VINICIUS SILVA JUNQUEIRA

GENOMIC INFORMATION FOR BREED DETERMINATION, MULTIBREED EVALUATION, AND ESTIMATION OF VARIANCE COMPONENTS IN LARGE POPULATIONS

Thesis submitted to the Breeding and Genetics Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRAZIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
J95g
2018

Junqueira, Vinícius Silva, 1985-
Genomic information for breed determination, multibreed evaluation, and estimation of variance components in large populations / Vinícius Silva Junqueira. – Viçosa, MG, 2018.
xii, 81 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Paulo Sávio Lopes.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Bovinos de corte - Genética. 2. Algoritmos genéticos. 3. Variação genética. 4. Polimorfismo de nucleotídeo único. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 636.20821

VINÍCIUS SILVA JUNQUEIRA

GENOMIC INFORMATION FOR BREED DETERMINATION, MULTIBREED EVALUATION, AND ESTIMATION OF VARIANCE COMPONENTS IN LARGE POPULATIONS

Thesis submitted to the Breeding and Genetics Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

APROVADA: 4 de junho de 2018.

Fernando Flores Cardoso
(Coorientador)

Fabyano Fonseca e Silva
(Coorientador)

Renata Veroneze

Daniela A. Lino

Renata Veroneze

Daniela Andressa Lino Lourenço

Paulo Sávio Lopes
(Orientador)

“Education is the most powerful weapon

that you can use to change the world”

Nelson Mandela

I dedicate this work to my parents

Gilson and Lucilene,

“You may never know what results come of your actions, but if you do nothing, there will be no results”

Mahatma Ghandi

ACKNOWLEDGMENTS

The Universidade Federal de Viçosa (UFV), especially the Genetics and Breeding Graduate Program and the Department of Animal Science, for the opportunity of carrying out the course;

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support;

Embrapa Pecuária Sul and University of Georgia for theoretical and technical support;

Conexão Delta G for providing the data and to make this work possible;

Professor Dr. Paulo Sávio Lopes, my advisor, for his valuable supervision, teaching, patience, friendship, advice and opportunities given to me;

My co-advisor Dr. Fernando Flores Cardoso for his supervision, teaching and friendship;

My co-advisor professor Dr. Fabyano Fonseca e Silva for his teaching, friendship, advice, opportunities, and friendship;

Professor Dr. Marcos Deon Vilela de Resende for his support, teaching and friendship;

Professor Dr. Robledo de Almeida Torres for friendship and teaching;

Researcher Dr. Marcos Jun it Yokoo for his helps, teaching, friendship, and support;

Researcher Dra. Bruna Pena Sollero for her support, teaching, friendship, and support;

Professor Daniela Andressa Lino Lourenço and Ignacy Misztal for giving me the opportunity to work as Research Fellow at University of Georgia, and to provide all support needed to live an amazing life experience with the group;

Researcher Yutaka Masuda for all friendship, teaching and for taking me through a higher level on Fortran programming;

Researcher Shogo Tsuruta for all conversations about hobbies, programming, and cultural language differences;

Professors, technicians and employees from the Genetics and Breeding Graduate Program and the Animal Science Department;

My family, Gilson Carlos Leite Junqueira, Lucilene Silva Junqueira for caring, support, comfort and also for believing in me all my life;

My sister and her husband, Andressa Junqueira Osorio and Luis Felipe Baumotte Osorio, for encouragement, support and good conversations;

My fiancée and future wife, Luciene Alves Pereira, for all love, keep believing in me and for all patience during graduate school;

My friends from UFV, Nadson, Leonardo Peixoto, Leonardo Bhering, Ali, Aline, André Mauric, Bruno, Daniele, Edson, Felipe, Giovani, Hinayah, Jeferson, Joashlleny, Lais, Leonardo Glória, Lucas, Luciano, Matilde, Renata, Rogério, Sara Coser, Janeo, Lidiane, for good times, jokes, pleasant workplace and for their friendship;

My friends from Bagé, Bruno, Ândrea, Rodrigo Azambuja, Rodrigo Costa, Patrícia, Robert, Rafael, Patrícia and Luiza;

My friends from United States Breno, Ivan, Heather, Dennis and Pete for good conversations at UGA;

For three special people – Corrie Brown, Jefão and Jeff Thompson – that were so present in my life during the time in Athens.

God to make it possible;

To everyone who directly or indirectly contributed to this work.

BIOGRAPHY

VINICIUS SILVA JUNQUEIRA, son of Gilson Carlos Leite Junqueira and Lucilene Silva Junqueira, was born in November 26th, 1985 in Brasília, Distrito Federal, Brazil.

In March, 2006, he started his undergraduate school in Veterinary Medicine in Agronomy and Veterinary Medicine at University of Brasilia (UnB). For almost all the five years in there he dedicated to have a deep understanding on infectious diseases and surgery. In May, 2011, he has completed the undergraduate school at UnB.

During 2011 and the front end of 2012 he professionally focused on surgery of large animals and to develop and implement management and financial guidelines for beef and dairy cattle farms.

In August, 2012, he started his graduate program at Genetics and Breeding Graduate Program by UFV to obtain his degree of Magister Scientiae in Genetics and Breeding. He presented his dissertation in July 25th, 2014.

In August, 2014, started his PhD at Genetics and Breeding Graduate Program by UFV, to obtain his degree of Doctor Scientiae in Genetics and Breeding presenting his thesis in June 4th, 2018.

SUMMARY

ABSTRACT

JUNQUEIRA, Vinícius Silva, D.Sc., Universidade Federal de Viçosa, June, 2018. **Genomic information for breed determination, multibreed evaluation, and estimation of variance components in large populations**. Advisor: Paulo Sávio Lopes. Co-advisors: Fernando Flores Cardoso and Fabyano Fonseca e Siva.

The knowledge on breed composition is of major importance under design of breeding schemes. With this respect, the estimation of such parameters must be as accurately as possible. Currently, most of genetic evaluation programs has been predicting breed composition based on pedigree datasets; but, such estimations only accounts for the expected (allele frequency) contributions across ancestors After the development and establishment of single nucleotide polymorphism (SNP) genotyping platforms on the last decade, an interest in genetic diversity studies has arisen and especially the study of individuals' origin. The objective of the present study was to evaluate the minimum required number of ancestry informative markers necessary to differentiate Hereford, Nelore, Brahman and Braford breeds genotyped with 777 K Illumina Bovine HD Bead Chip. In addition, we also compared the effects of different panels size on breed composition inference under different AIMs methods. To that, it was used the high-density Illumina Bovine HD BeadChip with more than 777 K SNPs to elucidate the structure of Hereford, Nelore, Brahman and Braford populations. Three different ancestry informative marker methods were used to distinguish such populations. Additionally, random marker selection was considered. Admixture software was used to infer breed composition using very low-density SNP panels assembled with AIMs. Our results suggest that is possible to assign individuals to populations with high confidence using less than 8 SNP markers selected per breed. Although millions of SNP markers have been identified, only few of them are needed to accurately infer ancestry in a cost-effective manner. Pedigree information is by nature incomplete and commonly not well established simply because many of the true genetic ties existent between individuals are not a priori known or they can be even wrong. Genomic era brought new opportunities when calculating relationships between individuals. The challenge under genomic approaches is the correct definition of genetic base by the use of pedigree and

genomic data. Genetic base may change as more individuals are included and are inadequately defined if populations are genetically structured. Metafounder concept relies on the definition of pseudo-individuals that describes some level of within and/or across genetic relationship between base population. The purpose of this study was to evaluate metafounder theory to estimate breeding values and the predictive ability under a single-step approach for a multibreed population. Three different scenarios were adopted to estimate variance components and to compute breeding values: pedigree-based model, single-step GBLUP and single-step GBLUP with addition of metafounders. A total of 28 different metafounders were included in the ssGBLUP+metafounder model. In general, it was possible to note that genomic models were able to greater ability to predict the future performance. Among genomic models, the inclusion of metafounder information could increment even more the predictive ability under cross-validation approach. Restricted maximum likelihood (REML) is a popular method for parameter estimation. Because it uses the mixed model equations, it is resistant to selection bias and efficient implementations are currently available. When genomic information is available, two versions of REML may be applicable. When only genotyped animals have phenotypes, genomic REML can be applied with a genomic relationship matrix. When only a fraction of animals is genotyped, a single-step REML is applicable. In general, it is of interest to include many genotyped animals in parameter estimation and into evaluations, to account for genomic selection or pre-selection. The aim of this study was to investigate to what extent generations truncation affects estimates for a simulated population under selection. The use of less generations reduced the ability of pedigree-based model in estimating the benchmark heritability (0.30). The decrease in heritabilities based on genomic information was less than using only pedigree relationships. Genomic models provided greater correlations than pedigree-based model; on average 25 points. Single-step genomic models do not require a deeper pedigree relationship to estimate reliable variance components and breeding values. The use of APY algorithm does not affect the estimation of variance components. An extra of 2 ungenotyped generations are sufficient to compute reliable variance components; as well as breeding values and accuracies.

RESUMO

JUNQUEIRA, Vinícius Silva, D.Sc., Universidade Federal de Viçosa, junho de 2018. **Informações genômicas para determinação racial, avaliação genética multirracial e estimação de componentes de variância em grandes populações**. Orientador: Paulo Sávio Lopes. Coorientadores: Fernando Flores Cardoso e Fabyano Fonseca e Silva.

A disponibilidade de uso de informações genômicas trouxe grandes oportunidades de aumento do ganho genético em sistemas produtivos de gado de corte. Apesar dos benefícios já conhecidos, implementação em larga escala nas condições nacionais ainda é um grande desafio principalmente pelo relativo alto custo de genotipagem. Uma alternativa economicamente viável é o desenvolvimento de painéis de marcadores SNP customizados para objetivos de melhoramento estrategicamente estabelecidos para características de interesse. A implementação dessa proposta tem maior impacto para os animais jovens. O objetivo desse estudo foi identificar o menor número necessário de marcadores do tipo SNP para diferenciar animais das raças Hereford, Nelore, Brahman e Braford genotipados com o painel 777K chip HD para bovinos. Adicionalmente, comparou-se o impacto na predição da proporção racial utilizando-se diferentes painéis reduzidos de marcadores do tipo SNP. Para isso, foram utilizados quatro diferentes métodos para a seleção de marcadores altamente informativos para a diferenciação racial. O software Admixture foi utilizado para os cálculos de proporção racial utilizando os painéis customizados. Os resultados observados nesse estudo sugerem a possibilidade de definir indivíduos às respectivas raças utilizando um painel de 24 marcadores do tipo SNP (isto é, 8 marcadores por raça pura). Informações de pedigree são por natureza incompletas e comumente não são bem definidas porque varias das ligações genéticas existentes não são conhecidas. A genômica trouxe grandes oportunidades para o cálculo do parentesco entre os indivíduos de uma população. Um dos principais desafios em implementações genômicas é a correta definição da população referência para o uso simultâneo das informações de pedigree e genômica. O conceito de metafundadores é baseado na definição de pseudo-indivíduos que descrevem os relacionamentos entre e dentre os indivíduos da população base. O objetivo

desse estudo foi avaliar os impactos do uso de metafundadores ao estimar valores genéticos e sua habilidade preditiva utilizando a metodologia single-step GBLUP (ssGBLUP) em uma população multirracial. Três diferentes cenários foram adotados nesse estudo para a estimação de componentes de variância e predição dos valores genéticos: BLUP tradicional, ssGBLUP e ssGBLUP com inclusão de metafundadores. Um total de 28 metafundadroes foram definidos no modelo ssGBLUP+metafundadores. De forma geral, os modelos genômicos apresentaram maior habilidade preditiva. Sendo o modelo com inclusão de metafundadores o que apresentou maior habilidade preditiva. O método da máxima verossimilhança restrita (REML) é um método comumente utilizado para a estimação de componentes de variância. Por ser implementado em modelos mistos, apresenta estimativas corrigidas para efeitos de seleção. De forma geral, todos os animais genotipados são utilizados nos cálculos para a predição dos valores genéticos. O objetivo desse estudo foi avaliar quantas gerações são necessárias para acurada estimação de componentes de variância com o algoritmo para animais provados e jovens (APY) em uma população simulada com restrições de seleção. O uso de menor número de gerações reduziu a habilidade do modelo BLUP em estimar a herdabilidade simulada (0.30). A redução na estimação da herdabilidade pelos modelos genomicos são menores do que os modelos baseados em informações de pedigree. Os modelos genômicos apresentaram em média maior correlação que os modelo BLUP. Os resultados desse estudo sugerem que não é necessário grande número de gerações para acurada estimação dos componentes de variância e dos valores genéticos. O algoritmo APY não afeta a estimação dos componentes de variância. Duas gerações extras de animais não genotipados são suficientes para acurado cálculo dos componentes de variância, valores genéticos e também acurácia de predição dos valores genéticos.

