

Efeitos dos erros de genotipagem na estimação dos valores genéticos genômicos: um estudo de simulação

Fernanda Schneberger^{1*}, Rafael Espigolan¹, Alisson Stefany Acero Valderrama¹, Bárbara da Conceição Abreu Silva¹, Fernando de Oliveira Bussiman³, Caroline Assis Almeida¹, Felipe Eguti de Carvalho¹, José Bento Sterman Ferraz¹, Elisângela Chicaroni Mattos¹, Fernando Baldi², Joanir Pereira Eler¹

¹ Departamento de Medicina Veterinária, Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo (ZMV-FZEA/USP), Pirassununga, SP, Brasil.

² Departamento de Zootecnia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista “Júlio de Mesquita Filho” (FCAV/UNESP), Jaboticabal, Brasil

³ Departamento de Nutrição e Produção Animal, Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo (VNP-FMVZ/USP), Pirassununga, SP, Brasil.

*Autor correspondente: fschneberger@usp.br

Resumo: A seleção genômica permitiu ganhos de acurácia na predição do valor genético de candidatas à seleção mais jovens, com maior conhecimento da covariância genética entre os indivíduos. A existência de erros na informação genotípica é inevitável, com consequências pouco conhecidas. Foram analisados dados simulados de 50.000 animais, 6.000 genótipos em alta densidade, para uma característica com herdabilidade de 0,32 e variância fenotípica igual a 1,0. As predições foram obtidas via ssGBLUP para cinco cenários de taxas de erros genotípicos (0%, 1,0%, 3,0%, 5,0% e 7,0%). Os erros genotípicos causam diminuição da variância entre os relacionamentos genômicos, vies na estimação dos valores genéticos e diminuição da acurácia de predição genômica quanto maior a taxa de erro presente nos genótipos.

Palavras-chave: Acurácia, bovinos, seleção genômica, SNP, ssGBLUP

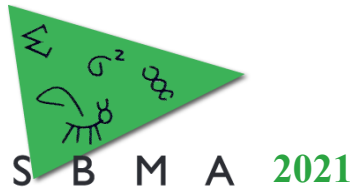
Effects of genotyping errors on the estimation of genomic breeding values: a simulation study

Abstract: Genomic selection allowed for accuracy gains in the predicted breeding values of younger candidates for selection, with the genetic covariance between individuals better known. The existence of errors in genotypic information is inevitable, with consequences that are little known and clarified. Simulated data from 50,000 animals, 6,000 genotypes in high density, for a heritability trait of 0.32 and phenotypic variance 1.0, were analyzed. Predictions were obtained via ssGBLUP from five genotypic error rates scenarios (0%, 1.0%, 3.0%, 5.0% and 7.0%). Genotypic errors may cause a decrease in variation in genomic relationships, bias in the estimation of breeding values and decrease in the accuracy of genomic prediction as higher the error rate is.

Keywords: Accuracy, cattle, genomic selection, SNP, ssGBLUP

Introdução

A seleção genômica (SG) transformou profundamente muitos programas de melhoramento genético ao redor do mundo. Com o método, pois com a metodologia foi possível inserir informações de marcadores moleculares às estimativas de parentesco e covariância genética entre indivíduos (Meuwissen et al., 2001) resultando em melhorias na acurácia de predição dos animais jovens e sem progênie. A aplicação da informação genômica requer que genótipos densos sejam obtidos para formação de grupos de referência (animais com informação fenotípica e de marcadores) para avaliação dos candidatos à seleção, por isso, milhares de animais são genotipados anualmente (Ventura et al., 2019), o que gera custos financeiros consideráveis. Na obtenção da matriz de parentesco realizado (Legarra et al., 2009), na primeira etapa da seleção genômica, são estimados os efeitos dos marcadores genéticos bialélicos, do tipo polimorfismos de nucleotídeo único (SNP), e a qualidade dessa informação pode ter muitas consequências nas etapas seguintes (Pimentel et al., 2015). A adoção dos métodos moleculares tornou mais evidente a distância genética entre animais aparentados, com o possível indicativo de erros de pedigree ou de genotipagem, quando os valores desviam expressivamente da média esperada. Considerando os erros de genotipagem, nem todas as suas origens estão claras, ou são inevitáveis, com



