplo, proporcionam o melhor aproveitamento produtivo dos recursos disponíveis, sendo fator condicionante para o resultado do uso de técnicas produtivas mais modernas, respondendo de forma mais eficiente e atingindo precocemente o peso ideal ao abate (FERRAZ; ELER, 1999). A produtividade da pecuária é influenciada tanto pela genética do animal quanto pelo ambiente de criação, por isso, o melhoramento genético prioriza obter maior desempenho animal ao menor custo, principalmente nutricional.

As principais ferramentas do melhoramento genético são os métodos de seleção e estratégias de acasalamento, que nas últimas décadas passaram por grandes e valiosas transformações, saindo de uma seleção visual, como formato de orelha, inserção de cauda e formato de cabeça, para uma seleção visando

características economicamente viáveis, tais como precocidade sexual, ganho de peso, acabamento de carcaça, conversão alimentar, e resistência à parasitas (BARROS et al., 2006; SILVA et al., 2016).

O método tradicional (BLUP e Modelo Animal) reúne dados de desempenho produtivo e reprodutivo do animal, grau de parentesco calculado via pedigree e pressupõe o conhecimento das variâncias e co-variâncias populacionais (SCHAEFFER, 2010). Desta forma permite ajustar as soluções para as variações genéticas entre grupos contemporâneos e até entre diferentes rebanhos (HENDERSON; QUAAS, 1976).

Com as descobertas e avanços de marcadores moleculares, desenvolveu-se a seleção genômica (SG) (MEUWISSEN; HAYES; GODDARD, 2001). Este método propõe a seleção com base nos valores genômicos (GEBV - *Genomic Estimated Breeding Values*) preditos a partir de informações fenotípicas e de marcadores genéticos, do tipo polimorfismo de base única (SNP - *Single Nucleotide Polymorphisms*), abrangendo grande parte do genoma. A principal mudança é que a matriz de relacionamentos genômicos pode ser considerada como um estimador “melhorado” de parentescos baseado em informações genotípicas observadas ao invés das estimações via pedigree (LEGARRA et al., 2014).

Consideráveis benefícios da SG são vistos na avaliação de fêmeas (SILVA et al., 2020). O aumento na intensidade de seleção das características reprodutivas reflete em maiores taxas de ganho genético por unidade de tempo e possíveis reduções no custo de produção. Este aumento da intensidade de seleção é possível quando mais animais são genotipados do que podem ser fenotipados (HUANG et al., 2012; LOURENCO et al., 2015).

Há, no entanto, um desafiador trabalho na pecuária de corte tropical, em que são comuns conjuntos de dados mais complexos se comparados, por exemplo, ao gado de leite, em decorrência da grande quantidade de dados parentais ausentes, poucas linhagens, e grande influência dos efeitos maternos (LEGARRA et al., 2014; PICCOLI et al., 2018). Outro fator é que a pecuária leiteira é dominada por inseminação artificial, enquanto que, a monta natural predomina no sistemas, em grande parte extensivos, de produção de gado de corte. Então, muito mais touros são necessários no gado de corte e, consequentemente, muitos mais candidatos precisam ser genotipados para determinar seu mérito genético, e, o custo da genotipagem se torna um fator importante nos programas de melhoramento.

No Brasil utilizam-se informações genômicas de populações de bovinos da raça Nelore como referência para a construção dos painéis de baixa densidade, pois esta raça é mais representativa no rebanho nacional e possui maior percentual de SNPs relacionados ao objetivo de seleção comum (produção de carne) (FERRAZ et al., 2008; 2020). No entanto, isto pode contribuir para baixa precisão de imputação quando o alvo são raças com objetivos de seleção diferentes, como raças leiteiras e de duplo propósito (TORO OSPINA et al., 2021).

Assim, se a escolha dos marcadores para compor o painel é ponderada genericamente, sem especificar os objetivos de seleção, características e arquitetura genética das diferentes raças, pode ocorrer um viés que desfavorece alguns grupos, tanto quanto menor a densidade do chip (PÉREZ O'BRIEN et al., 2014). Uma entre as estratégias bem sucedidas na adoção de painéis de baixa densidade, seguida de imputação, é a customização dos chips considerando os marcadores com maiores valores para desequilíbrio de ligação, à curta distância, e valores altos de alelo de menor frequência (MAF) para específicas raças (BERNARDES et al., 2019; BOISON et al., 2014; CARVALHEIRO et al., 2014; FERRAZ et al., 2020; TORO OSPINA et al., 2021).

Hipótese

A presença dos erros genotípicos afeta as acurácias das predições genômicas e causam viés nos valores genéticos genômicos estimados, subestimando os mesmos, refletindo nas mudanças de ranqueamento de indivíduos candidatos à seleção genética.

Objetivos

Objetivos gerais

Com base no exposto, o presente trabalho busca simular taxas de erros genotípicos em dados simulados de bovinos de corte que mimetizem erros de imputação para averiguar se os mesmos conferem riscos à confiança da seleção genômica.

Objetivos específicos

- Simular uma população de animais com fenótipos e genótipos em alta densidade para uma característica com herdabilidade de 0,32, e nível de desequilíbrio de ligação similar ao de zebuínos de corte;
- Mimetizar os erros genotípicos descritos para algoritmos de imputação, sugerindo o alelo mais frequente para ocupar as posições de alelos menos frequentes;
- Realizar a avaliação genética para a predição do mérito genético do animais considerando cinco cenários de taxas de erros genotípicos (0%, 1%, 3%, 5% e 7%) utilizando o método ssGBLUP (*single-step genomic Best Linear Unbiased Prediction*);
- Investigar a intensidade de alteração do coeficiente de regressão do TBV sobre o GEBV, do coeficiente de correlação de Pearson entre TBV e GEBV e do coeficiente de correlação de Spearman entre TBV e GEBV em todos os cenários de erros.

Referências

BARROS, J. B. G. *et al*. Estimativas de adaptabilidade e estabilidade fenotípicas em características produtivas em bovinos de corte mestiços criados em diferentes ambientes. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 58, n. 4, p. 590-598, 2006.

BERNARDES, P. A. *et al*. Evaluation of imputation accuracy using the combination of two high-density panels in Nelore beef cattle. **Scientific Reports**, London, v. 9, n. 1, p. 1-10, 2019.

BOISON, S. A. *et al*. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. **Livestock Science**, Amsterdam, v. 166, p. 176-189, 2014.

CARVALHEIRO, R. *et al*. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, London, v. 46, n. 1, p. 1-11, 2014.

FERRAZ, J. B. S.; ELER, J. P. Seleção de zebuínos para características produtivas. *In*: SIMPÓSIO DE PRODUÇÃO DE GADO DE CORTE, 1., 1999, Viçosa. **Resumos** [...]. Viçosa: UFV, 1999. p. 29-49.

FERRAZ, J. B. S.; ELER, J. P. Parceria público x privada no desenvolvimento de pesquisa em melhoramento genético animal. **Revista Brasileira de Zootecnia**, Viçosa, v. 39, p. 216-222, 2010.

FERRAZ, J. B. S. *et al*. Development and evaluation of a low-density single-nucleotide polymorphism chip specific to Bos indicus cattle. **Animal Production Science**, Clayton, v. 60, n. 15, p. 1769-1776, 2020.

FERRAZ, J. B. S. *et al*. Practical use of molecular markers in beef and dairy zebu cattle in Brazil. **Acta Scientiae Veterinariae**, Porto Alegre, v. 36, n. Suppl. 2, 2008.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS – FAO. **FAOSTAT**. Disponível em: <https://www.fao.org/faostat/en/#compare>. Acesso em: 12 set. 2021.

HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relatives' records. **Journal of Animal Science**, Champaign, v. 43, n. 6, p. 1188-1197, 1976.

HUANG, Y. *et al*. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, London, v. 44, art. 25, 2012.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Diretoria de Pesquisas, Coordenação de Agropecuária. **Pesquisa da Pecuária Municipal**. 2020.

KANASHIRO, I. M.; FRAISSE, C. W. Sustainability initiatives driving supply chain: Climate governance on beef production system. **Journal of Technology Management & Innovation**, Santiago, v. 10, n. 1, p. 215-224, 2015.

LEGARRA, A. *et al*. Single Step, a general approach for genomic selection. **Livestock Science**, Amsterdam, v. 166, p. 54-65, 2014.

LOURENCO, D. A. L. *et al*. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science**, Champaign, v. 93, n. 6, p. 2653-2662, 2015.

MALAFAIA, G. C. *et al*. The Brazilian beef cattle supply chain in the next decades. **Livestock Science**, Amsterdam, v. 253, art. 104704, 2021.

MARTHA JUNIOR, G. B.; ALVES, E. R. A.; CONTINI, E. Pecuária brasileira e a economia de recursos naturais. **Perspectiva: Pesquisa Agropecuária**, Brasília, n. 1, maio. 2011.

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

PÉREZ O'BRIEN, A. M. *et al*. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, Amsterdam, v. 166, p. 121-132, 2014.

PICCOLI, M. L. *et al*. A comprehensive comparison between single-and two-step GBLUP methods in a simulated beef cattle population. **Canadian Journal of Animal Science**, Ottawa, v. 98, n. 3, p. 565-575, 2018.

SCHAEFFER, L. R. **Linear models and animal breeding**. Guelph: University of Guelph, 2010.

SILVA, R. M. O. *et al*. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. **Journal of Animal Science**, Champaign, v. 94, n. 9, p. 3613-3623, 2016.

SILVA, R. P. *et al*. Genetic parameters and genomic regions associated with calving ease in primiparous Nellore heifers. **Livestock Science**, Amsterdam, v. 240, art. 104183, 2020.

TORO OSPINA, A. M. *et al*. Assessing the accuracy of imputation in the Gyr breed using different SNP panels. **Genome**, Ottawa, v. 64, n. 10, p. 893-899, 2021.

2 REVISÃO DE LITERATURA

Os fatos apresentados nesta revisão não possuem compromisso com perspectivas históricas e não buscam designar fundadores, sob o risco de não compreender a totalidade dos esforços científicos de homens e mulheres, por vezes pouco divulgados. A principal intenção é reconstruir processos e causas que levaram à concepção de ideias fundamentais durante o desenvolvimento do melhoramento genético animal, e que refletem a proposta deste estudo.

Genética Quantitativa

Paralelamente às descobertas moleculares, desenvolveu-se uma base matemática e estatística preocupada em estimar e predizer a resposta genética dos animais sob seleção artificial. Um dos primeiros apontamentos foi a hipótese básica de que um elevado número de genes, com pequenos efeitos aditivos, explicariam a herança das características que apresentam variação fenotípica contínua (FISHER, 1918). Conceitualmente baseada em experimentos mendelianos (GREGOR, 1866) nasceu a genética quantitativa.

Francis Galton foi um dos primeiros a realizar registros biométricos com aplicação de técnicas estatísticas de análise de regressão e correlação. Demonstrou em seu trabalho “Hereditary Genius” (GALTON, 1891) que desempenhos de indivíduos que desviam da média da população também são observados nos seus descendentes, na mesma direção, mas com magnitude menor.

As primeiras teorias descritas são os índices de seleção, partindo dos trabalhos de Karl Pearson (1903), que derivou médias e variâncias condicionais para uma suposta distribuição normal multivariada. Na genética, muito se deve à invenção do princípio de máxima verossimilhança e uso da análise de variância aplicada à estimação dos componentes de variância empregados por Fisher (1912; 1922; 1925). A formalização do índice de seleção iniciou-se com Smith (1936) e Hazel (1942; 1943). Seus trabalhos eram aplicados às plantas e aos animais, respectivamente. Hazel foi responsável por esclarecer a questão da seleção de múltiplas características, assim como definiu o que eram as correlações genéticas e ambientais e sugeriu como elas poderiam ser estimadas. Ele também enfatizou a importância da ponderação por valores econômicos (HAZEL; LUSH, 1942).

Henderson (1963) apresentou uma prova simples de que os pesos econômicos poderiam ser aplicados ao valor genético calculado para características únicas, mas utilizando registros correlacionados provenientes de medições em outras características, tanto no indivíduo quanto em seus parentes. Este procedimento permitia que cada produtor tivesse seus próprios pesos econômicos, distintos entre os rebanhos, e também incluiu a variação que os pesos econômicos sofrem com o tempo.

Com o desenvolvimento da “Equação do melhorista” (LUSH, 1937), a predição genética passou a ser dada em termos de diferencial de seleção, sendo assim, “para cada unidade para a qual a média dos pais selecionados é superior à média da população, sua progênie provavelmente terá desempenho semelhantemente maior”. Isso seria literalmente verdadeiro se todos os genes tivessem ação aditiva, mas não ocorre exatamente assim, pois, uma proporção da variância genética é devido à dominância e epistasia (FALCONER; MACKAY, 1996).

Lush (1949) foi quem adotou a herdabilidade, no sentido estrito da genética quantitativa, como a razão da variância genotípica para a fenotípica e, também, o quadrado da correlação de Wright entre o valor genotípico e fenotípico. Ele inovou ao mensurar a acurácia de um preditor do valor genético para comparar planos de seleção alternativos. Seu trabalho delineou os princípios e métodos básicos para o melhoramento genético.

Charles R. Henderson (1953), aluno de Lush e Hazel, seguiu desenvolvendo outros métodos, primeiro, para estimar os parâmetros genéticos necessários para a aplicação desses princípios e, então, para predizer com maior precisão os valores genéticos dos animais disponíveis para seleção. Os métodos para estimação dos componentes de variância foram obtidos com base nas expectativas de quadrados mínimos, no entanto, dependeram de avanços na informática para amplificação do seu uso.

Grande resultado foi alcançado com as soluções das equações de modelos mistos de Henderson (1975; 1976; 1984) aliado ao modelo animal (QUAAS; POLLAK, 1980, discípulos de Henderson), que surgiu através da tentativa de combinar o poder dos mínimos quadrados, ou mínimos quadrados generalizados (GLS), para estimar efeitos fixos, com características do índice de seleção como “shrinkage”, além de combinar informações de parentes, avaliação de características

múltiplas e otimizar a média de indivíduos selecionados sob certas restrições, assim, o desempenho dos parentes seria incluído na estimativa do valor genético animal. Shayle R. Searle, também aluno de Henderson, foi um contribuidor importante para seu trabalho, publicando diversos artigos e livros sobre métodos para estimatição dos componentes de variância (SEARLE; GRUBER, 1971).

Em outros estudos, Robertson e Rendel (1950), propuseram uma fórmula geral para medir a taxa de progresso genético para gerações sobrepostas com diferentes intensidades de seleção provaram quantitativamente os méritos do teste de progênie aliado à adoção de inseminação artificial, além de estabelecer o método de comparação de grupos contemporâneos (ROBERTSON; RENDEL, 1954).