GENERAL INTRODUCTION

Brazil plays as one of the most important worldwide producers of animal protein for human consumption. The country already has the largest commercial cattle herd of the world. Several initiatives are in place to augment the income by increasing the volume of exported products (GOMES, 2017). Simultaneously, there is an effort to create an international safety standard and to continue the improvement of traits of economic importance.

The second effort is the kind of initiatives leaded by breeding which is focused to continuously improve traits of economic importance and to develop design breeding schemes with potential to increase productivity. Traditional breeding schemes have been used and was undoubtedly successful for many decades. Animal breeding has achieved genetic gains by estimating the genetic merit of selection candidates based on phenotype and pedigree information (HENDERSON, 1973; SCHAEFFER, 2006). However, there is a need to increase the rate of genetic gain. Traditional schemes may limit the annual genetic progress due to high cost and time taken to identify superior animals.

More recently, developments in high-throughput genotyping platforms have allowed scientists and breeders to design strategies for long-term genetic gain at a reduced cost and time. The development of molecular markers opened up breeding to new opportunities. Initial insights with the use of genomic data have been focused only on computation of more reliable genetic merit information. This effort provided a tremendous gain when selecting the

parents of next generation. However, the complete potential for the use of molecular markers is yet to come. There is still a big challenge under broad genomic implementation. Genotyping cost still is one of the main limiting factors for broad implementation of genomic models in commercial herds. A big impact and cost-efficient alternative would be the adoption of customs marker panels (HUANG et al., 2012). The main idea for customization of SNP panels relies on decreasing of genotyping costs by selection of single nucleotide polymorphism (SNP) markers which includes only those informative markers for target objectives. Breeding programs would be able to re-design scheme strategies to include even more genotyped animals. Indeed, this may support the selection of animals upon young ages, consequently, it is expected to see an increase on annual genetic gain rates.

Following the breeder's interest in crossbreeding, BLUP models in multibreed and admixed evaluations were extended to account for both intrabreed and interbreed additive effects, and non-additive genetic effects such as dominance (LO et al., 1993; LO et al., 1995; LO et al., 1997). Genomic information opened up an unlimited number of possibilities for multibreed genetic evaluation. Now it is possible to evaluate the genetic merit of animals using a more reliable relationship information. Traditionally, such information was being calculated using expectations. Marker data permits the assessment of true alleles shared between animals and breeds. This is of major impact under genetic evaluation programs because mendelian sampling can be directly accessed.

Recently, a new paradigm emerged around genomic evaluations. An increasing availability of marker data is challenging the current methods for

breeding values prediction. The main challenge is related with genomic matrix operations (e.g., factorization, inversion) needed for variance component and accuracy estimation. New algorithms and data manipulation still needs to be developed or improved. Some approaches are already being reported in literature (FERNANDO et al., 2014; MISZTAL et al., 2014; MASUDA et al., 2016; MASUDA et al., 2017). These kinds of approaches are especially important under multi-trait and multibreed genetic evaluation; and they are usually implemented under mixed model equations. It would be straightforward only solve the equations without of genomic data. However, it can be challenging if a large number of genotyped animals are included in genetic evaluations.

The objectives in this study were identify SNP markers applied on breed of origin and breed composition for a multibreed population composed of Hereford and Braford beef cattle, to evaluate the impact of genomic information on multibreed genetic evaluation and, finally, to evaluate variance component estimation under sparse inversion in large populations.

REFERENCES

FERNANDO, R. L.; DEKKERS, J. C. M.; GARRICK, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. **Genetics Selection Evolution**, v. 46, n. 1, p. 50, 2014.

GOMES, J. R. **Brasil deve elevar exportação de carne bovina em 2018 com novos mercados**. Disponível em: < https://br.reuters.com/article/businessNews/idBRKBN1E82EI-OBRBS >. Acesso em: 8 de Maio de 2018.

HENDERSON, C. R. Sire evaluation and genetic trends. **Journal of animal science**, v. 1973, p. 10-41, 1973.

HUANG, Y.; HICKEY, J. M.; CLEVELAND, M. A.; MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, v. 44, n. 1, p. 25, 2012.

LO, L. L.; FERNANDO, R. L.; CANTET, R. J. C.; GROSSMAN, M. Theory for modelling means and covariances in a two-breed population with dominance inheritance. **Theoretical and Applied Genetics**, v. 90, n. 1, p. 49-62, 1995.

LO, L. L.; FERNANDO, R. L.; GROSSMAN, M. Covariance between relatives in multibreed populations: additive model. **Theoretical and Applied Genetics**, v. 87, n. 4, p. 423-430, 1993.

_____. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. **Journal of animal science**, v. 75, n. 11, p. 2877-2884, 1997.

MASUDA, Y.; MISZTAL, I.; LEGARRA, A.; TSURUTA, S.; LOURENCO, D. A. L.; FRAGOMENI, B. O.; AGUILAR, I. Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. **Journal of animal science**, v. 95, n. 1, p. 49-52, 2017.

MASUDA, Y.; MISZTAL, I.; TSURUTA, S.; LEGARRA, A.; AGUILAR, I.; LOURENCO, D. A. L.; FRAGOMENI, B. O.; LAWLOR, T. J. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for

US Holsteins with a large number of genotyped animals. **Journal of Dairy Science**, v. 99, n. 3, p. 1968-1974, 2016.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Using recursion to compute the inverse of the genomic relationship matrix. **Journal of Dairy Science**, v. 97, n. 6, p. 3943-3952, 2014.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 123, n. 4, p. 218-223, 2006.

CHAPTER 1

DEVELOPMENT OF A CUSTOM SNP PANEL TO ASSIGN BREED ORIGIN AND GENETIC COMPOSITION: THE CASE OF BRAFORD COMPOSITE BREED

Abstract: Understating the population structure has an immediate importance amongst areas of genetic. After the development and establishment of single nucleotide polymorphism (SNP) genotyping platforms on the last decade, an interest in genetic diversity studies has arisen and especially the study of individuals' origin. We used the high-density Illumina Bovine HD BeadChip with more than 777 K SNPs to elucidate the structure of Hereford, Nelore, Brahman and Braford populations. Three different ancestry informative markers (AIMs) methods were used to distinguish such populations. Additionally, random marker selection was considered. ADMIXTURE was used to infer breed composition using very low-density SNP panels assembled with AIMs. Our results suggest that is possible to assign individuals to subpopulations with high confidence using less than 8 SNP markers per breed. Under the adopted strategy to select AIMs, model-based ancestry estimation clearly separated the indicines from taurines. Although millions of SNP markers have been identified, only few of them are needed to accurately infer ancestry in a cost-effective manner.

Keywords: admixture, cattle, chromosome, genetic diversity, genomic region

INTRODUCTION

The use of molecular markers in breeding programs is already a reality for the majority of economically important species. Huge investments have been applied on development of more efficient high-throughput genotyping platforms aiming to decrease the genotyping costs. However, genotyping cost still is one of the main limiting factors for broad implementation of genomic models in commercial herds. HUANG et al. (2012) suggests that a cost-efficient alternative would be the adoption of reduced customs marker panels. The idea relies on decreasing of genotyping costs by selecting of single nucleotide polymorphism (SNP) markers for target objectives.

Under livestock conditions, the estimation of breed composition and breed assignment for young animals could be performed by the use of reduced low-cost SNP panel. The knowledge on breed composition is of major importance under design of breeding schemes. With this respect, the estimation of such parameters must be as accurately as possible. Currently, most of genetic evaluation programs has been predicting breed composition based on pedigree datasets; but, such estimations only accounts for the expected (allele frequency) contributions across ancestors (FRKONJA et al., 2012).

In literature is already reported some methods for selection of informative markers (DING et al., 2011; WILKINSON et al., 2011; FRKONJA et al., 2012; HULSEGGE et al., 2013; HWA et al., 2016). These methods are focused on the identification of markers that exhibit large allele frequency differences between populations under study (ROSENBERG et al., 2003). Over the last years some studies have demonstrated that thousands of

individual SNPs distributed throughout the genome may have very large differences in allele frequencies between populations depending on its origin (PRICE et al., 2007). These SNP markers are commonly referred as ancestry informative markers (AIMs). Independently of the chosen method, they intend to identify well distributed informative markers throughout genome are capable to distinguish individuals between subpopulations (MANTA et al., 2013). A variety of reported studies has been revealing that measures such as pairwise Wright's F_{ST} (F_{ST}), Informativeness for Assignment (I_n) and Absolute allele frequency difference (Delta, δ) are to be proven effective and useful to detect ancestry informative markers (CHIANG et al., 2010).

The objective of the present study was to evaluate the minimum required number of AIMs necessary to differentiate Hereford, Nelore, Brahman and Braford breeds genotyped with 777 K Illumina Bovine HD Bead Chip. In addition, we also compared the effects of different panels size on breed composition inference under different AIMs methods.

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because dataset was generated from multiple previous studies.

Genotype data

Brazilian and U.S. samples from Braford (n=124), Nelore (n=21), Brahman (n=30) and Hereford (n=116) were initially collected and genotyped using the 777K SNP Illumina Bovine HD Bead Chip.

Quality control

The quality control (QC) was performed within each progenitor breed (Hereford, Nellore and Brahman) using different criteria for SNPs and for samples. The criteria used to exclude SNPs were Hardy-Weinberg Equilibrium Chi-Square (HWE) (10^{-7}), call rate (CR) (<98%) and SNPs in the same position. In addition, SNP markers that are located in sex chromosomes or with unknown position were also excluded. Minor allele frequency (MAF) was not used as SNP exclusion criteria because we intended to identify informative markers independently of their frequency. In fact, monomorphic markers play an important role due to the possibility to be used to distinguish populations, which is desired for studies that intend to infer breed composition (JUDGE et al., 2017). The QC applied for samples was identical samples (IS) (> 99.5% similarity), call rate (< 90%) and heterozygosity deviation above three standard deviation (HD).

After editing, a total of 598,558; 510,090 and 624,912 SNP markers and 113; 21 and 29 samples remained for Hereford, Nelore and Brahman animals, respectively. Assuming that each breed selected a different amount and a different sort of markers, only SNPs shared between purebreds were kept for further analysis. Therefore, a total of 481,509 markers were selected to compose the full SNP panel used to apply the AIMs methods.

Marker imputation

To obtain a greater number of available genotypes, marker imputation was performed after the quality control check. This step was applied within each breed and the analysis were carried out using the FImpute software version 2.2 (SARGOLZAEI et al., 2011). FImpute uses an overlapping sliding window algorithm to efficiently exploit relationships similarities between target and reference individuals. Assuming the set of selected markers after QC, only Herefords and Braford had missing genotypes, 0.12% and 0.51%, respectively.

Methods to detect informative markers

In this study, we adopted the Absolute allele frequency difference (Delta, δ), the Pairwise Wright's F_{ST} (F_{ST}) and the Informativeness for Assignment (I_n) as AIMs measures. All analysis were performed using TRES software (KAVAKIOTIS et al., 2015).

The cutoff values applied on these statistics have been subjective and vary among the methods and published studies (HALDER et al., 2008). Because of the subjective pattern on the marker selection, we decided to select the top-ranked SNPs instead of evaluating different cutoff thresholds. In addition, we evaluated how often the same set of SNPs was selected among methods and their ability to reproduce consistent results on breed inference. Moreover, the Spearman's rank correlation of top-ranked SNPs between methods was also performed.

After the calculation of marker's score for origin determination, a sort of fifty top-ranked markers were selected within purebred totaling a custom panel of 150 SNPs for each AIM methods. Further constraints were applied in these custom panel aiming to keep only those informative markers (see next sections for more details).