taxas muitas vezes desconhecidas. O crescimento das coletas de dados e do uso de painéis de marcadores do tipo SNP distintos agravam esse problema, inclusive, a adoção da imputação dos genótipos de menor para maior densidade (Ventura et al., 2019), sendo estes os principais fatores da ocorrência de erros genotípicos e inconsistências mendelianas em bancos de dados. Portanto, o objetivo deste trabalho foi avaliar o efeito de diferentes taxas de erros genotípicos sobre viés e acurácia dos valores genéticos genômicos obtidos via ssGBLUP, utilizando dados simulados em bovinos de corte.

Material e Métodos

As populações foram simuladas utilizando o software QMSim versão 2.0, resultando em um total de 15 gerações, das quais as 5 gerações mais recentes (50.000 animais) foram analisadas. Genótipos para 6.000 indivíduos, contendo 600.000 marcadores do tipo SNP e 1000 QTL distribuídos entre os 29 autossomos (40 cM a 146 cM de comprimento) foram simulados. Todo o processo de simulação foi realizado em 5 replicatas e considerando uma característica com herdabilidade de 0,32 e variância fenotípica de 1,0. A taxa de substituição de touros e matrizes foi mantida constante ao longo das 15 gerações em 60% e 20%, respectivamente, e o critério de descarte foi o menor valor genético estimado (EBV), obtido via método BLUP (*Best Linear Unbiased Prediction*). O valor genético verdadeiro (TBV) 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. Cinco taxas de erros genotípicos (0%, 1,0%, 3,0%, 5,0% e 7,0%) foram simuladas nos genótipos simulados após o controle de qualidade dos dados. Os erros foram inseridos nas posições de marcadores em que, o alelo de menor frequência calculado fosse presente (MAF, < 0,45), foi feita a substituição do genótipo homocigoto pelo heterocigoto ou pelo homocigoto oposto, na mesma proporção, de acordo com as taxas de erros desejadas, que foram aplicadas por linha do arquivo, ou seja, em cada animal. Foram considerados os 6.000 genótipos aleatórios de animais pertencentes às três gerações simuladas mais recentes, via ssGBLUP (*single-step Genomic Best Linear Unbiased Prediction*).

Resultados e Discussão

A simulação teve por objetivo mimetizar algoritmos de imputação, que normalmente sugerem o haplótipo mais frequente durante a imputação (Pimentel et al., 2015). As taxas de erros simuladas correspondem à diferença entre 1 e o valor de acurácia de imputação, considerado como percentual de genótipos imputados corretamente, e são semelhantes aos observados em estudos com dados reais (Carvalho et al., 2014). Foram comparados na Tabela 1 os efeitos dos erros genotípicos na avaliação genética de dois grupos, um com todos os animais genotipados e outro com apenas 100 indivíduos superiores, considerando que estes são mais importantes para a seleção e melhoramento, e normalmente deixam mais descendentes. O erro no genótipo não causou alteração na estimação da herdabilidade, com valores de 0,27 e 0,26, na ausência e presença de 7,0% de erros, respectivamente.

Entre todos os indivíduos, o desvio padrão dos elementos fora da diagonal da matriz genômica foram 0,042, 0,039, 0,034, 0,030 e 0,028 para as taxas de erros de 0%, 1,0%, 3,0%, 5,0% e 7,0%, respectivamente, indicando que a presença de erros genotípicos pode causar redução da variação entre os relacionamentos genômicos. No grupo dos 100 melhores animais observamos diminuição de 0,400 para 0,333 no valor de correlação de ranking e de 0,437 para 0,345 no valor de correlação entre TBV e GEBV, causados pela taxa de 7,0% de erros. A alteração mais acentuada para esse grupo foi sobre o valor da acurácia BIF de predição (Item 5 da Tabela 1), que diminuiu de 45,41% para 34,28% na presença da maior taxa de erros, uma diferença de 11,13 pontos percentuais. De acordo com Pimentel et al. (2015) bons candidatos à seleção podem ser penalizados durante a avaliação se o genótipo contiver erros genotípicos semelhantes a estes.