Almejando os benefícios da heterose, Dickerson (1973) discutiu meios para uso de cruzamentos de linhagens consanguíneas e entre raças, evidenciando o benefício de maior diversidade genética nos produtos e a importância da distância genética entre os indivíduos, assim como, do desempenho individual, materno e paterno, com ênfase na utilização de fêmeas mestiças e touros sintéticos ou terminais.

Muitas outras contribuições foram feitas para os procedimentos de estimação dos componentes de variância de dados desbalanceados, que resultaram nos métodos de máxima verossimilhança (HARTLEY; RAO, 1967) e de mínima variância, ou MIVQUE (RAO, 1971), além da máxima verossimilhança restrita, ou REML (PATTERSON; THOMPSON, 1971). O REML tem a propriedade de estimar os parâmetros da população base, embora apenas os dados decorrentes da seleção estejam disponíveis. O método de máxima verossimilhança restrita livre de derivadas, ou DFREML, foi obtido pela geneticista Karin Meyer (1992), tornando-se um dos mais utilizados atualmente.

Van Vleck (1976) desenvolveu procedimentos de seleção para consideração simultânea dos efeitos genéticos diretos e maternos. Com poucas diferenças sobre as práticas atuais, até o final dos anos 1970, a estimação do valor genético dos indivíduos resultava de um modelo animal multicaracterística com efeitos genéticos diretos e maternais, além de efeitos permanentes de meio, solucionados a partir de dados produtivos e genealógicos.

Seleção Tradicional

As principais características econômicas na produção animal são determinadas por um conjunto de genes com pequeno efeito individual, ou seja, são quantitativas e seu processo de seleção não é tão simples. A maioria das características são complexas, muito influenciadas pelo ambiente e apresentam distribuição contínua.

Inicialmente, Henderson (1975,1977) desenvolveu a metodologia de avaliação baseada na teoria dos modelos mistos, conhecida como Melhor Predição Linear Não-Viesada (BLUP), que utiliza um modelo associado, o Modelo Animal (QUAAS; POLLAK, 1980), o qual inclui o efeito genético aditivo como efeito aleatório, devido à segregação gênica. O modelo expressa matematicamente uma tentativa de compreender processos biológicos relativos a crescimento, sobrevivência, reprodução, entre outros. Sua função é auxiliar na explicação das observações e inferir componentes de variância.

O método BLUP combina informações de desempenho fenotípico relativo do animal e/ou de seus parentes, grau de parentesco calculado via pedigree e pressupõe o conhecimento das variâncias e co-variâncias populacionais (SCHAEFFER, 2010). Permite obter a solução dos efeitos fixos simultaneamente ao efeito aleatório animal, assim como, minimizar o quadrado médio dos erros, e, para isso, as correlações entre os valores genéticos são consideradas.

O método representa um dos maiores avanços na avaliação genética de populações animais, uma vez que, a maioria é composta de diferentes rebanhos e indivíduos com ou sem observações em algumas características (dados desbalanceados), então ele permite ajustar as soluções para as diferenças genéticas entre grupos contemporâneos, possibilitando a aplicação da metodologia a grandes conjuntos de dados, ou seja, avaliação genética em larga escala (HENDERSON; QUAAS 1976). Outro ponto importante, é o uso da matriz de parentesco, a matriz **A**. Com suas propriedades, proporciona menor viés nas predições dos méritos genéticos, o qual ocorre devido à seleção, assim pode melhorar a correlação entre valores genéticos e, dependendo da herdabilidade da característica, promover ganhos de acurácia nas avaliações. A incorporação da inversa da matriz de parentesco nas equações de modelos mistos foi fundamental para a utilização do BLUP no melhoramento animal (HENDERSON, 1976).

Portanto, obtém-se resultados mais acurados devido à inclusão das informações de parentes, em que o desempenho de cada indivíduo contribui para a

avaliação de seus familiares (ELER et al., 1995), e menos viesados pela consideração da correlação entre os valores genéticos, na estimação dos efeitos fixos. Quanto mais próximas as predições estão dos valores verdadeiros, significa maior acurácia e menor a variância do erro de predição (MRODE, 2014). As variâncias do erro de predição podem ser diferentes entre indivíduos, pois aqueles com mais informações contém erros de predição menores.

Entretanto, informações incorretas na matriz de parentesco levam à resultados viesados. O impacto da correção do pedigree vai além do número de erros, mas também depende da diferença existente entre o valor fenotípico do indivíduo e a média da família onde ele foi atribuído incorretamente, pois afeta negativamente as estimativas de herdabilidade e geram implicações significativas para a detecção de efeitos maternos. Se a maioria dos erros de linhagem forem paternos, os descendentes tenderão a se parecer mais com suas mães do que com seus pais, resultando em superestimação dos efeitos maternos (POSTMA; CHARMANTIER, 2007).

O impacto da correção do pedigree na análise BLUP não depende apenas do número de erros, mas também de quanta diferença existia entre o valor fenotípico do indivíduo e a média da família onde o indivíduo foi atribuído incorretamente (VAN VLECK, 1970a). Isso ocorre porque, na análise BLUP tradicional, a regressão dos fenótipos individuais ocorre em direção à média parental da família definida na matriz **A**. Quando o fenótipo do indivíduo erroneamente rotulado é semelhante à média da família na qual esse indivíduo foi erroneamente atribuído, o valor genético estimado será menos viesado do que em uma situação onde a diferença entre o valor fenotípico e a média da família é grande (MUNOZ et al., 2014).

Seleção Genômica

Marcadores Moleculares

Desde que as endonucleases de restrição foram identificadas (ARBER, 1974; NATHANS, 1979; NATHANS; SMITH, 1975), houve a possibilidade do corte do DNA em locais específicos e detecção dos chamados polimorfismos de comprimento de fragmentos de restrição (RFLPs), realizado com a metodologia proposta por Southern et al. (1975).

Soller e Beckmann (1983) descreveram possíveis aplicações aos recém descobertos polimorfismos, do tipo RFLPs, tal como alguns marcadores são usados atualmente, para a construção de relacionamentos genéticos muito mais precisos, identificação dos parentais, e seleção de animais mais jovens, que ainda não possuíam progênie. Os marcadores se tornaram essenciais no mapeamento de genes com influência nas características quantitativas (QTLs) como produção de leite, crescimento, eficiência alimentar, resistência a doenças, fertilidade e qualidade de carne e carcaça (CRAWFORD, 2001).

Os marcadores moleculares, revelando milhares de polimorfismos do DNA, são agora peças-chave na genética animal. Ao utilizar marcadores moleculares para estudos genéticos, do ponto de vista biológico e molecular, o procedimento de genotipagem deve ser o mais simples e de menor custo possível para gerar maior quantidade de dados de genotipagem. Do ponto de vista estatístico, de acordo com o tipo de análise a ser realizada, é necessário considerar, por exemplo, ação de dominância, neutralidade, posições no mapa ou independência genética dos marcadores. Um marcador SNP (polimorfismo de nucleotídeo único) é apenas a mudança de uma base em uma sequência de DNA, normalmente com uma alternativa de dois nucleotídeos possíveis para determinada posição. Para que esse polimorfismo seja considerado um SNP, determina-se que o alelo menos frequente deve ter uma frequência de 1% ou mais (NELSON et al., 2004).

Em qualquer análise estatística, existe uma forte ligação entre os dados e o tratamento estatístico. Um ponto importante são os erros de genotipagem que podem ser aqueles inerentes aos problemas de manipulação humana e podem ser superados com um planejamento cuidadoso dos procedimentos laboratoriais, inclusão de controles bem definidos e o aumento do grau de automação. Ou ainda, durante a leitura dos SNPs, por exemplo, podem ocorrer a não detecção de um dos dois alelos, caso em que um indivíduo heterozigoto será genotipado como homozigoto, ou o inverso, que é a falsa genotipagem de um homozigoto como heterozigoto (VIGNAL et al., 2002).

Erros de Genotipagem e Imputação

Os erros de genotipagem são praticamente inevitáveis, e ocorrem em quase todos os conjuntos de dados, exceto naqueles muito pequenos. Os fatores que

podem influenciar as taxas de erro nos marcadores SNPs são amplamente inexplorados, e muitas vezes esses erros permanecem desconhecidos ou não-observados (CALUS et al., 2011; SAUNDERS; BROHEDE; HANNAN, 2007).

Os erros, mesmo quando ocorrem em taxas extremamente baixas, podem inviabilizar muitas análises genéticas, como atribuições de parentesco e estudos de associação genômica. No entanto, poucos métodos eficazes estão disponíveis para estimar a taxa de ocorrência de erros de genotipagem e para identificar individualmente genótipos errôneos em um locus. Métodos baseados em pedigrees são poderosos, robustos e precisos. Para isso, são imprescindíveis pedigrees conhecidos e completos, como ocorrem em populações de animais de produção sob seleção (WANG et al., 2016).

Entre as consequências da existência de erros de genotipagem está a possibilidade dos dados se afastarem das leis de herança Mendeliana, nas quais muitas análises genéticas populacionais são baseadas. A chamada inconsistência Mendeliana ocorre quando os dados do genótipo e a atribuição de paternidade estão em desacordo (SOBEL; PAPP; LANGE, 2002). Por exemplo, quando um animal é identificado como homozigoto para um alelo (AA ou 0), enquanto seu pai é homozigoto para o outro alelo (BB ou 2), ou seja, os dois animais têm genótipos homozigotos "opostos". Isso pode ocorrer devido a erros no pedigree, erros de genotipagem ou mistura de amostras de DNA e, em casos muito raros, de mutações (POMPANON et al., 2005). Embora nem todos os erros de genotipagem resultem em erros Mendelianos, há uma relação linear entre o número de erros de genotipagem e as taxas de conflitos (SAUNDERS; BROHEDE; HANNAN, 2007). Essas inconsistências relacionadas aos marcadores moleculares são amplamente definidas como alguns dos principais tipos de erros de genotipagem e em certos níveis podem ter um impacto considerável. Mesmo ocorrendo em uma taxa de 1,03%, podem causar, por exemplo, exclusões de parentesco (PANETTO et al., 2017).

Inevitavelmente, com o crescimento dos conjuntos de dados e o uso crescente de genotipagem de alto rendimento (HIDALGO et al., 2021), as taxas de erros tendem a aumentar. As atribuições de parentesco são particularmente vulneráveis a problemas de genotipagem, uma vez que, na relação pai e filho, os indivíduos devem compartilhar pelo menos um alelo idêntico em estado em cada locus (GORDON; HEATH; OTT, 1999). Um único erro no par de alelos pode causar

uma incompatibilidade de genótipo e, portanto, uma falsa exclusão de parentesco, não importa quantos outros loci sejam genotipados corretamente. Pesquisadores mais cautelosos geralmente descartam marcadores com uma taxa de erro moderada (CALUS; MULDER; BASTIAANSEN, 2011), entretanto, essas taxas de erro são mais ou menos arbitrárias e podem ser muito baixas, desperdiçando informações valiosas do marcador, ou muito altas, potencialmente prejudicando análises de pedigree (PANETTO et al., 2017). Portanto, fica evidente que os erros de determinação do parentesco podem afetar a taxa de ganho genético em populações sob seleção, pois reduzem a correlação entre os valores genéticos verdadeiros e preditos (VAN VLECK, 1970b).

Outro importante fator é distinguir entre erros de pedigree, erros de genotipagem e mutações. Os erros de pedigree envolvem a identificação incorreta de relações familiares e afetam sistematicamente mais de um genótipo (DOUGLAS; SKOL; BOEHNKE, 2002). Tanto os erros de genotipagem quanto os erros de pedigree podem afetar negativamente o poder de um estudo de mapeamento genético e estimativas de desequilíbrio de ligação (POMPANON et al., 2005; WANG et al., 2016). A probabilidade de observar uma inconsistência de herança Mendeliana depende: dos membros do pedigree que são genotipados, do tipo de erro e em quem ocorre, e da distribuição de frequência do marcador (STRUCKEN et al., 2016). Para marcadores bialélicos, do tipo SNPs, na ausência de dados de genótipo parental, um erro de genotipagem não pode ser detectado entre irmãos, pois apenas três genótipos são possíveis e mesmo com dados completos, frequentemente, tais erros são de difícil detecção (CALUS; MULDER; BASTIAANSEN, 2011).

Outra fonte de erro considerada e muito distinta, são os erros de imputação. Em situações rotineiras nas análises genéticas de animais, tornou-se frequente obter amostras de populações genotipadas em diversas densidades, com tendência para uso dos chips menos densos (PAUSCH et al., 2013; WANG et al., 2016). Em vista de otimizar custos e manter a qualidade da avaliação genética, os painéis são combinados (CALUS et al., 2014; GODDARD, 2008; HABIER; FERNANDO; DEKKERS, 2009). Jovens candidatos são genotipados em menor densidade, cobrindo o genoma uniformemente, e genótipos de *loci* não identificados são imputados usando informações de uma população de referência em painel denso (MARCHINI; HOWIE, 2010; PIMENTEL et al., 2015). A ideia principal dos métodos

de imputação é rastrear haplótipos compartilhados que exibem alto desequilíbrio de ligação. No nível do haplótipo, parentes próximos compartilham haplótipos mais longos que têm menor frequência na população (BOUWMAN et al., 2014). Parentes distantes compartilham haplótipos mais curtos que geralmente têm maior frequência (MEUWISSEN; GODDARD, 2010). A imputação e o faseamento são mais precisos ao usar informações de parentes próximos (ou seja, haplótipos longos com frequência geralmente baixa) do que informações de parentes distantes (HICKEY et al., 2012; SARGOLZAEI et al., 2014). Logo, um eficiente método de imputação é explorar a genealogia ou as relações entre os indivíduos, procurando haplótipos do mais longo ao mais curto (HEIDARITABAR et al., 2015). Em gado de corte isso pode significar a genotipagem de todos os touros com o painel de alta densidade e das progênies em baixa densidade, posteriormente, a progênie poderia ter seus genótipos imputados (CARVALHEIRO et al., 2014).

A imputação precisa de alelos raros é uma tarefa desafiadora, e um fator importante nesse processo é o número de ancestrais genotipados que compoem o grupo de referência da imputação (CALUS et al., 2014). Alelos raros podem contribuir substancialmente para o que é comumente chamado de "herdabilidade ausente", ou seja, eles podem ser responsáveis por uma parte substancial da variância genética (YANG et al., 2010). Além disso, conforme diminui a frequência do alelo menor (MAF), os métodos de associação tornam-se mais sensíveis à presença de erros no genótipo (BOISON et al., 2014). Portanto, a precisão da imputação de variantes com baixo MAF é de importância e de interesse. As variantes menos frequentes tendem a ser recentes e estão associadas a haplótipos mais longos, sua maior presença tem sido reportada nos genótipos de animais zebuínos como Nelore, Gir e Brahman (PÉREZ O’BRIEN et al., 2014).