Pairwise Wright's F_{ST}

The F_{ST} measure is used as a parameter for describing the genetic diversity among populations (WEIR and HILL, 2002). The method describes the variance of allele frequency among populations. As consequence of genetic drift and artificial or natural selection, one specific allele could be favored over others. Thereby, such locus possibly might be important to be used to genetically differentiate subpopulations.

$$F_{ST} = \frac{\left(p_{1j} - p_{2j}\right)^2}{\left(p_{1j} + p_{2j}\right)\left[2 - \left(p_{1j} + p_{2j}\right)\right]}$$

here, p_{ij} is the allele frequency of i^{th} population and j^{th} reference allele. The F_{ST} values range between 0 and 1, in which greater values denotes greater genetic differentiation in that chromosome region.

Absolute allele frequency difference (Delta, δ)

Delta (SHRIVER et al., 1997) is one of the most used AIMs measure

and is one of the easiest to be applied. The measure can be easily calculated as $\delta = |p_{11} - p_{21}|$. The method is defined as the absolute difference between allele frequencies of a genomic region. As well as F_{ST} measure, Delta values also ranges between 0 and 1.

Informativeness for Assignment (I_n)

This measure was proposed by ROSENBERG et al. (2003) and has widely been used to detect highly informative ancestry markers across populations (OLIVEIRA et al., 2015).The measure can be defined as follows:

$$I_n = \sum_{j=1}^{N} \left(-p_j \, log_2 \, p_j + \sum_{i=1}^{K} \frac{p_j \, log_2 \, p_{ij}}{K} \right)$$

This formula gives the expected logarithm of the likelihood ratio whose numerator is the likelihood that an allele is assigned to one of the populations and the denominator is the likelihood that an allele is assigned to the average population. Unlike F_{ST} and δ, which varies between 0 and 1, I_n has a possible maximum value defined as $log \, K$, where K is the number of populations.

Random marker selection

There are several published researches reporting random marker selection as a powerful strategy to detect genetic stratification (PRICE et al., 2006). In this sense, random marker selection was also included as scenario

for SNP selection.

Construction of a custom SNP panel

Since one of our objectives is the identification of the minimum number of SNP markers usable to infer breed composition as accurate as possible, we present a simple strategy to compute the marker score. All three adopted methods in this study estimates the informativeness importance of each marker by pairwise comparisons between breeds. However, we are dealing with three purebred populations (Hereford, Nelore and Brahman) and one crossbred population (Braford). In such circumstances, it is possible that marker identification may be impaired if all purebreds are considered as different populations. Then, the loci (i.e., marker) score is assumed as an average of pairwise comparisons in order to produce only one value for each marker. As alternative, AIM identification was performed assuming a multi-step approach, in which those informative markers within each purebred population were scored at a time. For example, to obtain the ancestry informative markers for Hereford it was assumed that Nelore and Brahman are a unique population. A similar procedure was considered to select markers from Nelore and Brahman breeds. Since Braford breed results from crossbreeding between Hereford, Nelore and Brahman, the identification of markers that are capable to distinguish purebreds would provide a breed composition estimation.

Linkage disequilibrium constraint

The exclusion of redundant SNP markers based on linkage disequilibrium (LD) was performed after the calculation of the AIMs methods. LD analysis was done considering only a very low-density SNP panel of 150 most informative markers for each AIMs method. For this purpose, we adopted the r^2 statistical measure calculated using PLINK software version 1.07 (PURCELL et al., 2007). The parameters used were *--indep-pairwise 50 5 0.5*, in which the first two are the window size in terms of number of SNPs and the number of SNPs to shift the window at each step; and the third parameter represents the r^2 threshold. For those SNP pairs that exhibited strong association (> 0.5), only the marker with lowest MAF was kept for further analysis. As result, it remained 125, 121 and 120 highly informative markers to compose F_{ST}, δ and I_n panels, respectively.

Individual assignment analysis

There are several approaches that can be used to evaluate genetic assignment (PAETKAU et al., 1995; RANNALA and MOUNTAIN, 1997; CORNUET et al., 1999). In this study, an R code was developed to implement the method proposed by PAETKAU et al. (1995), which determines how likely an individual's genotype may be originated from the population in which it was sampled. Let p_{ijk} denote the frequency of the k^{th} allele ($k = 1, 2$) at the j^{th} locus ($j = 1, \dots, J$) in the i^{th} population ($i = 1, \dots, I$). Let $g_{jkk'}$ denote a diploid genotype and let the Mendelian transmission probability of $g_{jkk'}$ arising in the

i^{th} population be

$$T(g_{jkk'}|i) = \begin{cases} p_{ijk}^2 & if\ k = k' \\ 2p_{ijk}p_{ijk'} & if\ k \neq k' \end{cases}$$

where a genotype is homozygous if $k = k'$ or heterozygous otherwise, under the assumption of random union of gametes. Next, let *g* denote an multilocus genotype. The likelihood of a diploid genotype occurring in a particular population, $T(g|i)$, was estimated as above. Under the assumption of independence between the *J* loci

$$log_{10}(T(g|i)) = \sum_j log_{10}(T(g_{jkk'}|i))$$

To assess the performance of the breed assignment procedure, log-likelihood ratios (LLR) were calculated as the log-likelihood difference of the population of origin and the others population as

$$LLR = log_{10}[T(g|i_A)] - log_{10}[T(g|i_B)]$$

Different stringency thresholds were applied as confidence levels of assignment precision. The four stringency levels adopted were LLR > 0, LLR > 1, LLR > 2 and LLR > 3 (WILKINSON et al., 2011; HULSEGGE et al., 2013). The correct assignment of an individual genotype to its known origin occurred when the calculated LLR was greater than the selected stringency level for all populations. If the LLR was lower than the selected stringency level at least in

one comparison, the animal failed to be assigned to its breed of origin.

Breed composition inference

After the identification of the most informative SNPs through the described AIMs methods, we evaluated the breed composition prediction ability using the custom panel. For this purpose, different quantities of top-ranked SNP markers were used until reach the total number of markers included in the custom panel. In this step, we used the Admixture software (ALEXANDER and LANGE, 2011), which is an ancestry estimation computational tool.

RESULTS

Comparison of the AIMs methods

Histograms of the average estimates of genetic information contained in each marker are presented for each method in Figure 1. It was observed a different distribution amongst F_{ST} and I_n showing a large amount of SNP markers with zero values (i.e., non-informative markers). Despite δ scores reflect a more symmetric shape than the other measures, all methods' distribution demonstrates some degree of asymmetry (positive-skewness trend). We noted that, independently of the method, most of SNPs contained low to medium levels estimates of genetic information. This could be accessed by the observed mean (median) ancestry informative scores of 0.23 (0.21), 0.34 (0.34) and 0.16 (0.14) from F_{ST}, δ and I_n methods, respectively.

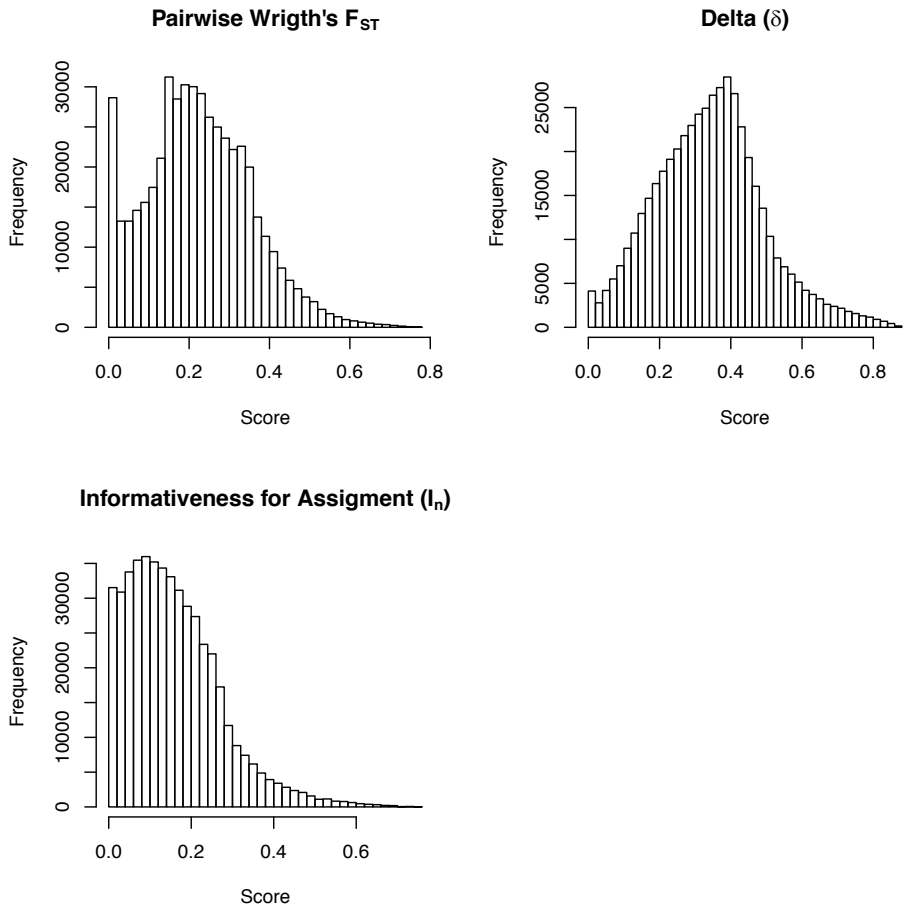

Figure 1. Histograms of the average estimates of genetic information contained in each SNP marker for the pairwise Wright's FST, Delta (δ) and Informativeness for Assignment (In) methods.

Most of 125 top-ranked markers overlapped across methods (102) and the greatest level of Spearman's rank correlation (1.00) were observed between δ and I_n (Table 1). The number of shared markers between F_{ST} and δ, and between F_{ST} and I_n were, respectively,79.20% and 80% of the total.

Table 1. Number of overlapping markers (upper-triangle) and the Spearman's rank correlation (lower-triangle) between the SNP scores under different ancestry informative markers measures.

	F_{ST}	Delta	I_n
F_{ST}		99	100
Delta	0.935		102
I_n	0.933	1.000	

F_{ST} = Pairwise Wright's F_{ST}; Delta = Absolute allele frequency difference (δ); I_n = Informativeness for Assignment.

Genotypic frequencies under different marker selection strategy

It was decided to present only the results of the first top 38 markers, because the addition of more markers did not provide a significant improvement (not shown). These markers were displayed in a scatter-plot for each method (Figure 2). Results of random marker selection are also shown as negative control. This plot shows the most frequent homozygous genotype. For example, if in a specific locus from Hereford the allele A is the most frequent, we used the genotypic frequency of genotype AA to contrast the genotypic frequency of the same genotype from Nellore, Brahman and Braford.

The Figure 2 shows how well AIMs methods were capable to identify the 38 highly informative SNP markers. Considering one purebred at a time, it is possible to note a well-defined behavior to differentiate its genotypic frequency from the others. As expected, Braford genotypic frequencies

behaved fluctuating between the purebreds. A quite different trend was showed for markers randomly selected.