Apesar das altas correlações dentro do grupo de 6000 animais, é notável a tendência linear dos efeitos das taxas de erros de genotipagem nas análises dos 100 melhores animais. A inclinação do coeficiente de regressão do TBV sobre o GEBV foi calculada para analisar o viés no GEBV, e observamos tendência de subestimação para os 100 animais superiores, em todos os níveis de erros. 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.

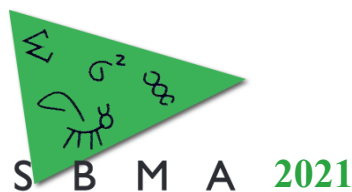


Tabela 1. Mudanças na classificação de ranqueamento, na correlação entre TBV e GEBV, na regressão do TBV sobre o GEBV e acurácia de predição dos animais genotipados

Grupos	¹ Taxa de erro	² ρ Ranking	³ r(TBV,GEBV)	⁴ b(TBV,EBV)	⁵ ACC (%)
⁶ G1	0%	0,400	0,437	0,301	45,41
	1,0%	0,382	0,427	0,293	42,13
	3,0%	0,352	0,395	0,267	37,42
	5,0%	0,341	0,369	0,248	34,76
	7,0%	0,333	0,345	0,227	34,28
⁷ G2	0%	0,845	0,850	1,133	43,48
	1,0%	0,844	0,849	1,134	40,48
	3,0%	0,838	0,844	1,130	36,10
	5,0%	0,830	0,840	1,123	33,54
	7,0%	0,820	0,830	1,120	32,84

¹Erros genotípicos; ²Coeficiente de correlação de Spearman entre TBV e GEBV, classificados de acordo com o TBV; ³Correlação de Pearson entre TBV e GEBV; ⁴Coeficiente de regressão do TBV sobre o GEBV; ⁵ACC: acurácia BIF do valor genético genômico; ⁶G1: Grupo dos 100 melhores animais genotipados, classificados de acordo com o TBV; ⁷G2: Grupo de 6,000 animais genotipados.

Conclusão

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.

Agradecimentos

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 pela concessão da bolsa de estudo de Mestrado. Ao grupo de pesquisa do Laboratório de Melhoramento Animal e Biotecnologia “Dr. Gordon Dickerson” da FZEA USP. Ao Dr. Mehdi Sargolzaei por conceder o software QMSim v2. Ao Dr. Mario Calus pelo conselho.

Literatura citada

Carvalho 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**, v. 46, n. 1, p. 1-11, 2014.

Legarra A.; Aguilar I.; Misztal I. A relationship matrix including full pedigree and genomic information. **Journal of dairy science**, v. 92, n. 9, p. 4656-4663, 2009.

Meuwissen T.H.E.; Hayes B.J.; Goddard M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.

Pimentel E.C.G.; Edel C.; Emmerling R.; Götz K.U. How imputation errors bias genomic predictions. **Journal of dairy science**, v. 98, n. 6, p. 4131-4138, 2015.

Ventura R.V.; Brito L.F.; Oliveira G.A.; Daetwyler H.D.; Schenkel F.S.; Sargolzaei M.; Vandervoort G.; Fonseca E Silva F.; Miller S.P.; Carvalho M.E.; Santana M.H.A.; Mattos E.C.; Fonseca P.; Eler J.P.; Ferraz J.B.S. A comprehensive comparison of high-density SNP panels and an alternative ultra-high-density panel for genomic analyses in Nelore cattle. **Animal Production Science**, v. 60, n. 3, p. 333-346, 2019.