Muitos estudos têm se dedicado ao uso de dados imputados, por exemplo, Carvalheiro et al. (2014) utilizando dados de animais zebuínos, Piccoli et al. (2017) com bovinos das raças Braford e Hereford, VanRaden et al. (2013) com animais da raça Holandês, e, Yoshida et al. (2018) com genótipos da espécie salmão do Atlântico. Os autores apontam resultados onde o uso de dados imputados têm efeito sobre a acurácia dos valores genéticos genômicos estimados. Além disso, de acordo com Pimentel et al. (2015) também pode ocorrer subestimação do mérito genético de animais superiores, entre outras consequências. O autor aponta que o ganho de

acurácia está relacionado ao aumento do número de animais no grupo de treinamento, o qual mais que duplicou.

GBLUP (Genomic BLUP)

No modelo clássico da genética quantitativa, o valor fenotípico é controlado por um número infinito de genes, cada um com um efeito infinitesimal, além de fatores não genéticos ou ambientais (HENDERSON, 1976). Sob este modelo não é possível estabelecer o genótipo dos indivíduos em loci específicos e selecionar o indivíduo com os alelos mais desejáveis. Para obter maior precisão na estimação do valor genético, Meuwissen et al. (2001) propuseram uma abordagem diferente que eles chamaram de seleção genômica. Os valores genéticos genômicos (GEBVs) são calculados como a soma dos efeitos dos marcadores espalhados por quase todo o genoma, portanto, potencialmente capturando a maioria dos QTLs que contribuem para a variação em uma característica. Entretanto, a maioria das variantes causais são desconhecidas e predições mais precisas em programas de seleção genômica são impulsionadas em grande parte pelo tamanho da população de referência (GODDARD, 2009) e pelo desequilíbrio de ligação (LD) entre marcadores e *loci* de características quantitativas (QTL) (DAETWYLER et al., 2012).

O uso de informações fenotípicas e pedigree tem sido, por muitas décadas, a base de todo o progresso genético alcançado pela produção animal, e nem mesmo a adoção da genômica tornou-as obsoletas, pelo contrário. Os avanços mais marcantes do uso de marcadores moleculares são maior precisão na estimação dos relacionamentos genéticos, redução da variância do erro predito, aumento de acurácia, determinação e correção de parentescos (MISZTAL; LOURENCO; LEGARRA, 2020), e a identificação de QTLs que afetam características economicamente importantes (BRITO et al., 2011).

Os esforços iniciais buscaram incorporar as informações de marcadores ao BLUP (FERNANDO; GROSSMAN, 1989; HALEY; VISSCHER, 1998; LANDE; THOMPSON, 1990). O modelo GBLUP proposto por Van Raden (2008) permitiu estimar valores genéticos e efeitos de milhares de marcadores simultaneamente, usando métodos lineares. Em substituição ao parentesco tradicional via pedigree, as predições para animais genotipados são obtidas com base no parentesco genômico,

ou seja, durante a construção da matriz de relacionamentos são considerados a proporção de alelos compartilhados entre os animais e frequências alélicas.

Matriz de Parentesco Genômico (Matriz “G”)

As relações entre os indivíduos obtidas por pedigree são mais convenientemente chamadas de "numeradores do parentesco de Wright" (QUAAS, 1976) ou ainda, parentesco aditivo. Isso é igual à covariância padronizada (não a correlação) entre os valores genéticos aditivos de dois indivíduos. São essas medidas de parentesco definidas a partir de uma população base constituída por "fundadores", que são considerados não relacionados e portadores de alelos diferentes em QTLs importantes (VANRADEN, 2007). Por isso que as relações estimadas usando pedigrees são estritamente positivas. No entanto, este não é o caso quando consideramos informações de marcadores ou QTL (NEJATI-JAVAREMI; SMITH; GIBSON, 1997).

Relacionamentos genômicos são inferidos quantificando o número de alelos compartilhados entre dois indivíduos, informados pelos SNPs (TORO; GARCÍA-CORTÉS; LEGARRA, 2011). A princípio esses relacionamentos podem ser entendidos como “idêntico por estado” (IBS), se dois alelos escolhidos aleatoriamente, um de cada indivíduo, são idênticos, independentemente da origem. Entretanto, VanRaden (2008) propôs em seu método modificações para refletir a covariância entre indivíduos nos termos do relacionamento “idêntico por descendência” (IBD), ou seja, os alelos compartilhados, em algum momento, pertenceram a um ancestral comum, equivalente ao que ocorre no pedigree. Essas definições para IBD estariam totalmente corretas, primeiro, se o genoma considerado fosse infinito, fato que na realidade não ocorre, portanto os relacionamentos via pedigree estarão corretos, na média. Segundo, o pedigree não é infinito, limita-se pelo número de gerações observadas (THOMPSON, 2013).

Desse modo, a matriz “G” é computada de forma que na diagonal estão os números de *loci* homozigotos de cada indivíduo, e fora da diagonal estão os números de alelos compartilhados entre os membros da população (VANRADEN, 2007). Esses valores são uma aproximação dos relacionamentos genéticos mais acurada que aqueles obtidos pelo pedigree, por permitir considerar as diferenças entre indivíduos devido à amostragem mendeliana, sendo uma das principais

propriedades para a predição genômica ser considerada melhor que análises tradicionais (FERNANDES JÚNIOR et al., 2016).

ssGBLUP (Single-Step Genomic BLUP)

O desenvolvimento do ssGBLUP vem do fato de que normalmente apenas uma pequena parte dos animais, em uma determinada população, é genotipada (LOURENÇO et al., 2014). Desta forma, a melhor abordagem para evitar várias etapas nos procedimentos de obtenção dos GEBVs seria combinar pedigree e relações genômicas em uma única matriz (chamada de **H**), e utilizá-la nas Equações de Modelo Misto.

Legarra, Aguilar e Misztal (2009) mostraram que a relação do pedigree pode ser vista como uma relação esperada e a relação genômica como a relação observada. Com a derivação da distribuição conjunta dos valores genéticos de todos os animais (genotipados ou não) conseguiu extender (ou imputar) a informação genômica para animais não genotipados. Isso significa que no ssGBLUP as relações de pedigree para animais não genotipados são aprimoradas pelas informações genômicas de seus parentes. Posteriormente, todas as ferramentas baseadas nas equações do modelo misto do BLUP, como a máxima verossimilhança restrita (PATTERSON; THOMPSON, 1971), podem ser facilmente convertidas ao método de passo único.

As vantagens do ssGBLUP incluem simplicidade de uso, ajuste simultâneo de informações genômicas, estimativa de efeitos fixos, diminuição do erro tipo II e forte controle do erro tipo I (inerente ao BLUP), melhor correção para estrutura populacional na predição dos efeitos dos marcadores e acurácia relativamente maior do que métodos de várias etapas (AGUILAR et al., 2010; LEGARRA et al., 2014; LOURENÇO et al., 2014; TSURUTA; MISZTAL; LAWLOR, 2013). Possui, ainda, potencial para considerar o viés de pré-seleção visto que todo o pedigree, e as informações fenotípicas e genômicas podem ser incluídos no modelo (PATRY; DUCROCQ, 2011; WIGGANS; VANRADEN; COOPER, 2012), facilitando a extensão para qualquer modelo.

*Matriz "**H**"*

A chamada matriz "**H**" é definida como uma matriz de relacionamento construída com informações dos marcadores SNPs e de pedigree, na qual a informação genômica é projetada aos indivíduos que não são genotipados. A matriz **H** é uma matriz definida semi-positiva ou positiva (LEGARRA; AGUILAR; MISZTAL, 2009). Aguilar et al. (2010) e Christensen e Lund (2010) mostraram que embora a matriz **H** seja bastante complexa, sua inversa ($\mathbf{H}^{-1}$) é bastante simples.

Baseando-se em como a matriz **H** é construída, pressupõe-se que as duas matrizes (**G** e $\mathbf{A}_{22}$) devem ser compatíveis. Essa compatibilidade pode ser entendida como ambas as matrizes referentes à mesma base genética e à mesma variância genética. Contudo, as relações genômicas podem conter viés se a matriz **G** for construída com base em frequências de alelos diferentes daqueles da população base (VANRADEN, 2008). No entanto, as frequências alélicas da população base não são conhecidas pois os registros são recentes (ou seja, a população base *per se* é desconhecida). Embora essas frequências possam ser estimadas usando o método proposto por Gengler, Mayeres e Szydlowski (2007), essas estimativas não são muito precisas porque a população base está a várias gerações de distância dos indivíduos genotipados.

A maioria dos procedimentos estatísticos em métodos de avaliação genômica publicados até o momento presumem que todos os animais foram genotipados. No entanto, raramente é o caso e, na maioria das vezes, existem indivíduos próximos não genotipados com informações fenotípicas (GARRICK; TAYLOR; FERNANDO, 2009). Ignorar essas informações resulta em predições menos precisas e possível viés devido à seleção (VANRADEN; TOOKER; COLE, 2009). A compatibilidade entre o pedigree e as relações genômicas, pode gerar vieses ou perdas de precisão na seleção (VITEZICA et al., 2011).

Diminuir o viés é de suma importância na seleção para estimativas precisas da tendência genética e também na comparação dos animais ao longo das gerações (HENDERSON et al., 1959). O interesse no método do passo único aumentou conforme a seleção genômica foi se intensificando, uma vez que o BLUP via pedigree e, assim, os métodos de várias etapas geram estimativas de EBV mais viesadas, quando comparadas às estimativas obtidas por ssGBLUP. (PATRY; DUCROCQ, 2011).

Referências

AGUILAR, I.; MISZTAL, I.; JOHNSON, D. L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T. J. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, Lancaster, v. 93, n. 2, p. 743-752, 2010.

ARBER, W. DNA modification and restriction. **Progress in Nucleic Acid Research and Molecular Biology**, New York, v. 14, p. 1-37, 1974.

BOISON, S. A. NEVES, H. H. R.; PÉREZ O'BRIEN, A. M.; UTSUNOMIYA, Y. T.; CARVALHEIRO, R.; SILVA, M. V. G. B.; SÖLKNER, J.; GARCIA, J. F. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. **Livestock Science**, Amsterdam, v. 166, p. 176-189, 2014.

BRITO, F. V.; BRACCINI NETO, J.; SARGOLZAEI, M.; COBUCI, J. A.; SCHENKEL, F. S. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, London, v. 12, n. 1, p. 1-10, 2011.

BOUWMAN, A. C.; HICKEY, J. M.; CALUS, M. P.; VEERKAMP, R. F. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. **Genetics Selection Evolution**, London, v. 46, n. 1, p. 1-11, 2014.

CALUS, M. P. L.; MULDER, H. A.; BASTIAANSEN, J. W. M. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. **Genetics Selection Evolution**, London, v. 43, n. 1, p. 1-12, 2011.

CALUS, M.; BOUWMAN, A.; HICKEY, J.; VEERKAMP, R.; MULDER, H. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. **Animal**, Cambridge, v. 8, n. 11, p. 1743-1753, 2014.

CARVALHEIRO, R.; BOISON, S. A.; NEVES, H. H. R.; SARGOLZAEI, M.; SCHENKEL, F. S.; UTSUNOMIYA, Y. T.; O'BRIEN, A. M. P.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSEL, C. P.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, London, v. 46, n. 1, p. 1-11, 2014.

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

CRAWFORD, A. M. A review of QTL experiments in sheep. **Proceeding of the Association for the Advancement of Animal Breeding and Genetics**, [*s.l.*], v. 14, p. 33-38, 2001.

DAETWYLER, H. D.; KEMPER, K. E.; VAN DER WERF, J. H. J.; HAYES, B. J. Components of the accuracy of genomic prediction in a multi-breed sheep population. **Journal of Animal Science**, Champaign, v. 90, n. 10, p. 3375-3384, 2012.

DICKERSON, G. E. Inbreeding and heterosis in animals. **Journal of Animal Science**, Champaign, v. 1973, p. 54-77, 1973.

DOUGLAS, J. A.; SKOL, A. D.; BOEHNKE, M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. **The American Journal of Human Genetics**, [*s.l.*], v. 70, n. 2, p. 487-495, 2002.

ELER, J. P.; VAN VLECK, L. D.; FERRAZ, J. B.; LOBO, R. B. Estimation of variances due to direct and maternal effects for growth traits of Nelore cattle. **Journal of Animal Science**, Champaign, v. 73, n. 11, p. 3253-3258, 1995.

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

FERNANDO, R. L.; GROSSMAN, M. Marker-assisted selection using best linear unbiased prediction. **Genetics Selection Evolution**, London, v. 21, p. 246-477, 1989.

FERNANDES JÚNIOR, G. A.; ROSA, G. J., VALENTE, B. D., CARVALHEIRO, R., BALDI, F., GARCIA, D. A., ... & DE ANDRADE, W. B. Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution**, London, v. 48, n. 1, art. 7. 2016.

FISHER, R. A. On an absolute criterion for fitting frequency curves. **Messenger of Mathematics**, [*s.l.*], v. 41, p. 155-156, 1912.

FISHER, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. **Transactions of the Royal Society of Edinburgh**, Edinburgh, v. 52, n. 2, p. 399-433, 1918.

FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society of London**. Series A, Containing Papers of a Mathematical or Physical Character, London, v. 222, p. 309-368, 1922.

FISCHER, R. A. **Statistical methods for research workers**. Edinburgh: Oliver Boyd, 1925.

GALTON, F. **Hereditary genius**. D. Appleton, 1891.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, London, v. 41, n. 1, p. 1-8, 2009.

GENGLER, N.; MAYERES, P.; SZYDLOWSKI, M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. **Animal**, Cambridge, v. 1, n. 1, p. 21-28, 2007.

GODDARD, M. E. The use of high density genotyping in animal health. **Developments in Biologicals**, Basel, v. 132, p. 383-389, 2008.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, Dordrecht, v. 136, n. 2, p. 245-257, 2009.

GORDON, D.; HEATH, S. C.; OTT, J. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. **Human Heredity**, [*s.l.*], v. 49, n. 2, p. 65-70, 1999.

GREGOR, M. **Versuche uber pflanzenhybriden**. Verhandlungen des naturforschenden Verein Brunn. Abh. IV, 1866.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. Genomic selection using low-density marker panels. **Genetics**, Austin, v. 182, n. 1, p. 343-353, 2009.

HALEY, C. S.; VISSCHER, P. M. Strategies to utilize marker-quantitative trait loci associations. **Journal of Dairy Science**, Lancaster, v. 81, p. 85-97, 1998.

HARTLEY, H. O.; RAO, J. N. K. Maximum-likelihood estimation for the mixed analysis of variance model. **Biometrika**, [*s.l.*], v. 54, n. 1-2, p. 93-108, 1967.

HAZEL, L. N.; LUSH, J. L. The efficiency of three methods of selection. **Journal of Heredity**, [*s.l.*], v. 33, n. 11, p. 393-399, 1942.

HAZEL, L. N. The genetic basis for constructing selection indexes. **Genetics**, Austin, v. 28, n. 6, p. 476-490, 1943.