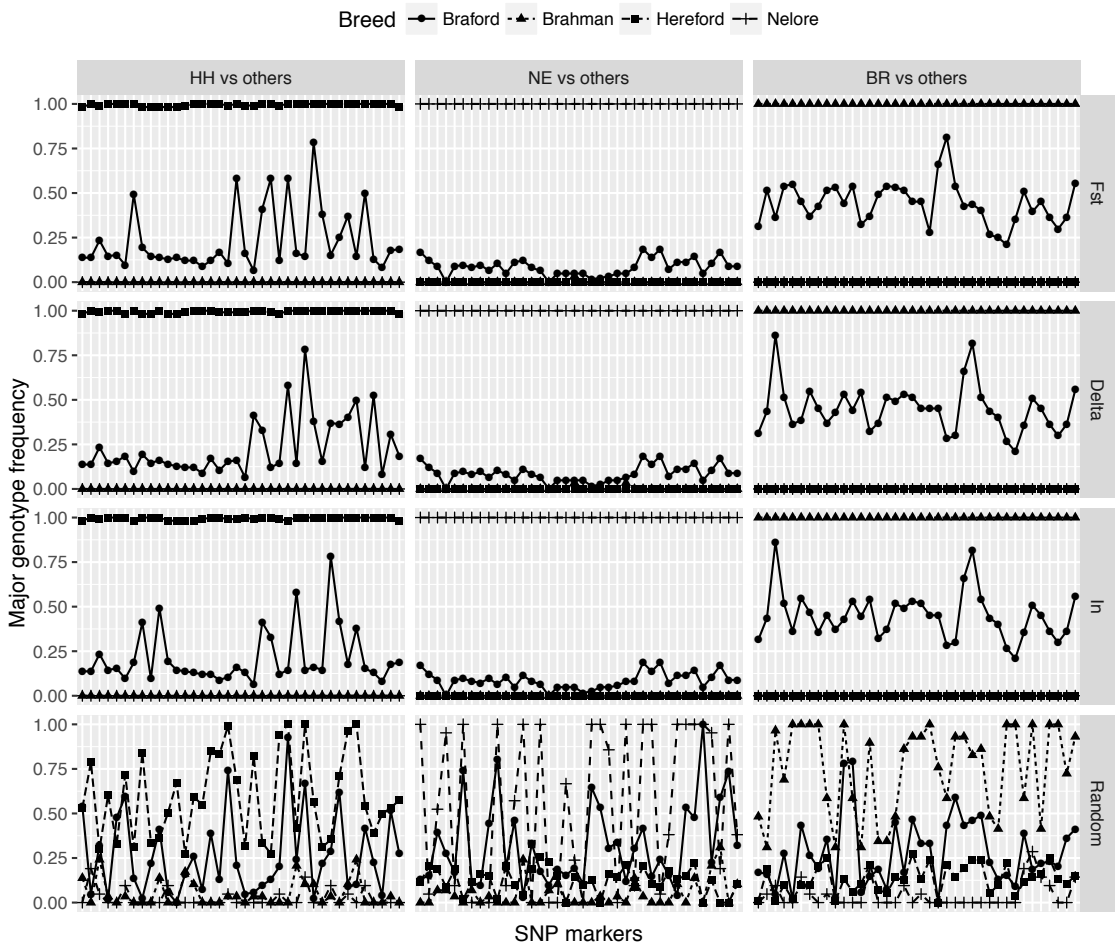

Figure 2. Genotypic frequency between 38 informative markers selected under different methods regarding the ability to distinguish a specific purebred to the other purebreds. HH = Hereford, NE = Nelore, BR = Brahman, Fst = Pairwise Wright's F_{ST}, Delta = Absolute allele frequency difference (δ), In = Informativeness for Assignment, Random = random marker selection

Overall assignment precision evaluation

To evaluate the average percentage of correct assignment to the known

breed of origin, we calculated the LLR by adding, progressively, one marker at a time until reach the number of markers in custom panels. The results of the average correct assignment of 287 individuals are presented in Figure 3. Assuming LLR > 0 as threshold we achieved more than 70% of correct assignment using 2 markers per breed independently of the method. It was possible to accomplish 100% of correct assignment using only 8 highly informative markers with high confidence (LLR > 3). The random marker selection showed the worse performance at the four confidence thresholds. This strategy was not able to deliver more than 95% of correct assignment using 87, 105 and 120 markers at LLR > 0, LLR > 1 and LLR > 2 confidence thresholds, respectively. The I_n measure was the best, it reached 100% of correct assignment using the lowest number of markers (5 markers per purebred) at LLR > 3 (Table 2).

Table 2. Minimum number of SNP markers per purebred required for achieve 100% of breed correct assignment at the four confidence thresholds by each SNP selection method.

	Confidence thresholds			
Method	**LLR0**	**LLR1**	**LLR2**	**LLR3**
F_{ST}	2	3	6	8
Delta	5	6	7	8
I_n	2	3	4	5
Random*	-	-	-	-

F_{ST} = Pairwise Wright's F_{ST}; Delta = Absolute allele frequency difference (δ); I_n = Informativeness for Assignment.

* Random marker selection was not capable to achieve 100% of correct assignment to Hereford (96.46% was the greater observed value under LLR0 confidence threshold). LLR0: LLR > 0; LLR1: LLR > 1; LLR2: LLR > 2; LLR3: LLR > 3.

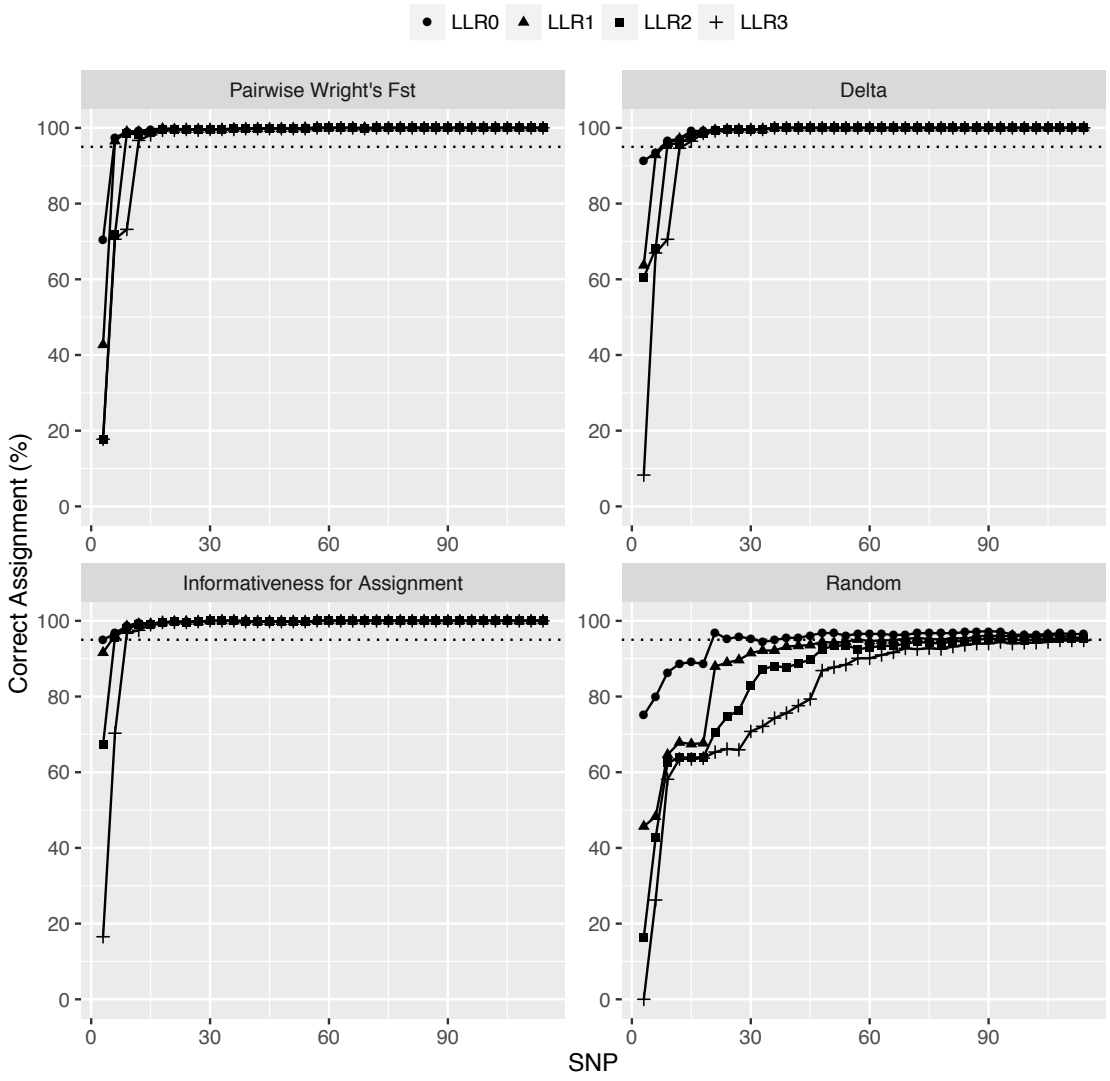

Figure 3. Average percentage of correct assignment considering the top-ranked SNP markers at the four stringency threshold levels for each selection method. LLR0: LLR > 0; LLR1: LLR > 1; LLR2: LLR > 2; LLR3: LLR > 3.

Individual assignment evaluation

The assignment power of the animals to its known breed of origin differed depending on the SNP selection method and number of markers used (Table 3). We noted that it was possible to achieve 100% of correct assignment using less than 8 highly informative markers in a variety of breeds and AIMs methods. Considering the ability to allocate animals to its breed of origin, random marker selection produced the worse performance exhibiting low assignment percentage for both Braford and Hereford. A different trend is shown, and a well-defined pattern is achieved for zebu breeds.

The success on Braford assignment required more markers than other breeds (18 markers per purebred). In general, the I_n measure demonstrated a very similar ability to allocate the animals to its breed compared to δ. Indeed, this is expected because 19 from 24 markers (i.e., very low-density SNP panel) were shared between these methods. Additionally, we noted that 15 SNPs are shared among all methods (Table 4).

Table 3. Assignment power on individual breed of origin at different confidence thresholds considering different methods to choose ancestry informative markers.

Method	**Number of SNPs per breed into panel**	**Confidence threshold**			
		LLR0	**LLR1**	**LLR2**	**LLR3**
Hereford					
	1	98.23	98.23	0.00	0.00
F_{ST}	3	100.00	100.00	98.23	0.00
	6	100.00	100.00	100.00	99.11

	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Delta	1	97.34	0.00	0.00	0.00
	3	99.11	97.34	97.34	0.00
	6	100.00	100.00	99.11	99.11
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
I_n	1	97.34	97.34	0.00	0.00
	3	100.00	100.00	97.34	97.34
	6	100.00	100.00	100.00	100.00
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Random	1	38.93	0.00	0.00	0.00
	3	85.84	5.30	0.00	0.00
	6	87.61	15.04	0.00	0.00
	8	92.03	74.33	27.43	0.88
	10	88.49	80.53	52.21	9.73
	18	88.49	85.84	82.30	68.14
Nelore					
F_{ST}	1	4.76	0.00	0.00	0.00
	3	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Delta	1	100.00	100.00	100.00	0.00
	3	100.00	100.00	100.00	100.00

	6	100.00	100.00	100.00	100.00
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
	1	100.00	100.00	100.00	0.00
	3	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00
I_n	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
	1	100.00	100.00	0.00	0.00
	3	100.00	100.00	100.00	90.47
	6	100.00	100.00	100.00	100.00
Random	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Brahman					
	1	100.00	0.00	0.00	0.00
	3	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00
F_{ST}	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
	1	100.00	100.00	96.55	0.00
	3	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00
Delta	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
I_n	1	100.00	100.00	96.55	0.00

	3	100.00	100.00	100.00	100.00
	6	100.00	100.00	100.00	100.00
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Random	1	100.00	82.75	65.51	0.00
	3	100.00	100.00	96.55	89.65
	6	100.00	100.00	100.00	100.00
	8	100.00	100.00	100.00	100.00
	10	100.00	100.00	100.00	100.00
	18	100.00	100.00	100.00	100.00
Braford					
F_{ST}	1	79.03	72.58	70.96	70.96
	3	95.96	95.96	95.96	92.74
	6	99.19	99.19	98.38	98.38
	8	98.38	98.38	98.38	98.38
	10	98.38	98.38	98.38	98.38
	18	100.00	100.00	100.00	100.00
Delta	1	67.74	54.83	45.16	33.06
	3	87.09	86.29	85.48	82.25
	6	96.77	95.96	95.16	95.16
	8	98.38	98.38	98.38	97.58
	10	98.38	98.38	98.38	98.38
	18	100.00	100.00	100.00	100.00
I_n	1	82.25	72.58	72.58	66.12
	3	94.35	94.35	94.35	89.51
	6	98.38	98.38	98.38	98.38
	8	99.19	99.19	98.38	98.38
	10	100.00	100.00	100.00	100.00
	18	99.19	99.19	99.19	99.19

Random	1	61.29	0.00	0.00	0.00
	3	59.67	53.22	53.22	52.41
	6	66.93	55.64	55.64	54.83
	8	88.70	81.45	70.96	63.70
	10	92.74	85.48	79.03	73.38
	18	95.96	91.93	91.93	85.48

F_{ST} = Pairwise Wright's F_{ST}; Delta = Absolute allele frequency difference (δ); I_n = Informativeness for Assignment. LLR0: LLR > 0; LLR1: LLR > 1; LLR2: LLR > 2; LLR3: LLR > 3.