HEIDARITABAR, M.; CALUS, M. P.; VEREIJKEN, A.; GROENEN, M. A.; BASTIAANSEN, J. W. Accuracy of imputation using the most common sires as reference population in layer chickens. **BMC Genetics**, London, v. 16, n. 1, p. 1-14, 2015.

HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, [*s.l.*], v. 9, n. 2, p. 226-252, 1953.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; VON KROSIGK, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, [*s.l.*], v. 15, n. 2, p. 192-218, 1959.

HENDERSON, C. R. Selection index and expected genetic advance. **Statisitical Genetics and Plant Breeding**, 1963.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, [*s.l.*], p. 423-447, 1975.

HENDERSON, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, [*s.l.*], p. 69-83, 1976.

HENDERSON, C. R. Best linear unbiased prediction of breeding values not in the model for records. **Journal of Dairy Science**, Lancaster, v. 60, n. 5, p. 783-787, 1977.

HENDERSON, C. R. **Applications of linear models in animal breeding**. Ontario: University of Guelph, 1984. 423 p.

HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relatives' records. **Journal of Animal Science**, Champaign, v. 43, n. 6, p. 1188-1197, 1976.

HICKEY, J. M.; CROSSA, J.; BABU, R.; DE LOS CAMPOS, G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. **Crop Science**, Madison, v. 52, n. 2, p. 654-663, 2012.

HIDALGO, J.; CESARANI, A.; GARCIA, A.; SUMREDDEE, P.; LARIOS, N.; MANCIN, E.; GARCÍA, J.G.; NÚÑEZ, R.; RAMÍREZ, R. Genetic background and inbreeding depression in romosinuano cattle breed in Mexico. **Animals**, Basel, v. 11, n. 2, p. 321, 2021.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, Austin, v. 124, n. 3, p. 743-756, 1990.

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

LEGARRA, A.; CHRISTENSEN, O. F.; AGUILAR, I.; MISZTAL, I. Single Step, a general approach for genomic selection. **Livestock Science**, Amsterdam, v. 166, p. 54-65, 2014.

LOURENCO, D. A. L.; MISZTAL, I.; TSURUTA, S.; AGUILAR, I.; EZRA, E.; RON, M.; SHIRAK, A.; WELLER, J. I. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. **Journal of Dairy Science**, Lancaster, v. 97, n. 3, p. 1742-1752, 2014.

LUSH, J. L. **Animal breeding plans**. Ames, Ia: Collegiate Press, Inc, 1937.

LUSH, J. L. Heritability of quantitative characters in farm animals. **Hereditas**, [*s.l.*], v. 35, n. S1, p. 356-375, 1949.

MARCHINI, J.; HOWIE, B. Genotype imputation for genome-wide association studies. **Nature Reviews Genetics**, London, v. 11, n. 7, p. 499-511, 2010.

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

MEUWISSEN, T.; GODDARD, M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. **Genetics**, Austin, v. 185, n. 2, p. 623-631, 2010.

MEYER, K. Variance components due to direct and maternal effects for growth traits of Australian beef cattle. **Livestock Production Science**, Amsterdam, v. 31, n. 3-4, p. 179-204, 1992.

MISZTAL, I.; LOURENCO, D.; LEGARRA, A. Current status of genomic evaluation. **Journal of Animal Science**, Champaign, v. 98, n. 4, p. skaa101, 2020.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. Wallingford: Cabi, 2014.

MUNOZ, P. R.; RESENDE JR, M. F.; HUBER, D. A.; QUESADA, T.; RESENDE, M. D.; NEALE, D. B.; WEGRZYN, J. L.; KIRST, M.; PETER, G. F. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic

parameters and genomic selection accuracy. **Crop Science**, Madison, v. 54, n. 3, p. 1115-1123, 2014.

NATHANS, D. Restriction endonucleases, simian virus 40, and the new genetics. **Science**, New York, v. 206, n. 4421, p. 903-909, 1979.

NATHANS, D.; SMITH, H. O. Restriction endonucleases in the analysis and restructuring of DNA molecules. **Annual Review of Biochemistry**, Palo Alto, v. 44, n. 1, p. 273-293, 1975.

NELSON, M. R. MARNELLOS, G.; KAMMERER, S.; HOYAL, C. R.; SHI, M. M.; CANTOR, C. R.; BRAUN, A. Large-scale validation of single nucleotide polymorphisms in gene regions. **Genome Research**, Cold Spring Harbor, v. 14, n. 8, p. 1664-1668, 2004.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, Champaign, v. 75, n. 7, p. 1738-1745, 1997.

PANETTO, J. C. C.; MACHADO, M. A.; DA SILVA, M. V. G.; BARBOSA, R. S.; DOS SANTOS, G. G.; DE MH LEITE, R.; PEIXOTO, M. G. C. Parentage assignment using SNP markers, inbreeding and population size for the Brazilian Red Sindhi cattle. **Livestock Science**, Amsterdam, v. 204, p. 33-38, 2017.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, [*s.l.*], v. 58, n. 3, p. 545-554, 1971.

PATRY, C.; DUCROCQ, V. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. **Journal of Dairy Science**, Lancaster, v. 94, n. 2, p. 1011-1020, 2011.

PAUSCH, H.; AIGNER, B.; EMMERLING, R.; EDEL, C.; GÖTZ, K. U.; FRIES, R. Imputation of high-density genotypes in the Fleckvieh cattle population. **Genetics Selection Evolution**, London, v. 45, n. 1, p. 1-10, 2013.

PEARSON, K. I. Mathematical contributions to the theory of evolution.—XI. On the influence of natural selection on the variability and correlation of organs. **Philosophical Transactions of the Royal Society A**, London**,** v. 200, n. 321-330, p. 1-66, 1903.

PÉREZ O'BRIEN, A. M.; MÉSZÁROS, G. A.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, J. F.; VANTASSELL, C. P.; CARVALHEIRO, R.; SILVA, M. V. B.; SÖLKNER, J. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, Amsterdam, v. 166, p. 121-132, 2014.

PICCOLI, M. L.; BRITO, L. F.; BRACCINI, J.; CARDOSO, F. F.; SARGOLZAEI, M.; SCHENKEL, F. S. Genomic predictions for economically important traits in Brazilian Braford and Hereford beef cattle using true and imputed genotypes. **BMC Genetics**, London, v. 18, n. 1, p. 1-14, 2017.

PIMENTEL, E. C. G.; EDEL, C.; EMMERLING, R.; GÖTZ, K. U. How imputation errors bias genomic predictions. **Journal of Dairy Science**, Lancaster, v. 98, n. 6, p. 4131-4138, 2015.

POMPANON, F.; BONIN, A.; BELLEMAIN, E.; TABERLET, P. Genotyping errors: causes, consequences and solutions. **Nature Reviews Genetics**, London, v. 6, n. 11, p. 847-859, 2005.

POSTMA, E.; CHARMANTIER, A. What ‘animal models’ can and cannot tell ornithologists about the genetics of wild populations. **Journal of Ornithology**, Heidelberg, v. 148, n. 2, p. 633-642, 2007.

QUAAS, R. L. Computing the diagonal elements and inverse of a large numerator relationship matrix. **Biometrics**, [*s.l.*], p. 949-953, 1976.

QUAAS, R. L.; POLLAK, E. J. Mixed model methodology for farm and ranch beef cattle testing programs. **Journal of Animal Science**, Champaign, v. 51, n. 6, p. 1277-1287, 1980.

RAO, C. R. Estimation of variance and covariance components – MINQUE theory. **Journal of Multivariate Analysis**, New York, v. 1, n. 3, p. 257-275, 1971.

ROBERTSON, A.; RENDEL, J. M. The use of progeny testing with artificial insemination in dairy cattle. **Journal of Genetics**, London, v. 50, n. 1, p. 21-31, 1950.

ROBERTSON, A.; RENDEL, J. M. The performance of heifers got by artificial insemination. **The Journal of Agricultural Science**, Cambridge, v. 44, n. 2, p. 184-192, 1954.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, London, v. 15, n. 1, p. 1-12, 2014.

SAUNDERS, I. W.; BROHEDE, J.; HANNAN, G. N. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. **Genomics**, Orlando, v. 90, n. 3, p. 291-296, 2007.

SCHAEFFER, L. R. **Linear models and animal breeding**. Guelph: University of Guelph, 2010.

SEARLE, S. R.; GRUBER, M. H. J. **Linear models**. New York: Wiley, 1971.

SMITH, H. F. A discriminant function for plant selection. **Annals of Eugenics**, London, v. 7, n. 3, p. 240-250, 1936.

SOBEL, E.; PAPP, J. C.; LANGE, K. Detection and integration of genotyping errors in statistical genetics. **The American Journal of Human Genetics**, Baltimore, v. 70, n. 2, p. 496-508, 2002.

SOLLER, M.; BECKMANN, J. S. Genetic polymorphism in varietal identification and genetic improvement. **Theoretical and Applied Genetics**, New York, v. 67, n. 1, p. 25-33, 1983.

SOUTHERN, E. M. *et al*. Detection of specific sequences among DNA fragments separated by gel electrophoresis. **Journal of Molecular Biology**, London, v. 98, n. 3, p. 503-517, 1975.

STRUCKEN, E. M.; LEE, S. H.; LEE, H. K.; SONG, K. D.; GIBSON, J. P.; GONDRO, C. How many markers are enough? Factors influencing parentage testing in different livestock populations. **Journal of Animal Breeding and Genetics**, Berlin, v. 133, n. 1, p. 13-23, 2016.

THOMPSON, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. **Genetics**, Austin, v. 194, n. 2, p. 301-326, 2013.

TORO, M. A.; GARCÍA-CORTÉS, L. A.; LEGARRA, A. A note on the rationale for estimating genealogical coancestry from molecular markers. **Genetics Selection Evolution**, London, v. 43, n. 1, p. 1-10, 2011.

TSURUTA, S.; MISZTAL, I.; LAWLOR, T. J. Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. **Journal of Dairy Science**, Lancaster, v. 96, n. 5, p. 3332-3335, 2013.

VANRADEN, P. M. Genomic measures of relationship and inbreeding. **Interbull Bulletin**, [*s.l.*], v. 37, p. 33-36, 2007.

VANRADEN, P. M. Efficient methods to compute genomic predictions**. Journal of Dairy Science**, Lancaster, v. 91, n. 11, p. 4414-4423, 2008.

VANRADEN, P. M.; TOOKER, M. E.; COLE, J. B. Can you believe those genomic evaluations for young bulls. **Journal of Dairy Science**, Lancaster, v. 92, n. E-Suppl 1, p. 314, 2009.

VANRADEN, P. M.; NULL, D. J.; SARGOLZAEI, M.; WIGGANS, G. R.; TOOKER, M. E.; COLE, J. B.; SONSTEGARD, T. S.; CONNOR, E. E.; WINTERS, M.; VAN KAAM, J. B. C. H. M.; VALENTINI, A.; VAN DOORMAAL, B. J.; FAUST, M. A.; DOAK, G. A. Genomic imputation and evaluation using high-density Holstein genotypes. **Journal of Dairy Science**, Lancaster, v. 96, n. 1, p. 668-678, 2013.

VAN VLECK, L. D. Misidentification and sire evaluation. **Journal of Dairy Science**, Lancaster, v. 53, n. 12, p. 1697-1702, 1970a.

VAN VLECK, L. D. Misidentification in estimating the paternal sib correlation. **Journal of Dairy Science**, Lancaster, v. 53, n. 10, p. 1469-1474, 1970b.

VAN VLECK, L. D. Selection for direct, maternal and grandmaternal genetic components of economic traits. **Biometrics**, [*s.l.*], p. 173-181, 1976.

VIGNAL, A.; MILAN, D.; SANCRISTOBAL, M.; EGGEN, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, London, v. 34, n. 3, p. 275-305, 2002.

VITEZICA, Z. G.; AGUILAR, I.; MISZTAL, I.; LEGARRA, A. Bias in genomic predictions for populations under selection. **Genetics Research**, Cambridge, v. 93, n. 5, p. 357-366, 2011.

WANG, Y.; LIN, G.; LI, C.; STOTHARD, P. Genotype imputation methods and their effects on genomic predictions in cattle. **Springer Science Reviews**, [*s.l.*], v. 4, n. 2, p. 79-98, 2016.

WIGGANS, G. R.; VANRADEN, P. M.; COOPER, T. A. Adjustment of all cow evaluations for yield traits to be comparable with bull evaluations. **Journal of Dairy Science**, Lancaster, v. 95, n. 6, p. 3444-3447, 2012.

YANG, J.; BENYAMIN, B.; MCEVOY, B. P.; GORDON, S.; HENDERS, A. K.; NYHOLT, D. R.; MADDEN, P. A.; HEATH, A. C.; MARTIN, N. G.; MONTGOMERY, G. W.; GODDARD, M. E.; VISSCHER, P. M. Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics**, New York, v. 42, n. 7, p. 565-569, 2010.

YOSHIDA, G. M.; CARVALHEIRO, R.; LHORENTE, J. P.; CORREA, K.; FIGUEROA, R.; HOUSTON, R. D.; YÁÑEZ, J. M. Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. **Aquaculture**, Amsterdam, v. 491, p. 147-154, 2018.

3 MATERIAL E MÉTODOS

Simulação dos Fenótipos

Os procedimentos de simulação foram realizados de acordo com estudos anteriores (BOISON et al., 2014; BRITO et al., 2011; PICOLLI et al., 2018). Os dados de pedigree, fenótipos e genótipos, incluindo marcadores e QTL espalhados pelo genoma foram simulados por meio do software QMSim versão 2.0 (SARGOLZAEI; SCHENKEL, 2009), de maneira que a população simulada apresentasse extensão e padrão de desequilíbrio de ligação compatíveis com os verificados nas populações de bovinos de corte *Bos taurus indicus* no Brasil (ESPIGOLAN et al., 2013). Primeiro, uma população histórica de 1.000 animais foi simulada, através de acasalamentos aleatórios, em que o tamanho populacional permaneceu constante durante 1.000 gerações, então, gradualmente, foi reduzida para 200 indivíduos após 1.020 gerações (efeito gargalo). Esta etapa teve como objetivo criar o desequilíbrio de ligação inicial e permitir o estabelecimento do equilíbrio de mutação-deriva. Para esses procedimentos, o número de indivíduos de cada sexo foi o mesmo e o sistema de acasalamento baseado na união aleatória dos gametas.