Table 4. Description of the highly informative SNP markers shared between ancestry informative markers (AIMs) measures.

SNP_Name	**#RefSeq**	**Chr**	**Position**	**Method**		
				F_{ST}	δ	I_n
BovineHD0100005101	rs135136717	1	16899067	1.00	1.00	1.00
BovineHD0300006220	rs133782326	3	19498973	1.00	1.00	1.00
BovineHD0300016882	rs137036969	3	55957304	1.00	1.00	1.00
BovineHD0300029358	rs137604291	3	102530413	1.00	1.00	1.00
BovineHD0400003175	rs136308245	4	10484060	1.00	1.00	1.00
BovineHD0600032730	rs134927227	6	115400094	1.00	1.00	1.00
BovineHD0800024621	rs137239842	8	82702393	1.00	1.00	1.00
BovineHD1000015876	rs134656281	10	53092803	1.00	1.00	1.00
BovineHD1600014567	rs137052668	16	52600950	1.00	1.00	1.00
BovineHD1800004598	rs136175363	18	14084653	1.00	1.00	1.00
BovineHD1900018697	rs137297147	19	35328429	1.00	1.00	1.00
BovineHD2300015477	rs136184268	23	4825558	1.00	1.00	1.00

BovineHD2400000053	rs135903052	24	357370	1.00	1.00	1.00
BovineHD2400017966	rs136386152	24	61834611	1.00	1.00	1.00
BovineHD2500012235	rs135204356	25	37936548	1.00	1.00	1.00

#RefSeq: reference number; Chr: chromosome; F_{ST}: pairwise Wright's F_{ST};δ: absolute allele frequency (Delta);I_n: informativeness for assignment.

Breed composition prediction

Until now, we have discussed the simplest situation when the objective is only the individual identification to its known breed origin. However, a complex scenario emerges in circumstances where markers will be applied for breed composition inference. Such complexity may arise when crossbreds under evaluation are resulting from a recent admixture, likely the case of Braford breed (BLACKBURN et al., 2017). The population structure inferred using a maximum likelihood approach considering 2 and 3 clusters is presented in Figure 4. It is straightforward to see that inferences on breed composition of full genotype panel are quite different from those when using a very low-density SNP panel. The results of full genotypic panel (481,509 markers) suggest that some Hereford animals shares more than 20% of its genetic background with Zebu breeds (K=2). A similar trend is also viewed for all Nellore animals, in which shares genetic content with major part of Hereford animals (K=3).

Independently of the number of populations (K) assumed *a priori*, we observed a similar pattern amongst AIMs methods. Despite the random marker selection was able to differentiate between Hereford and zebu breeds

(data not shown), it was not sufficient to capture the Braford and Hereford composition.

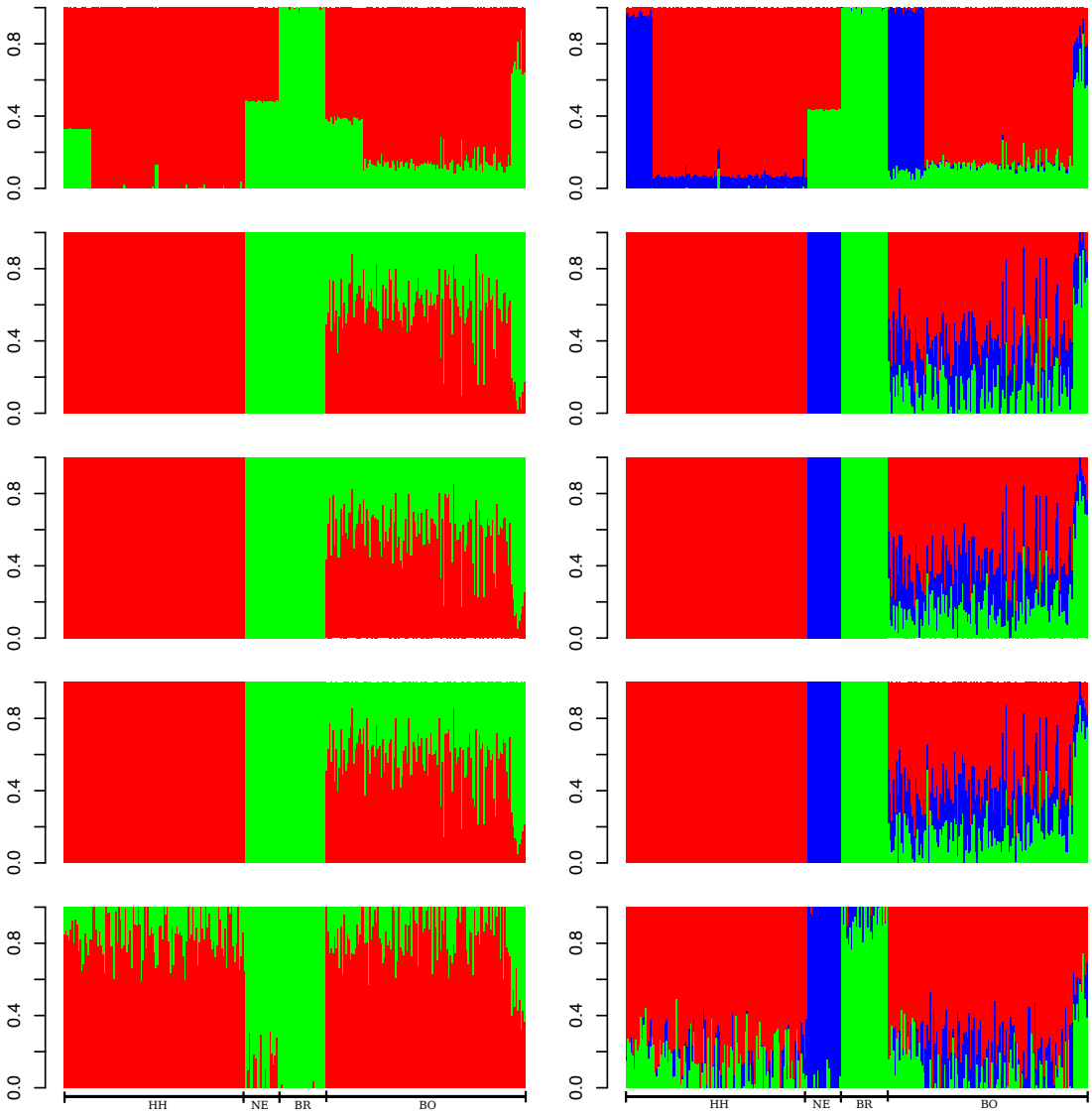

Figure 4. Population structure inferred using a maximum likelihood approach. Individual breed composition predicted using a complete SNP panel (first row) with 481,509 markers and a reduced marker panel with 8 markers per purebred proposed by Pairwise Wright's F_{ST}, Delta, Informativeness for Assignment and random markers selection methods, respectively. The left column presents a scenario assuming K=2 and the right column show the results of K=3. HH = Hereford, NE = Nelore, BR = Brahman and BO = Braford.

DISCUSSION

Our goal was to evaluate the minimum number of markers required to correctly assign Hereford, Nellore, Brahman and Braford animals to its known breed of origin and to evaluate breed inferences using the Ancestry Informative Markers methods. For this purpose, we adopted three widely used methods which aims to select markers that are capable to distinguish populations [Pairwise Wright's F_{ST} (F_{ST}), Delta (δ) and Informativeness for Assignment (I_n)].

Several reported researches elucidated methods to identify breed of origin (ROSENBERG et al., 2003). However, neither of them discussed the effects on the breed inferences specially when dealing with recent composite (admixture populations) breeds.

Comparisons of informative marker selection methods

The high score Spearman's rank correlation between AIMs measures suggests a similar trend amongst methods (Table 1). Such outcome would indicate that adopted methods have selected informative markers. A similar trend was reported by WILKINSON et al. (2011), in which detected high correlation estimates and also a large degree of overlapped markers between F_{ST} and δ.

The majority of the markers are useful to distinguish or to allocate individuals to populations. Despite similar reported outcomes here, in general, it is not well established which method provides the best sort of AIM. However, it is already published that the number of populations under consideration, the respective level of genetic differentiation between populations and the desired

confidence affects the assignment process (BAYE et al., 2011). Indeed, many questions addressed to what method and type of markers should be used to allocate individuals into populations are still unanswered (DING et al., 2011).

Overall assignment quality

The evaluation of the minimum number of markers to effectively assign animals to its known breed of origin was performed by calculating the average log-likelihood ratios (LLR) estimates among breeds within each AIMs measure. The LLR approach was used to evaluate the success on the assignment to breed of origin. As expected, the higher stringency threshold adopted, the higher the number of markers needed to assign individuals at high confidence levels (e.g. levels greater than 95% of assignment success). Thereby, the risk on the false positive decreases when greater confidence thresholds (LLR > 3) are assumed (HULSEGGE et al., 2013). The adoption of this approach as statistical measure of efficiency improves our certainty in situations where it is necessary assignment power evaluation.

The average values of LLR (Figure 3) suggests that the strategy used to identify AIMs demonstrated high ability to set individuals to its true population using only few markers. As expected, marker random selection shown limited and low efficiency to identify breed of origin because the method did not provide 100% of assignment in neither LLR level. One possible explanation is that there are relatively few genomic regions substantially differing among populations (HALDER et al., 2008). In that cases, it would be a difficult task to randomly find informative genomic regions.

Individual breed assignment

Our findings suggest that 8 markers per purebred were needed to reach 100% of assignment power regardless of the method (Table 3). However, using only a panel with 24 markers it was not possible to assign Braford animals with high certainty. This suggests that heterozygosity may affected the assignment panel ability (Figure 2). Differently for purebreds that have most of the selected homozygous genotypes (BAYE et al., 2011). Many factors can affect the success on the assignment process. Most of the human and animal studies (ROSENBERG et al., 2003; WILKINSON et al., 2011; FRKONJA et al., 2012) have been focusing on the identification of individuals' subpopulation of origin for group of individuals who has been an ancient segregation. In such scenarios, it is expected that have had occurred many recombination and mutative events among such individuals. All these biological processes are important on the development of phenotypic patterns particular to breeds or populations. Thereby, it is expected that most reported methods used to estimate global ancestry shows a good ability to identify the most informative markers. However, our study deals with a recent breed, resulted from crossings between Hereford and zebu cattle (e.g. Nellore and/or Brahman). A complex scenario arises for Braford breed because the breed association allows the registration of different breed composition, such as Braford (1/2, 3/4, 5/8, 7/8 and others). Such flexibility on crossing schemes inputs a high level of bias to correctly assign animals to that breed and to infer composition using molecular data. Perhaps, the close genetic relationship between Braford and purebreds is what determined that not all Braford animals had been correctly

set. This was also argued by Ding et al. (2011) when working with a simulated phased HapMap dataset.

It has been reported (Zeng et al., 2015) that most of the assignment issues has been occurring due to the lower level of pairwise genetic differentiation amongst breeds. Thus, breed genetic differentiation levels would act like a success determinant factor when a very low-density SNP panel is designed to set individuals to its true breed. This argument agrees with the lower assignment power to Braford individuals viewed in our results. The close genetic similarity of Braford to the other breeds justifies the lower assignment percentage found. Some authors argue that is possible that some animals will never be correctly assign; even when using a high number of markers, due to the lower genetic differentiation amongst the breeds. This is especially true when such populations results from recent admixture (Qin et al., 2010; Baye, 2011).