Em seguida, a população foi expandida selecionando aleatoriamente 100 fundadores do sexo masculino e 100 fundadores do sexo feminino da última geração da população histórica, acasalados aleatoriamente e sem seleção. A população de expansão conteve oito gerações, em que cada fêmea produziu cinco progênies e o crescimento do número de fêmeas foi exponencial. Ao final do processo, 400 machos e 10.000 fêmeas da última geração da população de expansão foram selecionados ao acaso, para mimetizar um gargalo que reflete a domesticação e a criação moderna. A população recente foi obtida após 15 gerações. Durante a simulação machos e fêmeas foram selecionados pelo maior EBV em cada geração e acasalados, e cada fêmea gerou uma progênie com 50% de probabilidade de ser macho, totalizando 277.700 animais. Foi simulada uma característica com herdabilidade de 0,32 e variância fenotípica igual a 1,0. A taxa de substituição de touros e matrizes foi mantida constante ao longo das gerações em 60% e 20%, respectivamente, e o critério de descarte foi o menor valor de EBV. Os valores

genéticos (EBV) foram estimados dentro do software QMSim pelo método BLUP (*Best Linear Unbiased Prediction*), utilizando as equações de Henderson para um modelo animal, considerando o valor genético aditivo verdadeiro.

Simulação dos Genótipos

O genoma simulado consistiu em 29 pares de autossomos, com comprimento variando de 40 cM a 146 cM, totalizando 2.333 cM e idêntico ao genoma bovino real baseado no ARS-UCD 1.2 (NCBI RefSeq accession GCF_002263795.1). Conteve, aproximadamente, 700.000 marcadores bi-alélicos e 1.000 QTL, igualmente espaçados, em que, a quantidade de ambos por cromossomo variou de 12.931 a 46.495 SNP e de 121 a 438 QTL, mimetizando a densidade do painel comercial Illumina BovineHD 777k. Nos QTL, a quantidade de alelos por loci variou, aleatoriamente, de 2 a 4 considerando os efeitos alélicos amostrados de uma distribuição gama com parâmetro de forma igual a 0,4 (HAYES; GODDARD, 2001). Assumiu-se que 100% da variância genética fora explicada pelos QTL. A fim de prosseguir com o estabelecimento do equilíbrio mutação-deriva nas gerações históricas, a taxa de mutação recorrente para marcadores e QTL determinadas foram 10^{-4}, para ambos. Aproximadamente 600.000 marcadores (MAF ≥ 0,01) e 1.000 QTL segregantes foram selecionados aleatoriamente a partir da última geração da população histórica para gerar dados genotípicos para a população de seleção. O valor genético verdadeiro (TBV - *True Breeding Value*) de cada indivíduo foi calculado como a soma dos efeitos de substituição alélica no QTL. Os fenótipos foram gerados pela adição de um resíduo ao TBV, amostrados aleatoriamente a partir de uma distribuição normal com média igual a zero e variância igual a 1,0. Todo o processo de simulação foi replicado 5 vezes. Os parâmetros da simulação são mostrados no Quadro 1.

Controle de Qualidade dos Genótipos

A consistência dos dados genômicos foi realizada pelo software preGSf90 (MISZTAL et al., 2002). Foram considerados somente os 29 autossomos bovinos (BTA – *Bos Taurus autossome*). O controle de qualidade dos genótipos foi feito em um processo iterativo considerando os seguintes critérios para exclusão de SNP:

MAF < 0,05; SNP com posição idêntica no genoma; SNP altamente correlacionado com outros SNP e/ou SNP com desvio de heterozigosidade maior de 0,15 em relação à heterozigosidade esperada de acordo com o equilibrio de Hardy-Weinberg. Após o controle de qualidade, genótipos contendo, aproximadamente, 600.000 marcadores para 6.000 animais foram disponibilizados para as análises subsequentes.

Simulação dos Erros Genotípicos

A simulação dos erros foi aplicada sobre os genótipos de 6.000 animais escolhidos aleatoriamente das três últimas gerações recentes simuladas, e nas seguintes proporções: 30%, 20% e 10%, com proporção de machos e fêmeas de 80% e 20%, respectivamente. A simulação dos erros foi executada em linhas de programação com auxílio do software R Studio e Linux. Primeiro, em cada coluna do arquivo, correspondente a um marcador bialélico, foram mensuradas as frequências alélicas e genotípicas. Então, obtiveram-se duas subamostras, uma para o alelo A e outra para o alelo B, contendo os marcadores que apresentaram esses alelos como alelo de menor frequência (MAF) com valor entre 0,05 e 0,45. Esse processo foi realizado em cada uma das 5 réplicas das populações simuladas. Na subamostra do alelo A, as substituições do genótipo AA para AB e BB, na mesma proporção, foi a fonte do erro, e na subamostra do alelo B, as substituições do genótipo BB para AB e AA, também em proporção igual, foi a fonte do erro. A quantidade de substituições variou de acordou com as taxas de erro desejadas (0%, 1%, 3%, 5% e 7%), e essas taxas foram aplicadas por linha do arquivo, ou seja, em cada animal, e aleatoriamente nas subamostras de alelos A e B. O número de substituições foi dividido entre as duas subamostras de MAF dos alelos A e B, pois os marcadores são distintos entre ambas, mas os indivíduos não são.

Análises Estatísticas

As análises foram realizadas utilizando os programas da família BLUPF90 (MISZTAL et al., 2002) disponíveis em http://nce.ads.uga.edu/wiki/doku.php, o software R (R Core Team, 2020) e o terminal do Linux (Ubuntu 18.04 LTS). O banco de dados simulados utilizado foi composto pelo pedigree contendo 50.000 animais,

50.000 dados fenotípicos e 6.000 dados genotípicos das gerações 11, 12, 13, 14 e 15 da população recente, com 5 replicatas. O procedimento do ssGBLUP (*single-step genomic* BLUP) proposto por Misztal, Legarra e Aguilar (2009) foi utilizado para a estimação dos GEBVs dos animais simulados. O modelo pode ser representado da seguinte forma (AGUILAR el al., 2010):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

Em que: $\mathbf{y}$ é o vetor do fenótipo, $\boldsymbol{\beta}$ é o vetor dos efeitos fixos, $\mathbf{u}$ é o vetor de efeitos genético aditivo de animal, $\boldsymbol{\varepsilon}$ é o vetor dos resíduos aleatórios, e $\mathbf{X}$ e $\mathbf{Z}$ são matrizes de incidência para $\boldsymbol{\beta}$ e $\mathbf{u}$. Considerando um modelo infinitesimal, Var($\mathbf{u}$) = $\mathbf{H}\sigma_u^2$, em que $\mathbf{H}$ é uma matriz de relacionamento combinada que integra a matriz de relacionamento genômico (**G**) com matriz de parentesco via pedigree (**A**) e σ_u^2 é a variância do efeito genético aditivo. A inversa da matriz **H** apresentada por Aguilar et al. (2010) é:

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

Em que: $\mathbf{G}^{-1}$ é a inversa da matriz de relacionamento genômico e $\mathbf{A}_{22}^{-1}$ é a inversa da matriz de parentesco baseada no pedigree para animais genotipados. Segundo VanRaden (2008), a matriz G pode ser obtida de pelo menos de três formas. Para este projeto, escolhemos o seguinte:

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2\sum_{j=i}^{m} p_j\,(1 - p_j)}$$

Em que: $\mathbf{M}$ é uma matriz de marcadores com n linhas (n = número de animais genotipados) e m colunas (m = número de marcadores), e $\mathbf{P}$ é uma matriz contendo: 2(p_j - 0,5), com p_j sendo a frequência do segundo alelo. Os elementos de $\mathbf{M}$ são definidos como 0 e 2 para ambos os homozigotos e como 1 para o genótipo heterozigoto.

A definição clássica de acurácia apresentada por Van Vleck (1993) é a correlação $r(TBV, GEBV) = \frac{cov(TBV,GEBV)}{\sqrt{var(GEBV)var(TBV)}}$ para um indivíduo, que é uma medida da magnitude de mudança esperada no GEBV, com crescente número de informações. O viés é definido como a diferença absoluta das médias $DMA = \frac{1}{n}\sum_{t-1}^{n} |GEBV - TBV|$ e a dispersão como o coeficiente de regressão (b_1) do TBV sobre o GEBV $b_1 = \frac{cov(TBV,GEBV)}{var(GEBV)}$. A regressão do GEBV tem um valor esperado $E(b_{TBV,GEBV}) = 1$ se não houver inflação quando $b_1 < 1{,}0$ (subestimação) ou deflação

quando b_1 > 1,0 (superestimação). Foi calculado o coeficiente de correlação de Spearman (ρ) entre GEBV e TBV, de acordo com o ranqueamento do TBV. Foi calculada a acurácia de predição $acc_{BIF} = 1 - \sqrt{\frac{VEP}{\sigma_g^2}}$ para os valores genéticos. Além disso, investigou-se a ocorrência de mudanças na variabilidade dos dados por comparação do desvio padrão médio (σ) dos elementos fora da diagonal da matriz de relacionamentos genômicos **G**, em todos os cenários.

Para as comparações dos efeitos dos erros genotípicos entre os cinco cenários (T0%, T1%, T3%, T5%, T7%) formou-se 04 grupos de validação dos GEBV, sendo estes: Todos (TD) onde foram considerados os GEBV dos 50 mil animais avaliados; Genotipados (GE) que considerou os GEBV dos 6 mil animais genotipados; 100 superiores 'Todos' (100TD) considerando os GEBV de 100 animais superiores, classificados entre os 50 mil animais avaliados, de acordo com o TBV; 100 superiores 'Genotipados' (100GE) incluindo os GEBV de 100 animais superiores, classificados entre os 6 mil animais genotipados, de acordo com o TBV.

Quadro 1 – Opções e parâmetros usados no processo de simulação

OPÇÕES PARA SIMULAÇÃO	PARÂMETROS
POPULAÇÃO HISTÓRICA	
FASE 1 (GERAÇÕES / ANIMAIS)	1000 / 1000
FASE 2 (GERAÇÕES / ANIMAIS)	1020 / 1000 ATÉ 200
POPULAÇÃO DE EXPANSÃO	
NÚMERO DE FUNDADORES MACHOS	100
NÚMERO DE FUNDADORES FÊMEAS	100
NÚMERO DE GERAÇÕES	8
NÚMERO DE FILHOS POR FÊMEA	5
POPULAÇÃO RECENTE	
NÚMERO DE FUNDADORES MACHOS	400
NÚMERO DE FUNDADORES FÊMEAS	10000
NÚMERO DE GERAÇÕES	15
NÚMERO DE FILHOS POR FÊMEA	1
PROPORÇÃO DE MACHOS NA PROGÊNIE	0,5
ACASALAMENTO	MAIOR EBV
TAXA DE REPOSIÇÃO DE TOUROS	0,6
TAXA DE REPOSIÇÃO DE MATRIZES	0,2
CRITÉRIO DE SELEÇÃO / DESCARTE	MAIOR EBV / MENOR EBV
MÉTODO DE ESTIMAÇÃO DO VALOR GENÉTICO	BLUP E MODELO ANIMAL
HERDABILIDADE	0,32
VARIÂNCIA FENOTÍPICA	1,0
GENOMA	
NÚMERO DE CROMOSSOMOS	29
COMPRIMENTO TOTAL	2333CM
NÚMERO DE MARCADORES	735965
DISTRIBUIÇÃO DOS MARCADORES	ESPAÇADOS IGUALMENTE
NÚMERO DE QTLS	1000
DISTRIBUIÇÃO DOS QTLS	ALEATÓRIO
MAF NOS MARCADORES	0,01
MAF NOS QTLS	0,01
EFEITOS ALÉLICOS ADITIVOS NOS MARCADORES	NEUTRO
EFEITOS ALÉLICOS ADITIVOS NOS QTLS	DISTRIBUIÇÃO GAMMA (0,4)
TAXA DE MARCADORES AUSENTES	0,000
TAXA DE ERROS NOS MARCADORES	0,000
TAXA DE MUTAÇÃO RECENTE	0,0001

Fonte: Própria autoria.

4 RESULTADOS E DISCUSSÃO

O erro genotípico ocorre quando existe inconsistência entre o genótipo observado e o verdadeiro, para o mesmo indivíduo, ou, entre os genótipos de pai e filho (CALUS; MULDER; BASTIAANSEN, 2011). Este trabalho de simulação almejou mimetizar algoritmos de imputação, que normalmente combinam informações de família e população em um método determinístico, que, sugere o haplótipo mais frequente para ocupar as posições de marcadores ausentes, compartilhados por parentes próximos ou ancestrais mais remotos (BOISON et al., 2015; SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). Deste modo assumem que há algum grau de relacionamento entre os indivíduos (VAN RADEN et al., 2013).

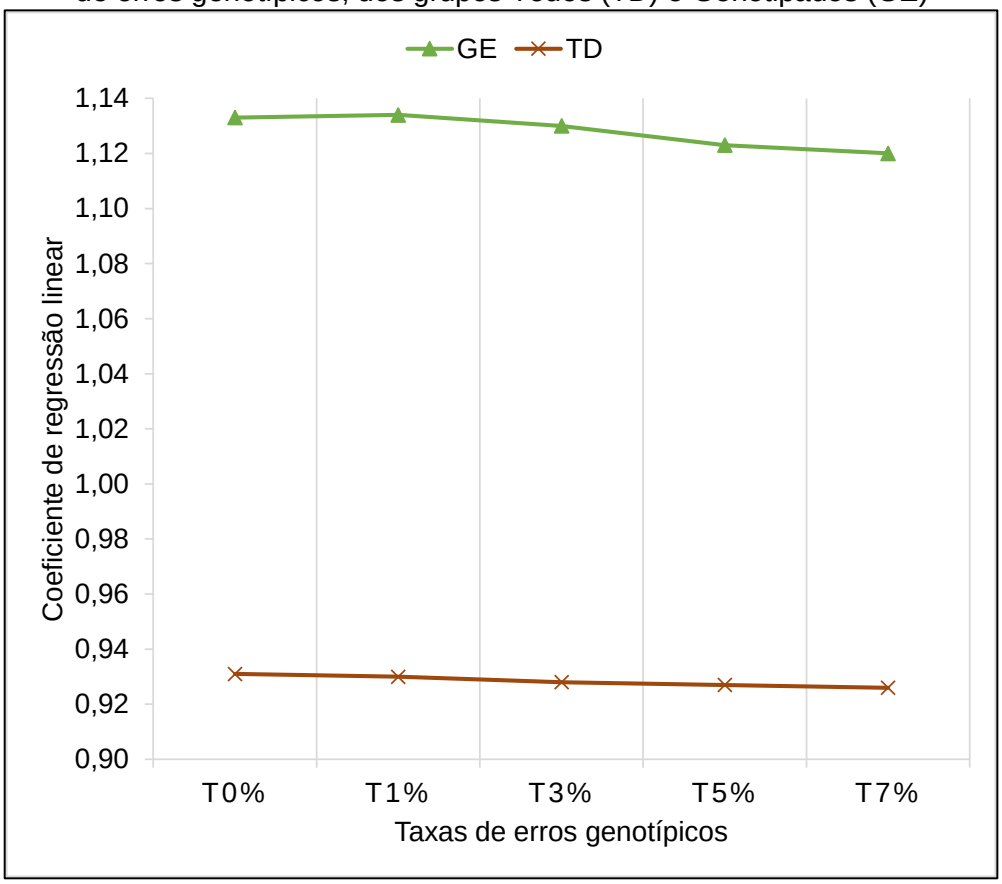
A confiabilidade da imputação é muito importante para garantir que os genótipos possam ser usados sem causar viés e prejudicar as análises subsequentes. As taxas de erros que foram simuladas correspondem à diferença entre 1 e o valor de acurácia de imputação, considerado como percentual de genótipos imputados corretamente, ou taxa de concordância, e são semelhantes aos observados em estudos com dados reais de bovinos das raças Nelore e Gir (BOISON et al., 2015; CARVALHEIRO et al., 2014; TORO OSPINA et al., 2021).