Breed composition prediction

Analysis of breed composition were performed using a maximum likelihood approach assuming 2 and 3 cluster as *a priori* knowledge. Breed composition inferred using ADMIXTURE software is shown in Figure 4. It is known that different artificial selection pressure was applied to Hereford and it was differently influenced by natural selection and random drift compared to Nellore or Brahman (O'Brien et al., 2015). This is true because Hereford breed is originated from United Kingdom; and Nellore breed, was originated from India (tropical continent) and for a long time has been genetically improved

under climate Brazilian conditions. Brahman is a breed resulted from crossings of different zebu cattle, in which Nellore was one of the foundation breeds. As a result of population structure inference, it is expected that Hereford animals have much more similarities among themselves than with Nellore or Brahman (O'Brien et al., 2015). Also, it is expected that Brahman may share some alleles with Nellore cattle. Thereby, independently of K assumed, AIMs methods (F_{ST}, δandI_n) shows the expected partition between Hereford and zebu breeds. This can be seen when using K=2 in which Nellore and Brahman are clustered together, and Hereford in the second one. Despite random method have had the ability to suggest that Nellore and Brahman as a unique genetic group, it was not capable to select markers to distinguish Hereford from Braford. This can be seen when evaluating the genetic content inferred of Hereford which shows some degree of genetic sharing with zebu breeds (Figure 4).

One of the most important step to define the panel size is the decision of how many markers should be included to provide reliable estimations about breed composition. Because little difference on Hereford proportion was observed when using panels with from 1 to 37 high informative markers per purebred, we suggest the optimum panel size as 24 markers (8 SNP markers per purebred). This was endorsed following the results of LLR approach in Table 2. Our panel size is similar to those obtained by Hulsegge et al. (2013) while selecting AIMs for different breeds.

There are two different paradigms on the ancestry estimation issues: global ancestry and local ancestry. While local ancestry is related to the identification of chromosome segments belonging different subpopulation,

global ancestry aims to calculate the proportion of ancestry from each contribution population (ALEXANDER and LANGE, 2011). DING et al. (2011) argue that the presence of chromosome (local) population ancestry can be a confounding factor and indeed can lead to global ancestry false positive finds. Future studies should focus on the evaluation of the local ancestry effects when estimating global ancestry aiming to understand the relationship between AIMs methods and chromosome important regions.

CONCLUSION

Except for Braford animals, it was possible to assign individuals to subpopulations with 100% of confidence using less than 8 SNP markers per breed. All AIMs methods shown a very similar performance to assign individuals to subpopulations and predicting breed composition. Our results suggest that random marker selection does not perform well to identify ancestry informative markers and to infer breed composition.

ACKNOWLEDGEMENTS

To CAPES and CNPq for providing scholarship during all Master and Ph.D graduate period.

REFERENCES

ALEXANDER, D. H.; LANGE, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. **BMC bioinformatics**, v. 12, n. 1, p. 246, 2011.

BAYE, T. M.; HE, H.; DING, L.; KUROWSKI, B. G.; ZHANG, X.; MARTIN, L. J. Population structure analysis using rare and common functional variants. In: BMC PROCEEDINGS, 2011, **Anais.** BioMed Central, 2011. p. S8.

BLACKBURN, H. D.; KREHBIEL, B.; ERICSSON, S. A.; WILSON, C.; CAETANO, A. R.; PAIVA, S. R. A fine structure genetic analysis evaluating ecoregional adaptability of a Bos taurus breed (Hereford). **PloS one**, v. 12, n. 5, p. e0176474, 2017.

CHIANG, C. W. K.; GAJDOS, Z. K. Z.; KORN, J. M.; KURUVILLA, F. G.; BUTLER, J. L.; HACKETT, R.; GUIDUCCI, C.; NGUYEN, T. T.; WILKS, R.; FORRESTER, T. Rapid assessment of genetic ancestry in populations of unknown origin by genome-wide genotyping of pooled samples. **PLoS genetics**, v. 6, n. 3, p. e1000866, 2010.

CORNUET, J. M.; PIRY, S.; LUIKART, G.; ESTOUP, A.; SOLIGNAC, M. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. **Genetics**, v. 153, n. 4, p. 1989-2000, 1999.

DING, L.; WIENER, H.; ABEBE, T.; ALTAYE, M.; GO, R. C. P.; KERCSMAR, C.; GRABOWSKI, G.; MARTIN, L. J.; HERSHEY, G. K. K.; CHAKORBORTY, R. Comparison of measures of marker informativeness for ancestry and admixture mapping. **BMC genomics**, v. 12, n. 1, p. 622, 2011.

FRKONJA, A.; GREDLER, B.; SCHNYDER, U.; CURIK, I.; SOELKNER, J. Prediction of breed composition in an admixed cattle population. **Animal genetics**, v. 43, n. 6, p. 696-703, 2012.

HALDER, I.; SHRIVER, M.; THOMAS, M.; FERNANDEZ, J. R.; FRUDAKIS, T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. **Human mutation**, v. 29, n. 5, p. 648-658, 2008.

HUANG, Y.; HICKEY, J. M.; CLEVELAND, M. A.; MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, v. 44, n. 1, p. 25, 2012.

HULSEGGE, B.; CALUS, M.; WINDIG, J.; HOVING-BOLINK, A.; MAURICE-VAN EIJNDHOVEN, M.; HIEMSTRA, S. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. **Journal of animal science**, v. 91, n. 11, p. 5128-5134, 2013.

HWA, H.-L.; WU, L. S. H.; LIN, C.-Y.; HUANG, T.-Y.; YIN, H.-I.; TSENG, L.-H.; LEE, J. C.-I. Genotyping of 75 SNPs using arrays for individual identification in five population groups. **International journal of legal medicine**, v. 130, n. 1, p. 81-89, 2016.

JUDGE, M. M.; KELLEHER, M. M.; KEARNEY, J. F.; SLEATOR, R. D.; BERRY, D. P. Ultra-low-density genotype panels for breed assignment of Angus and Hereford cattle. **animal**, v. 11, n. 6, p. 938-947, 2017.

KAVAKIOTIS, I.; TRIANTAFYLLIDIS, A.; NTELIDOU, D.; ALEXANDRI, P.; MEGENS, H.-J.; CROOIJMANS, R. P.; GROENEN, M. A.; TSOUMAKAS, G.; VLAHAVAS, I. TRES: identification of discriminatory and informative SNPs from population genomic data. **Journal of Heredity**, v. 106, n. 5, p. 672-676, 2015.

MANTA, F. S. N.; PEREIRA, R.; CAIAFA, A.; SILVA, D. A.; GUSMÃO, L.; CARVALHO, E. F. Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. **Annals of human biology**, v. 40, n. 1, p. 94-98, 2013.

OLIVEIRA, R.; RANDI, E.; MATTUCCI, F.; KURUSHIMA, J.; LYONS, L.; ALVES, P. Toward a genome-wide approach for detecting hybrids: informative SNPs to detect introgression between domestic cats and European wildcats (Felis silvestris). **Heredity**, v. 115, n. 3, p. 195, 2015.

PAETKAU, D.; CALVERT, W.; STIRLING, I.; STROBECK, C. Microsatellite analysis of population structure in Canadian polar bears. **Molecular ecology**, v. 4, n. 3, p. 347-354, 1995.

PRICE, A. L.; PATTERSON, N.; YU, F.; COX, D. R.; WALISZEWSKA, A.; MCDONALD, G. J.; TANDON, A.; SCHIRMER, C.; NEUBAUER, J.; BEDOYA, G. A genome wide admixture map for Latino populations. **The American Journal of Human Genetics**, v. 80, n. 6, p. 1024-1036, 2007.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nature genetics**, v. 38, n. 8, p. 904, 2006.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I.; DALY, M. J. PLINK: a tool set for whole-genome association and population-based linkage

analyses. **The American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007.

RANNALA, B.; MOUNTAIN, J. L. Detecting immigration by using multilocus genotypes. **Proceedings of the National Academy of Sciences**, v. 94, n. 17, p. 9197-9201, 1997.

ROSENBERG, N. A.; LI, L. M.; WARD, R.; PRITCHARD, J. K. Informativeness of genetic markers for inference of ancestry. **The American Journal of Human Genetics**, v. 73, n. 6, p. 1402-1422, 2003.

SARGOLZAEI, M.; CHESNAIS, J.; SCHENKEL, F. FImpute-An efficient imputation algorithm for dairy cattle populations. **Journal of Dairy Science**, v. 94, n. 1, p. 421, 2011.

SHRIVER, M. D.; SMITH, M. W.; JIN, L.; MARCINI, A.; AKEY, J. M.; DEKA, R.; FERRELL, R. E. Ethnic-affiliation estimation by use of population-specific DNA markers. **American journal of human genetics**, v. 60, n. 4, p. 957, 1997.

WEIR, B. S.; HILL, W. G. Estimating F-statistics. **Annual review of genetics**, v. 36, n. 1, p. 721-750, 2002.

WILKINSON, S.; WIENER, P.; ARCHIBALD, A. L.; LAW, A.; SCHNABEL, R. D.; MCKAY, S. D.; TAYLOR, J. F.; OGDEN, R. Evaluation of approaches for identifying population informative markers from high density SNP chips. **BMC genetics**, v. 12, n. 1, p. 45, 2011.

CHAPTER 2

A METAFOUNDER THEORY IN GENOMIC PREDICTION APPLIED TO BEEF CATTLE MULTIBREED POPULATION

Abstract: Pedigree information is by nature incomplete and commonly not well established simply because many of the true genetic ties existent between individuals are not a priori known or they can be even wrong. Genomic era brought new opportunities when calculating relationships between individuals. The challenge under genomic approaches is the correct definition of genetic base by the use of pedigree and genomic data. Genetic base may change as more individuals are included and are inadequately defined if populations are genetically structured. Metafounder concept relies on the definition of pseudo-individuals that describes some level of within and/or across genetic relationship between base population. The purpose of this study was to evaluate metafounder theory to estimate breeding values and the predictive ability under a single-step approach for a multibreed population. Three different scenarios were adopted to estimate variance components and to compute breeding values: pedigree-based model, single-step GBLUP and single-step GBLUP with addition of metafounders. A total of 28 different metafounders were included in the ssGBLUP+metafounder model. In general, it was possible to note that genomic models were able to greater ability to predict the future performance. Among genomic models, the inclusion of metafounder information could increment even more the predictive ability under cross-validation approach.

Keywords: breeding, founder, variance components, YAMS

INTRODUCTION

Pedigree information is by nature incomplete (and commonly not well established) simply because many of the true genetic ties existent between individuals are not a priori known or they can be even wrong (JUNQUEIRA et al., 2017). Besides that, pedigree is usually available for several livestock species and has been widely used to improve the reliability of breeding value estimation.

Genomic era brought new opportunities when calculating relationships between individuals, because genomic relationships are independent of pedigree knowledge. Thus, they are not affected by incorrect or incomplete pedigree records over generations. Several genomic methods are published in literature (MEUWISSEN et al., 2001; VANRADEN, 2008; AGUILAR et al., 2010; FERNANDO et al., 2014), but all of them implicit assumes that pedigree structure is absent (CHRISTENSEN, 2012) and the proposals are difficult to extend to several populations (HARRIS and JOHNSON, 2010; MISZTAL et al., 2013). The challenge under genomic approaches is the correct definition of genetic base. Usually the base population is assumed as the current/available sort of individuals. Consequently, the genetic base may change when individuals are included in the database or if populations are genetically structured (HARRIS and JOHNSON, 2010).

CHRISTENSEN (2012) provided some insights on how to create estimations of founder relationship. His suggestions works well when a single population is a priori assumed; however, extend to several founder populations is not straightforward. LEGARRA et al. (2015) reported a metafounder theory to consider relationships within and across founder populations. The paper

provides a generalization of unknown parent groups and Christensen's results. Metafounder concept relies on the definition of pseudo-individuals that add some level of within and/or across genetic relationship between base (i.e., founder) individuals in the population.

The purpose of this study was to test metafounder theory to estimate breeding values and to evaluate its predictive ability under a single-step approach for a multibreed population.