Aguns trabalhos têm demonstrado que os erros genotípicos (após a imputação de marcadores) podem ser diferentes entre os animais e por isso quanto maior o relacionamento com o grupo referência, maiores as chances de sucesso na imputação (CHEN et al., 2014; CHUD et al., 2015). A presença de um dos pais sugere maior compatilhamento de haplótipos longos e alelos de menor frequência (BOISON et al., 2017), e quanto à presença dos avós, também ocorrem benefícios para a imputação, mas em menor grau. Assim, para os animais geneticamente distantes do grupo de referência existem maiores chances que sejam-lhes atribuídos valores de covariância genética ilegítimos, se obtidos a partir dos dados imputados com baixa acurácia, surgindo como uma nova fonte de viés aos valores genéticos genômicos.

Nesse estudo de simulação, todos os animais tinham medidas fenotípicas, e dos 50 mil indivíduos considerados apenas 6 mil possuíam genótipo em alta densidade (~600K). A dispersão dos valores genéticos para os diferentes cenários de taxas de erros genotípicos, medida pelo coeficiente de regressão (b_1) do TBV sobre o GEBV, mostra que houve inflação e deflação do GEBV (Figuras 1 e 2). Para

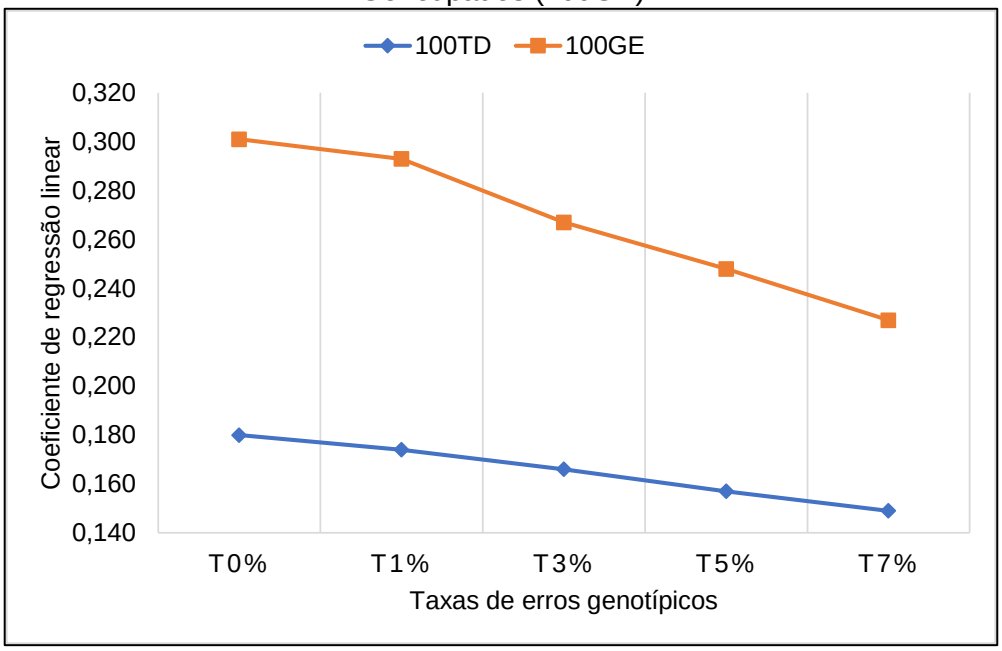
o grupo Todos (TD) mesmo com a presença de erros no genótipo, b_1 apesentou valores muito próximos de 1,0 (entre 0,931 e 0,926, para T0% e T7%, respectivamente) indicando ausência de inflação. Por outro lado, no grupo de animais Genotipados (GE) foram encontrados valores de b_1 entre 1,133 a 1,120, nos cenários T0% e T7%, respectivamente, que sugere deflação dos valores genéticos, ou seja, em geral, as predições genômicas foram menores em valores absolutos do que o esperado (TSURUTA et al., 2021). Esse resultado faz sentido à presença de animais genotipados apenas nas três últimas gerações (13, 14 e 15), em proporção de 10, 20 e 30% dos indivíduos, respectivamente, o que poderia ter criado mais viés nos valores genéticos (VITEZICA et al., 2011). Outra possibilidade é a baixa compatibilidade entre a matriz de relacionamento via pedigree (**A**) e a matriz de relacionamento genômico (**G**), pois ao combinar todas as informações disponíveis na matriz **H** do BLUP genômico de etapa única (ssGBLUP), **A** é construída usando todos os animais disponíveis no pedigree e **G** apenas os animais genotipados da população. A diferença no coeficiente do GEBV entre os grupos GE (b_1 = 1,133) e 100GE (b_1 = 0,301) mostra que houve inflação (b_1 < 1,0) no grupo 100GE mesmo quando não haviam erros genotípicos (T0%), apesar do grupo GE apresentar deflação (b_1 > 1,0). Uma justificativa é que nesse estudo foram simulados alta intensidade de seleção e acasalamento seletivo positivo, fatores que podem ser responsáveis por viés nas previsões genômicas e causar inflação (TSURUTA et al., 2021).

Figura 1 – Coeficiente de regressão do TBV sobre o GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)



Fonte: Própria autoria.

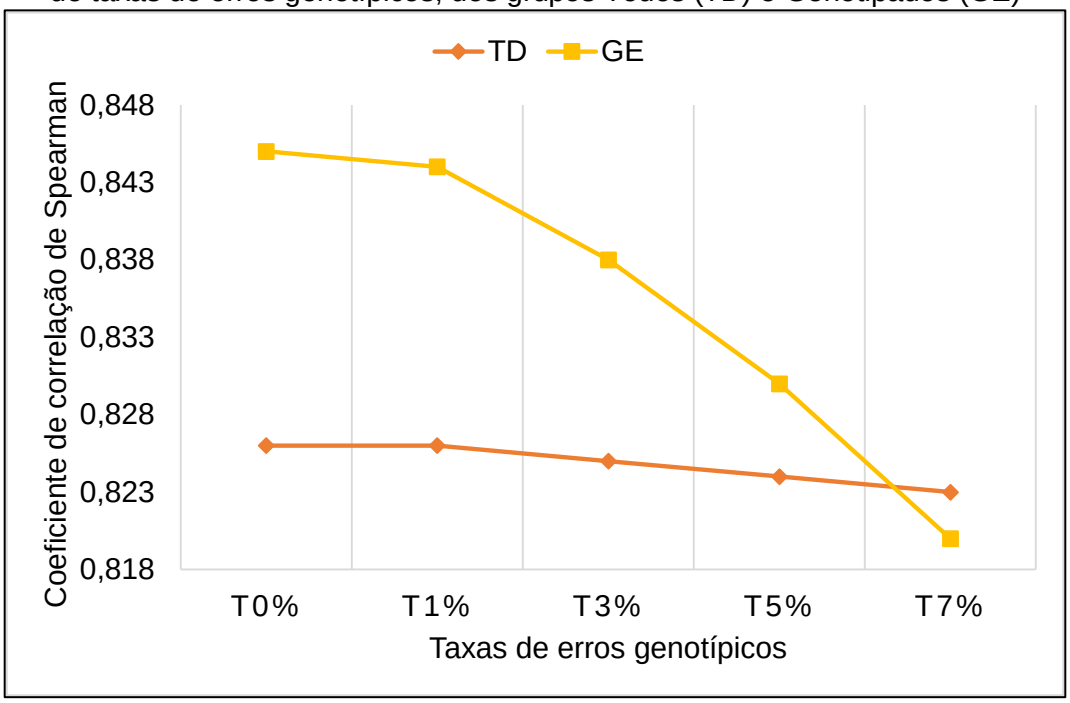
Figura 2 – Coeficiente de regressão do TBV sobre o GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)



Fonte: Própria autoria.

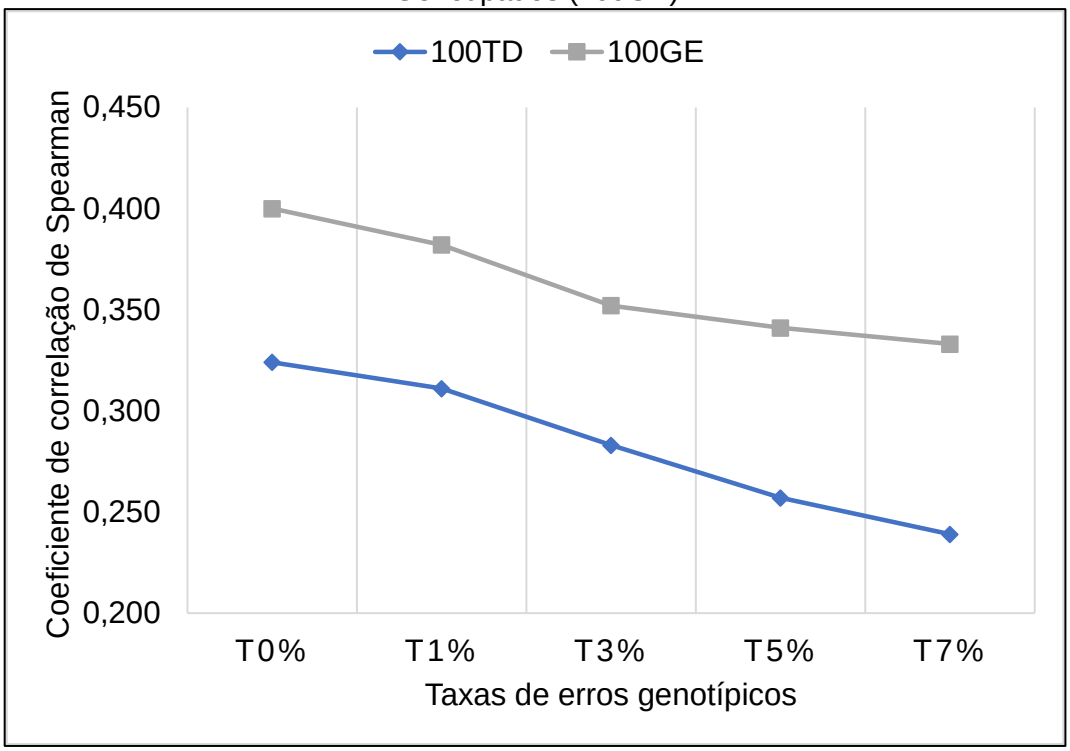
As taxas de erros genotípicos tiveram um efeito substancial sobre os coeficientes de correlação de Spearman entre TBV e GEBV nos grupos de animais superiores (100TD e 100GE), enquanto que nos grupos TD e GE percebeu-se que houve pouca alteração. Na figura 3, os coeficientes do grupo TD sofreu a menor redução (0,36%) com valores entre 0,826 (T0%) e 0,823 (T7%), já no grupo GE os coeficientes foram um pouco melhores, porém a redução (2,96%) foi mais acentuada e as correlações foram 0,845 em T0%, e 0,820 em T7%. Na figura 4, observamos que os valores dos coeficientes no grupo 100TD foram de 0,324 a 0,239 (redução de 26,23%) e no grupo 100GE foram de 0,400 a 0,333 (redução de 16,75%), ambos nos cenários T0% e T7%, respectivamente, indicando maior reclassificação entre indivíduos superiores geneticamente. Essas baixas correlações podem estar associadas à inflação observada nos GEBV desses grupos (Figura 2). Em um programa de melhoramento genético, o ranqueamento dos animais após a predição genética é uma etapa importante, pois é o momento em que os candidatos à seleção serão escolhidos e acasalados para gerar a próxima geração (multiplicação do material genético), então, a presença de erros genotípicos, como os erros de imputação (PIMENTEL et al., 2015) podem causar ranqueamentos inconsistentes e viesar a seleção dos seus melhores candidatos. Pode-se considerar que os animais superiores (100TD e 100GE) possuem, em média, os melhores haplótipos no que diz respeito aos seus efeitos para a característica estudada (FALCONER; MACKAY, 1996). Quando o alelo raro é favorável, esperamos que a maioria dos animais com esse alelo seja do grupo superior no ranqueamento, mas com os erros genotípicos alguns animais correm o risco de perder posições de classificação, para investigar e confirmar essa hipótese foi feita durante a simulação dos erros genotípicos, em 50% dos casos, a substituição do genótipo homozigoto do alelo mais frequente pelo homozigoto do alelo menos frequente.

Figura 3 – Coeficiente de correlação de Spearman entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)



Fonte: Própria autoria.

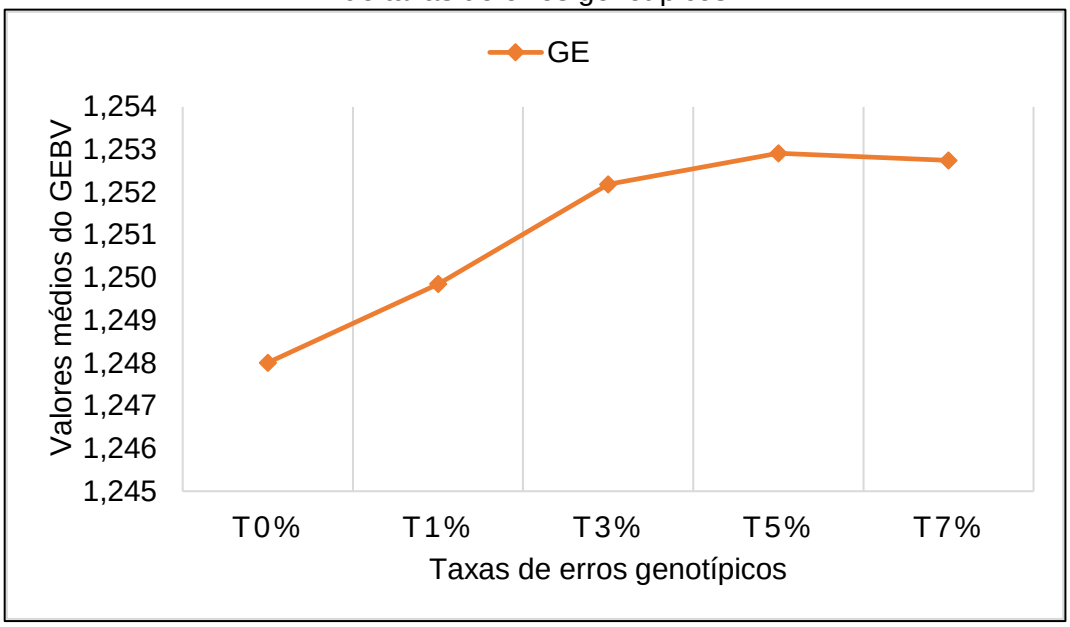
Figura 4 – Coeficiente de correlação de Spearman entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)



Fonte: Própria autoria.

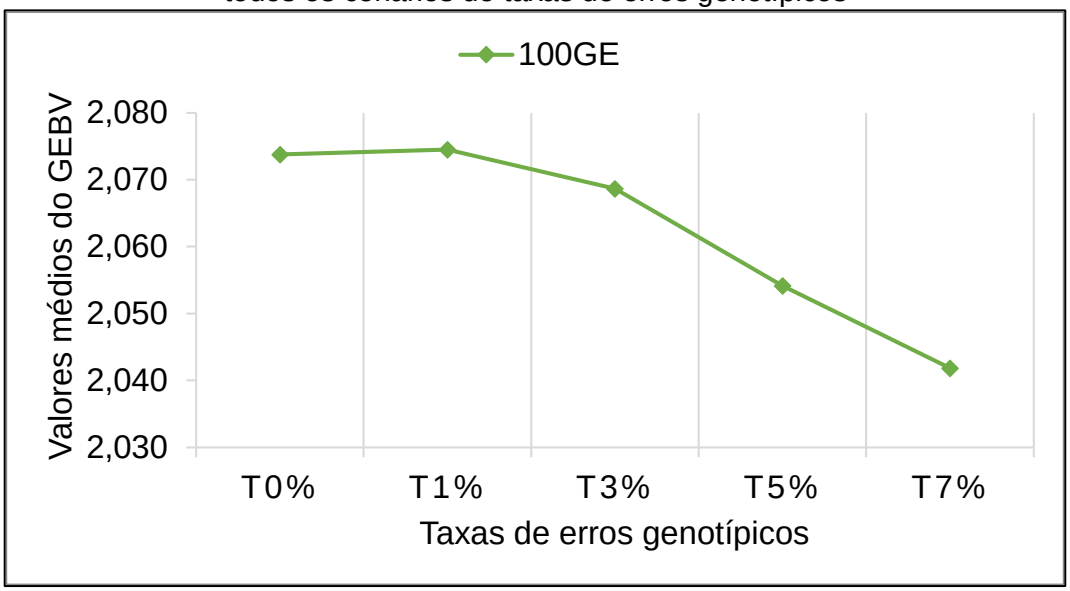
A média geral do GEBV no grupo GE apresentou um leve aumento (0,38%) em resposta às taxas de erros e no grupo dos animais superiores (100GE) houve uma pequena diminuição (1,54%) quanto maior a taxa de erros no genótipo, de acordo com as Figuras 5 e 6. Um fator de alteração pode ser que o GEBV dos animais mais jovens genotipados (com erros), foi calculado pelo GEBV inflado de indivíduos genotipados também com erros. Ainda, essas observações indicam que as mudanças no GEBV causadas por erros genotípicos podem seguir algum padrão sistemático com efeitos mais relevantes aos indivíduos geneticamente superiores. No trabalho de Pimentel et al. (2015), a presença de alteração nas estimativas tornou-se mais evidente para as características sob alta intensidade de seleção, que pode ser atribuído à tendência genética que separa o grupo de candidatos à seleção dos grupos de calibração, do modelo de predição, e de referência na imputação, aonde os jovens candidatos são, em média, melhores para a maioria das características em comparação com os grupos de referência.

Figura 5 – Valor médio do GEBV para os animais genotipados (GE), em todos os cenários de taxas de erros genotípicos



Fonte: Própria autoria.

Figura 6 – Valor médio do GEBV para os 100 animais superiores genotipados (100GE), em todos os cenários de taxas de erros genotípicos

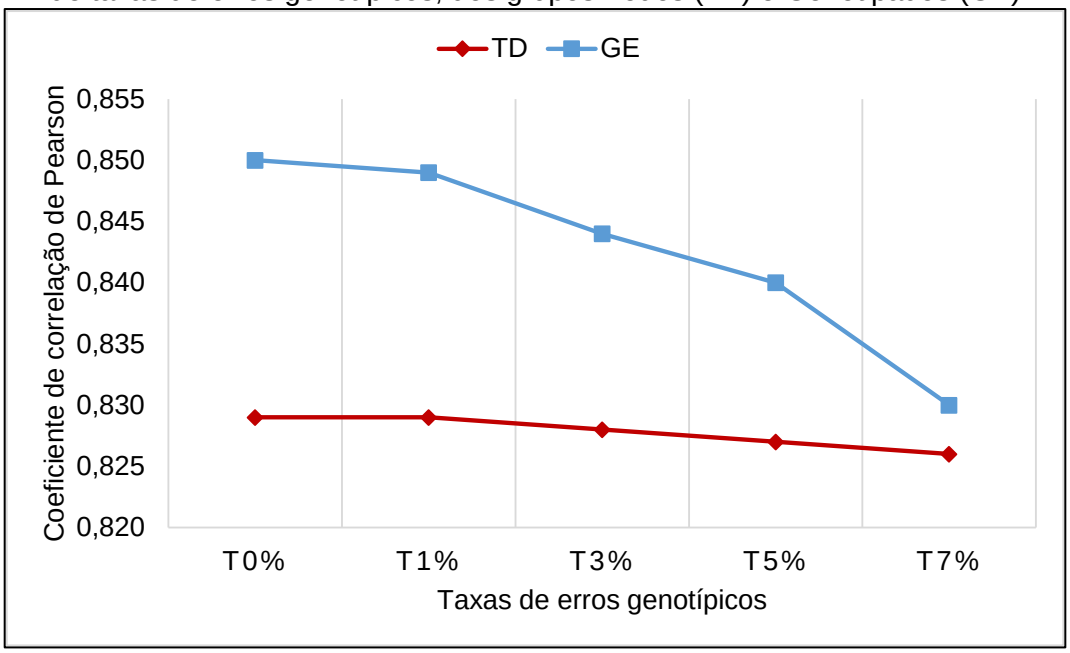


Fonte: Própria autoria.

Um conceito comum na análise da acurácia de predição dos GEBVs é o estudo da associação entre estes e o valor genético real (TBV – True Breeding Value) obtida por correlação, entretando o TBV é um dado inacessível nas situações de dados reais. Como a simulação permite conhecer o TBV, não foi necessário utilizar um pseudo fenótipo, que poderia ser o valor genético tradicional (EBV – Estimated Breeding Value), o EBV desregredido ou, ainda, o fenótipo corrigido para os efeitos fixos como variável resposta nas predições genômicas (MOROTA et al., 2014). Nas figuras 7 e 8 são apresentados os coeficientes de correlação de Pearson entre TBV e GEBV. A ocorrência de taxas de erros nos genótipos diminuiu a acurácia do GEBV, principalmente entre os animais genotipados (GE e 100GE). No grupo TD a correlação diminuiu 0,36% e no grupo GE 2,35%, no cenário de 7% de erros. Entre os animais superiores as acurácias diminuíram 16,34% e 21,05% com a taxa de 7% de erros, nos grupos 100TD e 100GE, respectivamente. O maior valor de correlação encontrado foi para o grupo GE (0,850) no cenário sem erros (T0%), porém, no cenário T7% o valor da acurácia desse grupo (0,830) foi muito próximo ao valor obtido no cenário T0% para o grupo TD (0,829). Sugere-se que ao considerar todos os animais (TD) os efeitos dos erros ficaram diluídos porque havia maior número de animais não genotipados (88%). Um dos principais objetivos da genotipagem é aumentar a acurácia das predições genômicas de candidatos à seleção, mas esse ganho pode ser perdido se são permitidas altas taxas de erros

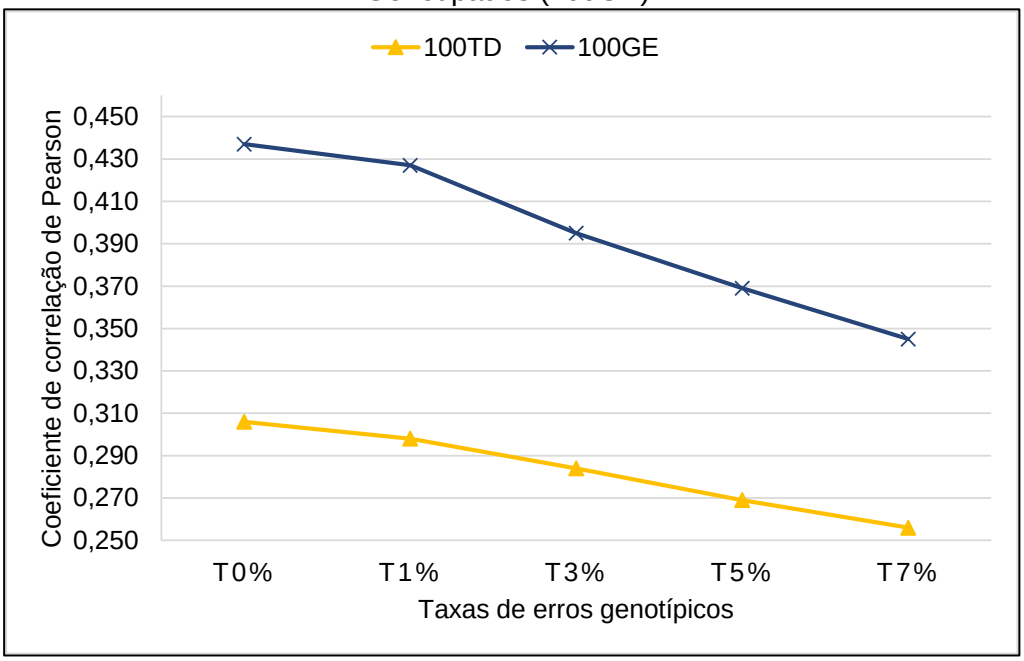
genotípicos, principalmente para características de baixa herdabilidade (AKBARPOUR; HOSSEIN-ZADEH; SHADPARVAR, 2021).

Figura 7 – Coeficientes de correlação de Pearson entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, dos grupos Todos (TD) e Genotipados (GE)



Fonte: Própria autoria.

Figura 8 – Coeficientes de correlação de Pearson entre TBV e GEBV, em todos os cenários de taxas de erros genotípicos, para os 100 animais superiores dos grupos Todos (100TD) e Genotipados (100GE)

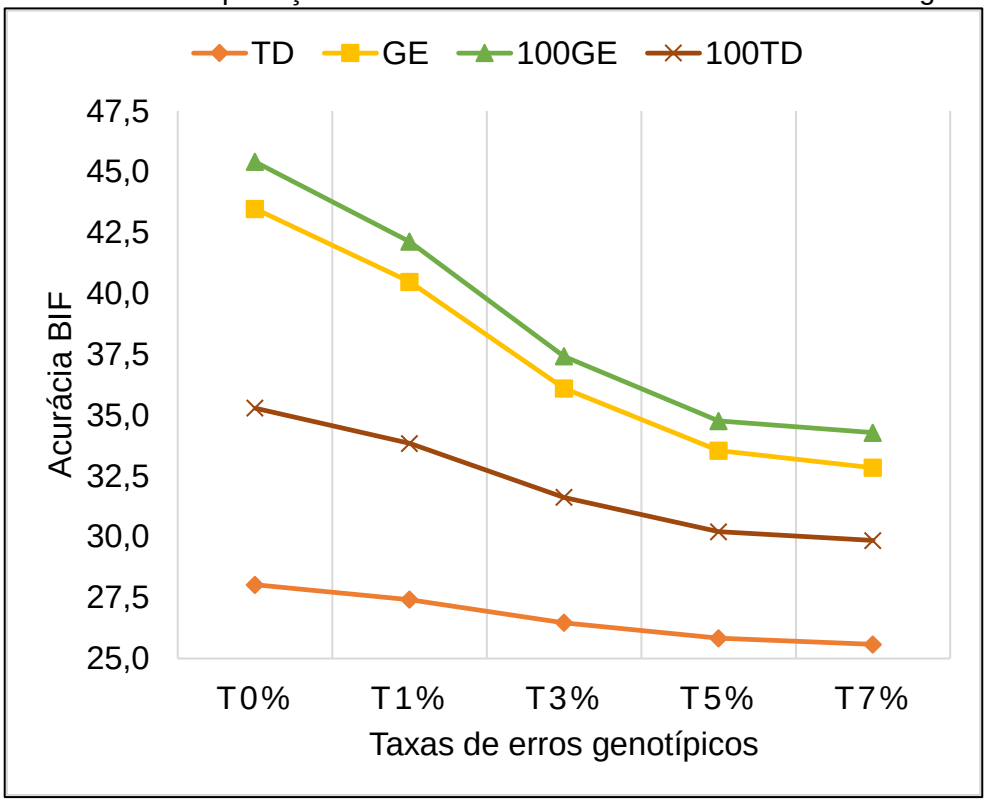


Fonte: Própria autoria.