MATERIAL AND METHODS

Phenotype, Genotype and Pedigree information

The data used for investigating the use of metafounders in genomic evaluations was provided by Conexão Delta G Breeding Program (Rio Grande do Sul, Brazil). All Hereford and Braford tick count records were derived from eight herds. Individuals kept in the data file were between 326 and 729 days old at the time of the count. The contemporary groups (CG) were formed by the combination of the effects of farm, sex, year of birth, management group and count date. Records of CG with less than five animals and counts above or below 3.5 standard deviations in comparison with the CG mean were discarded from the data file. After restrictions, 146 contemporary groups remained. The phenotypic file included records of 4,363 animals raised under extensive conditions and pedigree file included 12,755 animals. Of these phenotyped individuals, 2,188 animals had three subsequent tick counts, 1,934 had two counts and 241 had only one count. Therefore, the total number of records was 10,673 related to 2,369 Hereford animals, and 8,304 from

Braford with a maximum of ¾ of Zebu proportion. The heterozygosity effects and recombination loss were calculated as proposed by CARDOSO and TEMPELMAN (2004) and included as linear covariate.

Genotyping of 130 sires was performed using the high-density panel (BovineHD - Illumina bead chip with 777,962 SNPs), while the BovineSNP50 Illumina panel (54,609 SNPs) was used for 3,461 animals. The quality control criteria adopted for SNPs exclusion were the Hardy–Weinberg equilibrium chi-square test ($p = 10^{-7}$), genotype call rate (CR) (<98%), minor allele frequency (MAF) (<3%), near-perfect collinearity with another SNPs ($r > 0.98$) and SNPs in the same position. The criteria adopted to reject samples were CR <90%, heterozygosity deviation above three standard deviations, sex identification errors and identical genotypes between different individuals (more than 99.5% of similarity for all markers). After quality control, a total of 3,528 samples and 39,554 markers were retained for further analysis.

Metafounders relationship

Metafounder information was included in this study based on the methodology proposed by LEGARRA et al. (2015). In the paper, authors defined the pedigree-based matrix modified for populations under different structures (e.g., single and multiple base populations). The concept of metafounder relies on the definition of pseudo-individuals that add some level of within and/or across genetic relationship between base (i.e., founder) individuals in population. The main idea is the assumption that metafounder population have a common ancestral population. To that, authors suggested a

modified relationship matrix (**A**(Γ)) which includes metafounders as individuals. The Γ matrix is composed by the relationship between metafounders. In this study, a total of 28 metafounders were defined both based on breed of origin (i.e., Hereford, Braford and Zebu) or based on year of birth and gender combination. The latter definition was only adopted when breed of origin of base individuals were unknown. Recursive computations of **A**(Γ) follows usual rules (EMIK and TERRILL, 1949; KARIGL, 1981; AGUILAR and MISZTAL, 2008). The only required modification to include metafounders is the assumption of γ as the self-relationship for founders. Note that self-relationship for base animals are usually assumed as zero due to lack of historical information. The Γ matrix needs within- and across-founder relationship was estimated using molecular markers. In this study, generalized least square (GLS) was the method adopted to estimate Γ (GARCIA-BACCINO et al., 2017).

Scenarios

Three different scenarios were tested in this study: pedigree-based model (traditional BLUP), single-step GBLUP (ssGBLUP) and single-step GBLUP with addition of metafounders (ssGBLUP+metafounder). No restrictions were imposed to avoid or minimize inbreeding, as result a total of 130 inbred individuals were found. The maximum inbreeding coefficient found was 25% and an average value across inbred animals of 5.73%.

Aiming to reduce the computational time for variance components estimation in AI-REML, initial guesses of variance components were estimated

by Gibbs sampling using GIBBS2F90 (MISZTAL et al., 2002). A total of 100,000 iterations were generated, with the first 30,000 discarded as burn-in and every 10th sample included in the posterior analysis. Posterior means were used as starting values for AIREMLF90 software (MISZTAL et al., 2002) implemented on YAMS package. Average information REML algorithm achieved convergence after only a few (≤ 7) iterations.

The BLUP model was fitted using the regular relationship matrix constructed based on HENDERSON (1976) rules. Covariance genetic relationship matrix for both ssGBLUP and ssGBLUP+Metafounder were fitted using the approach presented by AGUILAR et al. (2011). The difference between these two last scenarios are restricted to the kind of information used as relationship matrix required to construct **H** matrix. While ssGBLUP adopted the same **A** matrix as BLUP, ssGBLUP+Metafounder fitted $\mathbf{A}(\Gamma)$ as described in the previous topic.

To compare the estimated variance components and genetic parameters between models, ssGBLUP+metafounder parameters needed to be multiplied by $(1-\gamma_i)$, corresponding to (co)variances among the unrelated breed animals (scaled) (LEGARRA et al., 2015). More specifically, the scaled genetic variances of Hereford (Braford) performance were $\sigma^2_{\mathbf{a(c)}}\left(1-{}^{\gamma_{\mathbf{a(c)}}}/_{2}\right)$; the scaled genetic covariance for crossbred performance were $\sigma^2_{\mathbf{ac}}\left(1-{}^{\gamma_{\mathbf{a}}}/_{2}\right)$. Heritabilities were calculated as usual using these scaled (co)variance genetic components. The additive correlations between Hereford and Braford were calculated as $r_a = \frac{\sigma^2_{\mathbf{ac}}\left(1-{}^{\gamma_{\mathbf{a}}}/_{2}\right)}{\sqrt{\sigma^2_{\mathbf{a}}\left(1-{}^{\gamma_{\mathbf{a}}}/_{2}\right)\sigma^2_{\mathbf{c}}\left(1-{}^{\gamma_{\mathbf{a}}}/_{2}\right)}}$. Finally, repeatability for Hereford and

Braford was calculated as $r_H = \frac{\sigma^2_{\mathbf{a}}\left(1-{}^{\gamma_{\mathbf{a}}}/_2\right)+\sigma^2_{pe_H}}{\sigma^2_{p_H}}$ and $r_B = \frac{\sigma^2_{\mathbf{c}}\left(1-{}^{\gamma_{\mathbf{a}}}/_2\right)+\sigma^2_{pe_B}}{\sigma^2_{p_B}}$. For BLUP and ssGBLUP models the same described formulas were used to compute heritabilities, additive correlation and repeatability, but using directly the (co)variances estimated by REML algorithm.

Statistical models

A repeatability bi-trait Mixed Model was used to estimate breeding values. The model can be seen as an incomplete model of WEI and VAN DER WERF (1994). The term incomplete is associated to the fact that phenotypes of Zebu are unknown. The model may be defined as follows

$$\begin{bmatrix}\mathbf{y}_H\\ \mathbf{y}_B\end{bmatrix} = \begin{bmatrix}\mathbf{X}_H & \mathbf{\Phi}\\ \mathbf{\Phi} & \mathbf{X}_B\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_H\\ \boldsymbol{\beta}_B\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_H & \mathbf{\Phi}\\ \mathbf{\Phi} & \mathbf{Z}_B\end{bmatrix}\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix} + \begin{bmatrix}\mathbf{P}_H & \mathbf{\Phi}\\ \mathbf{\Phi} & \mathbf{P}_B\end{bmatrix}\begin{bmatrix}\mathbf{pe}_H\\ \mathbf{pe}_B\end{bmatrix} + \begin{bmatrix}\mathbf{e}_H\\ \mathbf{e}_B\end{bmatrix}$$

here $\mathbf{y}_i$ is the vector of tick counts phenotypes from Hereford (H) and Braford (B) animals; $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{P}_i$ are known matrices that relates phenotypes to its respective fixed, additive and permanent environment effect levels. The vector of fixed effect ($\boldsymbol{\beta}_i$) is composed by an overall mean, contemporary groups; heterozygosity, recombination loss, age at tick counting, quadratic age at tick counting effects were fitted as covariables. Vector of permanent environment effect can be defined as $\mathbf{pe}_i \sim N\left(\mathbf{0}, \mathbf{I}\otimes\begin{bmatrix}\sigma^2_{pe_H} & \sigma_{pe_{HB}}\\ \sigma_{pe_{HB}} & \sigma^2_{pe_B}\end{bmatrix}\right)$; and the residual vector as $\mathbf{e}_i \sim N\left(\mathbf{0}, \mathbf{I}\otimes\begin{bmatrix}\sigma^2_{e_H} & \sigma_{e_{HB}}\\ \sigma_{e_{HB}} & \sigma^2_{e_B}\end{bmatrix}\right)$. Moreover, the vector of additive effects of BLUP

model was defined as $\begin{bmatrix}\mathbf{a}\\ \mathbf{c}\end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{A}^{-1} \otimes \begin{bmatrix}\sigma^2_{\mathbf{a}} & \sigma_{\mathbf{ac}}\\ \sigma_{\mathbf{ac}} & \sigma^2_{\mathbf{c}}\end{bmatrix}\right)$; where $\sigma^2_{\mathbf{a}}$ and $\sigma^2_{\mathbf{c}}$ is the additive variance for Hereford and Braford traits, respectively; and $\sigma_{\mathbf{ac}}$ is the additive covariance between breeds.

For genomic models, the $\mathbf{A}^{-1}$ matrix is replaced by $\mathbf{H}^{-1}$ and $\mathbf{H}(\boldsymbol{\Gamma})^{-1}$. Its additive (co)variance structure can be defined as follows

$$\text{var}\begin{bmatrix}\mathbf{a}_H\\ \mathbf{a}_B\\ *\\ \mathbf{c}_B\end{bmatrix} = \mathbf{H} \otimes \begin{bmatrix}\sigma^2_{\mathbf{a}} & \sigma_{\mathbf{ac}}\\ \sigma_{\mathbf{ac}} & \sigma^2_{\mathbf{c}}\end{bmatrix}$$

where $\mathbf{a}$ and $\mathbf{c}$ stands for breeding values from Hereford and Braford animals, respectively. Thus, $\mathbf{a}_H$ is the breeding value of Hereford animals for purebred performance; whether $\mathbf{a}_B$ stands for Hereford breeding values for crossbred performance. The vector $\mathbf{c}_B$ represents the Braford breeding values for crossbred performance; and * denotes artificial random values.

The general $\mathbf{H}^{-1}$ matrix can be defined as following

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & (0.95\mathbf{G} + 0.05\mathbf{A}_{22})^{-1} - \mathbf{A}_{22}^{-1}\end{bmatrix}$$

The relationship coefficient of $\mathbf{A}_{22}^{-1}$ was built based only on genotyped animals and their ancestors applying COLLEAU (2002) rules. The metafounder additive relationship, $\mathbf{H}(\boldsymbol{\Gamma})^{-1}$, was also constructed using the same approach, but differing on construction of pedigree-based relationship matrix. A unique genomic (**G**) matrix was used for both ssGBLUP and

ssGBLUP+Metafounder models. Consequently, it was implicit assumed that both breeds may have the same base allele frequencies.

Within breed predictive ability

In this study the predictive ability was used as measure of model ability to predict the future (unknown) performance. Here we intended to compare predictive ability of each breed when predicting its own performance. To that, phenotypic data was divided into five (5-fold) random different training and validation sets.

The predictive ability was represented as the average over the 5-fold random groups. Aiming to perform a fair comparison for each breed, twenty percent (20%) of phenotypes were removed for both breeds. Thus, a total of 172 and 700 animals were used as validation set for Hereford and Braford, respectively. The predictive ability was measured by cross-validation trial as the correlation between corrected phenotypes ($\hat{\mathbf{y}}^* = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$) and estimated breeding values within breeds. For example, Hereford predictive ability for its own performance was calculated as $\mathrm{cor}(\hat{\mathbf{y}}_H{}^*, \hat{\mathbf{a}}_H)$. Similarly, for Braford the predictive ability was computed as $\mathrm{cor}(\hat{\mathbf{y}}_B{}^*, \hat{\mathbf{c}}_B)$.

RESULTS AND DISCUSSION

Metafounder relationship and inbreeding

A total of 28 different metafounders were included in the ssGBLUP+metafounder model. Three metafounder groups were defined

based on breed of origin knowledge (Hereford, Braford and Zebu). Table 1 shows the number of male and female individuals included in each Hereford, Braford and Zebu metafounder group. For the remaining animals with unknown parent information, a total 25 metafounders were determined based on gender and year of birth. Twenty-three of them had less than 20 male and female animals. The remaining two metafounder groups were composed by 114 and 238 female animals.

Table 1. Number of male and female individuals included in each metafounder constructed based on breed of origin.