Observamos nesse estudo que a presença dos erros genotípicos pode levar à perda de variação genética nos relacionamentos genômicos e nos dados, indicada pela diminuição da acurácia de predição BIF (Figura 9), aumento da variância do erro de predição (Figura 10) e do desvio padrão dos elementos fora da diagonal da matriz de relacionamentos **G** (Figuras 12). Ao tornar maior a frequência dos genótipos mais comuns observados na amostra, essa substituição remove a variância genética dos dados, e esse pode ser o fator redutor das acurácias, em especial para os animais genotipados (GE e 100GE). A diminuição na acurácia de predição BIF pode estar diretamente relacionada ao aumento da quantidade de erros genotípicos e, consequentemente, da variância dos erros de predição. A variância dos erros de predição, comparada ao T0%, apresentou aumento no cenário T7% em todos os grupos: TD (3,73%), GE (41,11%), TD100 (10,83%) e GE100 (44,73%). A figura 11 apresenta o relacionamento linear (R^2 = 0,6703) entre o aumento das taxas de erros genotípicos e a diminuição da acurácia de predição BIF, no grupo de animais genotipados (GE), representado pela equação:

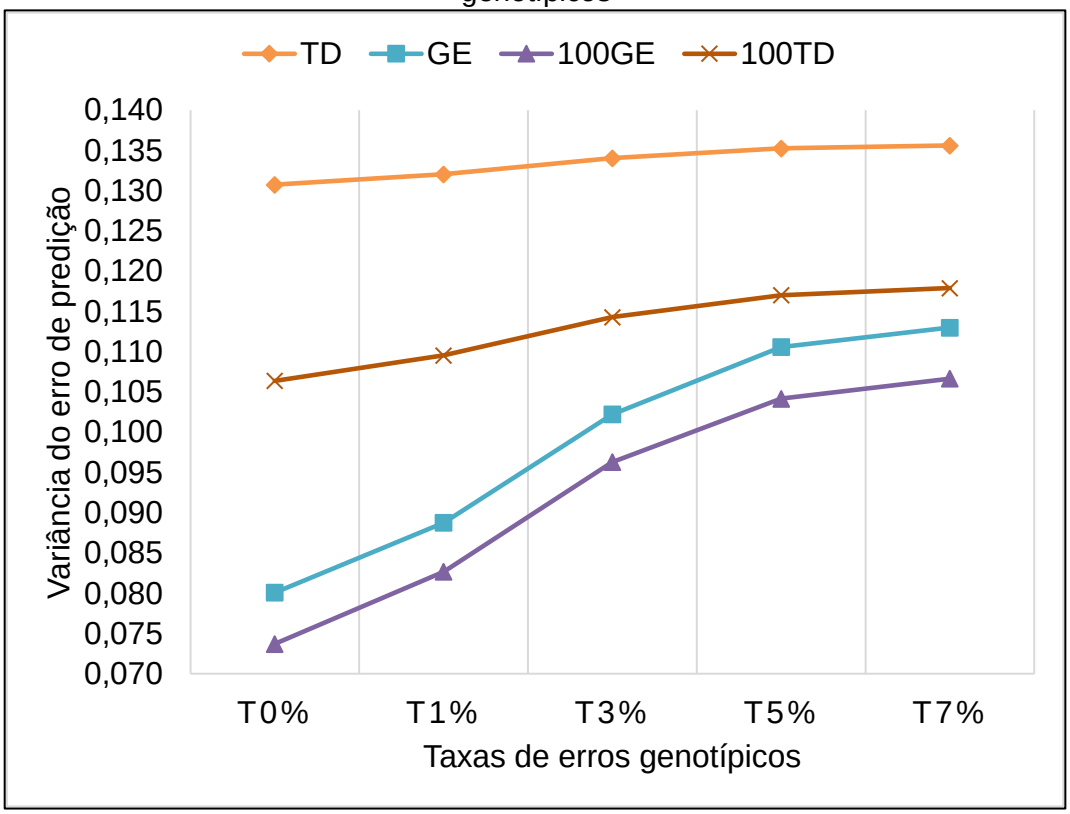
$$y = -0{,}3751 - 1{,}7788x$$

Figura 9 – Acurácia de predição BIF em todos os cenários de taxas de erros genotípicos



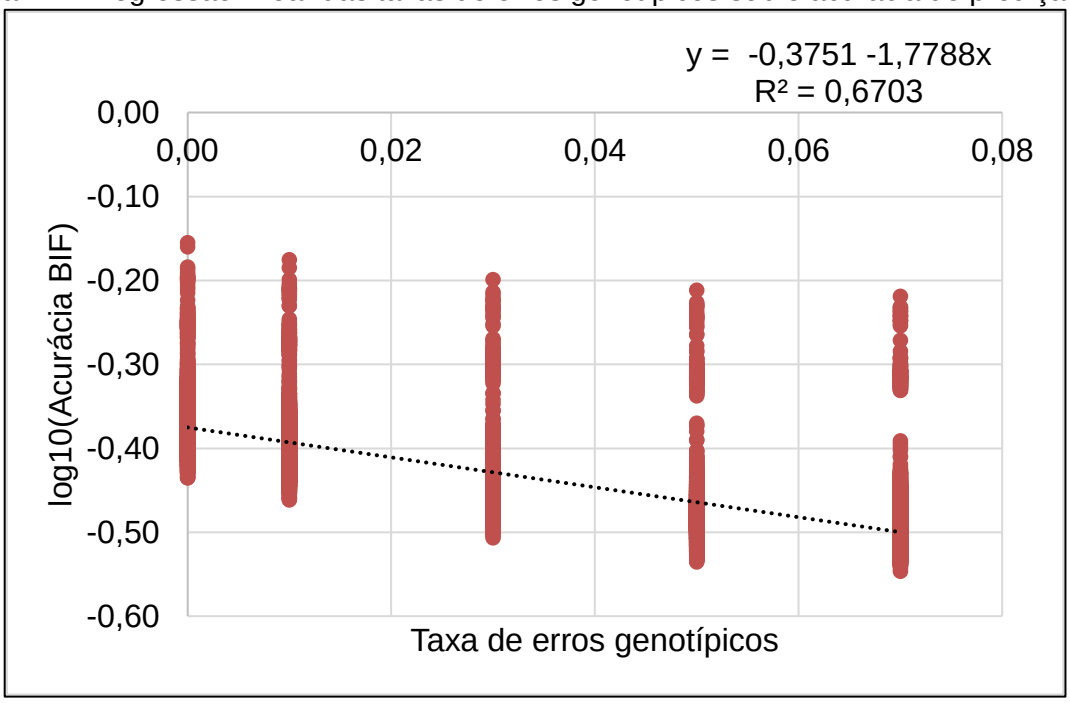
Fonte: Própria autoria.

Figura 10 – Variância do erro de predição, em todos os cenários de taxas de erros genotípicos



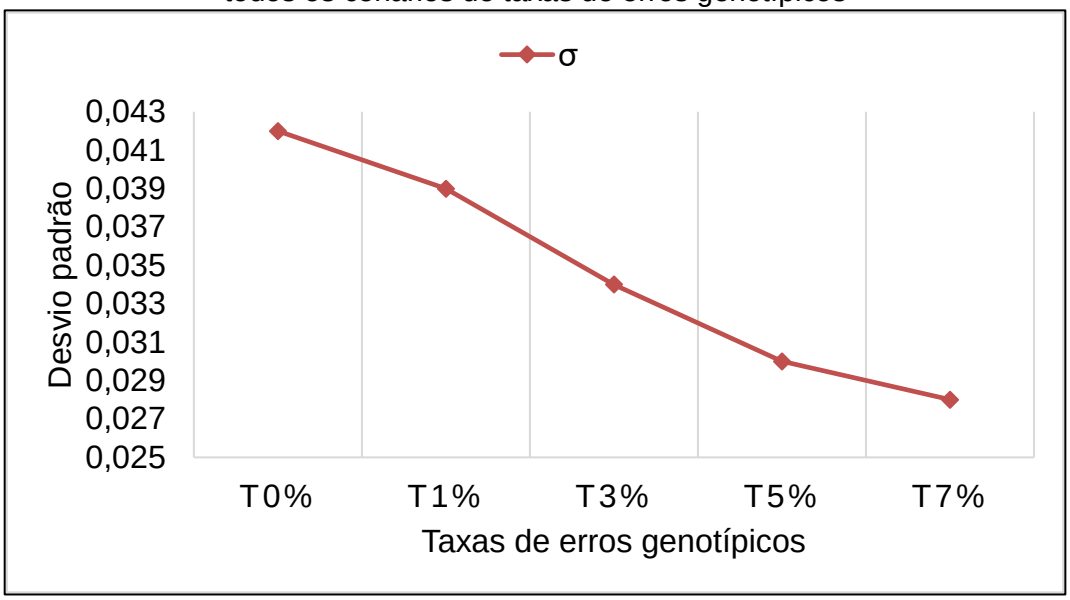
Fonte: Própria autoria.

Figura 11 – Regressão linear das taxas de erros genotípicos sobre acurácia de predição BIF



Fonte: Própria autoria.

Figura 12 – Desvio padrão (σ) dos elementos fora da diagonal da matriz genômica, em todos os cenários de taxas de erros genotípicos



Fonte: Própria autoria.

Para compreender a origem da influência dos erros e seu reflexo na acurácia de predição é necessário relembrar como foi computada a matriz genômica de parentesco, pois este trabalho propôs simular os erros a partir da alteração das frequências de alelos e genótipos menos frequentes. Com o uso de marcadores genéticos estima-se a proporção de segmentos cromossômicos compartilhados por indivíduos com base na informação de idêntidade por estado (IBS) dos alelos do marcador (AGUILAR et al., 2010). As correspondências de alelos raros neste locus serão muito informativas sobre a ancestralidade comum, dado que, a probabilidade de que um par de indivíduos, não relacionado, seja homozigoto para o alelo menor em um mesmo locus é mínima. A definição de Wright (1922) do coeficiente de parentesco via pedigree obtido como uma correlação sugere que um estimador de coeficiente de relacionamento realizado deve ser baseado na covariância ou correlação entre as pontuações de dosagem dos alelos (0, 1 ou 2) dos marcadores para cada par de indivíduos. A computação da matriz **Z** é uma função das frequências dos alelos, que confere maior contribuição ou peso para genótipos raros homozigotos, precisamente a ponderação ideal para refletir a relação de identidade por descendência (IBD).

Como $\mathbf{ZZ}^T$ é uma matriz de covariância, seus elementos ficam maiores à medida que são incluídos mais marcadores no cálculo. Para remover a influência da quantidade de marcadores da estimativa, calcula-se a covariância média por

marcador, dividindo esta matriz pela soma das variâncias em cada locus, que resulta na matriz denominada matriz de relacionamento genômico realizado, **G** (FORNI; AGUILAR; MISZTAL, 2011; VANRADEN, 2008).

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{2\Sigma \mathrm{p_i}(1 - \mathrm{p_i})}$$

Os elementos da matriz **G** estimam os valores realizados de duas vezes a probabilidade de que os indivíduos i e j sejam IBD para alelos amostrados aleatoriamente para um locus comum. Portanto informações errôneas de genótipos, que surgem após procedimentos de imputação, podem levar à parentescos ilegítimos entre animais (BOISON et al., 2017; PIMENTEL et al., 2015) e, consequentemente, indivíduos em famílias não relacionadas podem compartilhar mais alelos do que o esperado ao acaso. Os resultados desse estudem sugerem maior rigor quanto à qualidade das informações genotípicas utilizadas na predição dos valores genéticos genômicos e seleção genética animal.

5 CONCLUSÃO

Foi possível investigar que o aumento dos alelos mais frequentes, que ocorreu após a inserção dos erros, diminuiu o desvio dos valores genéticos em relação à média populacional, corroborando com as teorias da genética quantitativa. Animais superiores candidatos à seleção podem ser penalizados durante a avaliação se o genótipo contiver erros genotípicos semelhantes aos simulados neste trabalho. Observou-se um padrão linear em todas as mudanças causadas pelos erros genotípicos simulados, entre os 100 animais superiores. Erros de genotipagem podem causar perda de variância genética nos dados e induzir à menor acurácia de predição genômica.

Referências

AGUILAR, I. *et al*. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Animal Science**, Champaign, v. 93, p. 743-752, 2010.

AKBARPOUR, T.; HOSSEIN-ZADEH, N. G.; SHADPARVAR, A. A. Marker genotyping error effects on genomic predictions under different genetic architectures. **Molecular Genetics and Genomics**, Berlin, v. 296, n. 1, p. 79-89, 2021.

BOISON, S. A. *et al*. Imputation of non-genotyped individuals using genotyped progeny in Nellore, a Bos indicus cattle breed. **Livestock Science**, Amsterdam, v. 166, p. 176-189, 2014.

BOISON, S. A. *et al*. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. **Journal of Dairy Science**, Lancaster, v. 98, n. 7, p. 4969-4989, 2015.

BOISON, S. A. *et al*. Accuracy of genomic predictions in Gyr (Bos indicus) dairy cattle. **Journal of Dairy Science**, Lancaster, v. 100, n. 7, p. 5479-5490, 2017.

BRITO, F. V.; NETO, J. B.; SARGOLZAEI, M.; COBUCI, J. A.; SCHENKEL, F. S. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, London, v. 12, p. 80-89, 2011.

CALUS, M. P. L.; MULDER, H. A.; BASTIAANSEN, J. W. M. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. **Genetics Selection Evolution**, London, v. 43, n. 1, p. 1-12, 2011.

CARVALHEIRO, R. *et al*. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, London, v. 46, n. 1, p. 1-11, 2014.

CHEN, L. *et al*. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. **PloS One**, San Francisco, v. 9, n. 7, p. e101544, 2014.

CHUD, T. C. S. *et al*. Strategies for genotype imputation in composite beef cattle. **BMC Genetics**, London, v. 16, n. 1, p. 1-10, 2015.

ESPIGOLAN, R. *et al*. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC Genomics**, London, v.14, p.1-8, 2013.

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

FORNI, S.; AGUILAR, I.; MISZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. **Genetics Selection Evolution**, London, v. 43, p.1-7, 2011.

HAYES, B.; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, London, v. 33, n.3, p.209-229, 2001.

MISZTAL, I.; TSURUTA, S.; STRABEL, T.; AUVRAY, B.; DRUET, T.; LEE, D. H. BLUPF90 and related programs (BGF90). *In*: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2022, Montpellier. **Proceedings** […]. Montpellier, 2002. CD-ROM.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, Lancaster, v. 92, p. 4648-4655, 2009.

MOROTA, G. *et al*. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Frontiers in genetics**, v. 5, p. 56, 2014.

PICCOLI, M. L. *et al*. A comprehensive comparison between single-and two-step GBLUP methods in a simulated beef cattle population. **Canadian Journal of Animal Science**, v. 98, n. 3, p. 565-575, 2018.

PIMENTEL, E. C. G.; EDEL, C.; EMMERLING, R.; GÖTZ, K. U. How imputation errors bias genomic predictions. **Journal of Dairy Science**, Lancaster, v. 98, p. 4131-4138, 2015.

R Core Team: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2020.

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

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, London, v. 15, p. 478-10, 2014.

TORO OSPINA, A. M. *et al*. Assessing the accuracy of imputation in the Gyr breed using different SNP panels. **Genome**, Ottawa, v. 64, n. 10, p. 893-899, 2021.

TSURUTA, S.; LAWLOR, T. J.; LOURENCO, D. A. L.; MISZTAL, I. Bias in genomic predictions by mating practices for linear type traits in a large-scale genomic evaluation. **Journal of Dairy Science**, Lancaster, v. 104, n. 1, p. 662-677, 2021.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, Lancaster, v. 91, p. 4414-4423, 2008.

VANRADEN, P. M. *et al*. Genomic imputation and evaluation using high-density Holstein genotypes. **Journal of Dairy Science**, Lancaster, v. 96, n. 1, p. 668-678, 2013.

VAN VLECK, L. D. Variance of prediction error with mixed model equations when relationships are ignored. **Theoretical and Applied Genetics**, Berlin, v. 85, n. 5, p. 545-549, 1993.

VITEZICA, Z. G.; AGUILAR, I.; MISZTAL, I.; LEGARRA, A. Bias in genomic predictions for populations under selection. **Genetics Research**, Cambridge, v. 93, n. 5, p. 357-366, 2011.

WRIGHT, S. Coefficients of inbreeding and relationship. **The American Naturalist**, Chicago, v. 56, n. 645, p. 330-338, 1922.