Metafounder	Males	Females
Hereford	2,224	1,019
Braford	5,891	2,421
Zebu	34	34

Self- and across- relationship (Γ) between Hereford, Braford and Zebu breeds estimated by GLS are $\begin{pmatrix} 0.58 & 0.49 & 0.31 \\ 0.49 & 0.54 & 0.43 \\ 0.31 & 0.43 & 0.68 \end{pmatrix}$, respectively in column-wise order. As previously defined by LEGARRA et al. (2015), $\hat{\gamma}$ can be seen as self-relationships. The relationship coefficient between metafounders was larger than zero, suggesting some overlapping degree between ancestor populations. The estimation of metafounder relationship indicates that Hereford and Zebu breeds has some degree of overlapping generations. However, as previously stated, there is no genotypic information from zebu breed in this study; in fact, only a fraction of all zebu descendants from the population were used for further computations. Thus, the population under

study is a special case (HARRIS and JOHNSON, 2010) of metafounder theory (LEGARRA et al., 2015) where at least one of pure breeds is unknown, but genotypic information from crossings are available. Moreover, the SNP panel used in this analysis is a blend of different SNP-chips where the missing genotypes were imputed. Because of that, our intention is not drawn any assumption on how these Hereford and Zebu may have shared some amount of allelic content across generations. To that end, there are other approaches already published in literature (ALEXANDER and LANGE, 2011; DECKER et al., 2014).

Inbreeding coefficient ($\mathbf{H}_{ii}^{\Gamma} - 1$) distribution calculated in ssGBLUP+Metafounder scenario is shown in Figure 1. Based on estimated $\hat{\gamma}$, negative values were seen after $\hat{\gamma} - 1$ computations. By that, it is suggested that ancestor populations are large enough (i.e., large effective population size); indicating that gametes from historical (i.e., base) population were not identical. Still in metafounder model, it can be seen that average inbreeding departure from the traditional framework, i.e. centered in zero. Classical quantitative genetic theory postulates that inbreeding for individuals with known parents are function of parent's relationship. Founder individuals are typically assumed to be drawn from a large, unrelated, ancestral population mated at random. Consequently, inbreeding for founder animals are usually defined as zero due to lack of information. A different condition arises under metafounder theory where some degree of known relationship is initially assumed between ancestral populations. In this case, the probability that identical gametes are shared between individuals may increase; thus, inbreeding coefficients are upward shifted.

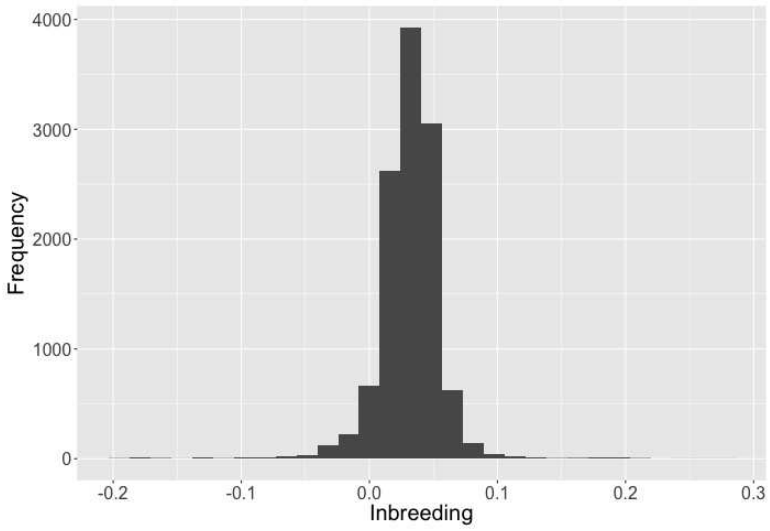

Figure 1. Inbreeding ($\boldsymbol{H}_{ii}^{\Gamma} - 1$) estimates from **H** matrix constructed based on metafounders.

Variance components, heritability and additive correlation

All variance components, heritability and additive correlation estimates are available in Table 2. As previous described, variance components of metafounder model were scaled to provide a fair comparison with BLUP and ssGBLUP models (i.e., where founders are assumed as unrelated). In general, it can be seen that additive, residual and phenotypic variance components estimated based on Hereford performance are smaller than Braford across different models. For permanent environment variance, genomic models estimated nearly the same magnitude across scenarios. A different behavior was shown on BLUP estimation where σ_{pe}^2 for Hereford is greater than for Braford breed.

When using genomic information, heritability estimates for Hereford are greater than traditional pedigree-based model. This suggests that it is possible

to achieve higher genetic gain rate over generations. Indeed, it is not expected to increase heritability estimates only due to inclusion of genomic information. However, the current pedigree is composed by many individuals with unknown parents. Consequently, there would be some level of additive variation which was not explained on BLUP model due to lack of relationship knowledge.

Some degree of change on variance components was expected as a consequence of relationship improvement in an animal model with repeated measures. Improvements in the additive relationships due to inclusion of genomic information (i.e., ssGBLUP) and metafounders (i.e., ssGBLUP+metafounder) will affect as such components are jointly estimated under frequentist theory. As observed by JUNQUEIRA et al. (2017), any improvement on additive relationships may cause a changes of additive and permanent environmental effects. Consequently, there is an increase in heritability estimates indicating the increment on success of applying direct selection. The main outcome would be the increase in the annual rate of genetic gain, mainly due to more reliable heritability estimates and better prediction of breeding value.

Additive correlation (r_a) between Hereford and Braford were 0.67, 0.45 and 0.62 for BLUP, ssGBLUP and ssGBLUP+Metafounder scenarios, respectively. The identification of existence of additive correlation is useful when designing breeding schemes and defining breeding objectives. Additive correlations are usually important to understand the magnitude of change on a first factor when applying some degree of constraint in the second factor. In the case of additive correlation between different breeds, our results that some genes responsible for the control of tick resistance are being expressed in both

purebreds and crossbreds. This result suggests that select Hereford for tick count resistance may also have a positive impact on Braford resistance.

At the end, the definition of a breeding scheme will depend on the business objective. In the case of beef cattle, there are farmers that produce sires (purebreds or crosbreds) and those farmers that purchase genetically superior sires to produce calves. Rare are cases where all the production system is performed within a farm due to structure limitations and farmers profile/business interest. However, independently of the business objective, the breeding schemes need to consider the desired outcome when deciding if additional effects (e.g., dominance) are required to achieve the goal. The design of breeding schemes for mixture populations has been a challenge for years and will continue to be if a business collaboration between farmers and associations is not efficiently implemented.

Table 2. Description of variance components, heritability and additive correlation of Hereford and Braford using multibreed pedigree and genomic information. Standard error is presented within parenthesis.

Parameters[1]	Model[2]					
	BLUP		**ssGBLUP**		**ssGBLUP+Metafounder**	
	Hereford	**Braford**	**Hereford**	**Braford**	**Hereford**	**Braford**
σ_a^2	0.003 (0.000)	0.027 (0.002)	0.009 (0.004)	0.018 (0.003)	0.012 (0.001)	0.020 (0.002)
σ_{pe}^2	0.018 (0.002)	0.006 (0.001)	0.013 (0.004)	0.013 (0.002)	0.012 (0.001)	0.013 (0.001)
σ_e^2	0.060 (0.000)	0.074 (0.001)	0.060 (0.002)	0.074 (0.002)	0.06 (0.002)	0.073 (0.001)
σ_p^2	0.081 (0.004)	0.106 (0.004)	0.082 (0.011)	0.105 (0.006)	0.084 (0.004)	0.106 (0.004)
h^2	0.04 (0.003)	0.25 (0.003)	0.11 (0.044)	0.17 (0.015)	0.14 (0.005)	0.19 (0.007)
r	0.26 (0.013)	0.31 (0.012)	0.26 (0.080)	0.30 (0.030)	0.29 (0.008)	0.31 (0.010)
r_a	0.67 (0.022)		0.45 (0.149)		0.62 (0.017)	

[1] σ_a^2: additive variance; σ_{pe}^2: permanent environment variance; σ_e^2: residual variance; σ_p^2: phenotypic variance; h^2: additive heritability; r: repeatability; r_a: additive correlation. [2]BLUP: pedigree-based model; ssGBLUP: single step genomic model; ssGBLUP+Metafounder: single step model with metafounder adjustment.

Predictive Ability

Mean predictive ability of 5-fold cross-validation approach is shown in Figure 2. This figure presents the predictive ability of each breed to predict its own performance. As expected, pedigree-based models shown the worse predictive ability when compared with both genomic models. When comparing the performance of genomic models, it is possible to see a positive increment on predictive ability due to inclusion of metafounders. The improvement on breeding values prediction is especially important for Hereford animals. Note there are less Hereford phenotypes and genotypes data than Braford data and any increase on accuracy estimation may have a direct impact under practical conditions when selecting breeding candidates. However, is important to keep in mind that these results might be influenced by the fact that only few Hereford samples were available. Perhaps, all allelic diversity present in Hereford population could not be captured; thus, further analysis is required to get a better understanding on the impacts of breeding values prediction using larger population.

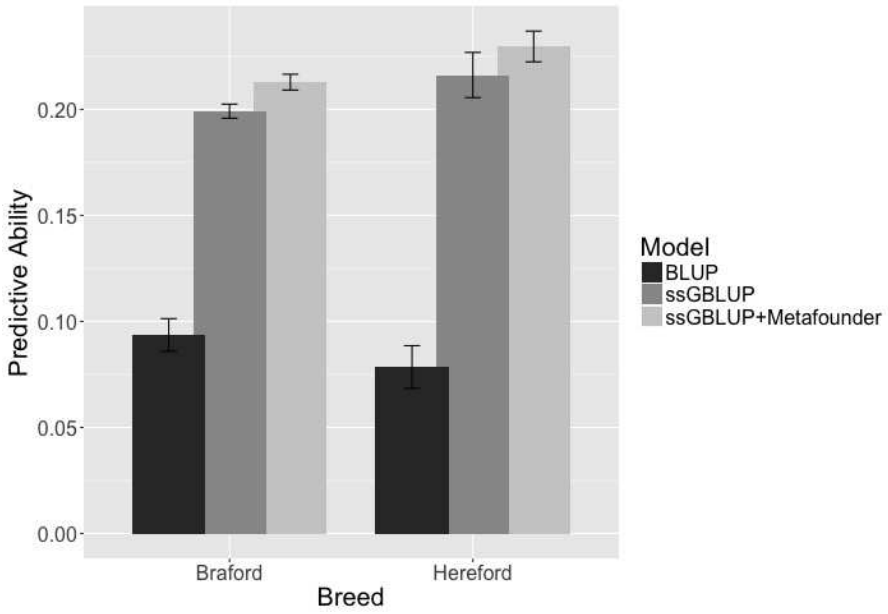

Figure 2. Mean predictive ability across of 5-fold clusters of Hereford and Braford own performance using pedigree (BLUP), single step (ssGBLUP) and single step with metafounders (ssGBLUP+Metafounder) models.

In accordance with XIANG et al. (2017), this study shows the potential of metafounders to positively increment the rate of genetic gain across generations due to an accuracy increase on breeding values prediction. Perhaps, the challenge for Brazilian breeding programs would be the availability of a large amount information to calculate $\boldsymbol{\Gamma}$. This study focused on the evaluation of the impact on breeding values due to inclusion of metafounder information. To that, all available genotypes were used to estimate $\boldsymbol{\Gamma}$. However, only a small fraction of the population is genotyped; which is also the reality of the most worldwide breeding programs. Thus, there is still a lack of knowledge on how impactful would be the number and relatedness of genotyped samples when estimating $\boldsymbol{\Gamma}$. Next studies should focus to have a better understanding on the number of genotyped samples

from different breeds may impact the prediction. Another subject would be the evaluation predictive ability when constructing $\mathbf{H}^{\Gamma}$ by using breed specific allele frequency.

CONCLUSION

The inclusion of genomic information provided greater predictive ability than pedigree-based models for both Hereford and Braford breeds. In addition, it can be seen inclusion of metafounders for genetic evaluation of beef cattle can positively impact the rate of genetic gain due to the increase of predictive ability.

ACKNOLWDGMENTS

To Andres Legarra and Ole F. Christensen for their helpful comments on the topic, and for sharing binaries and Fortran/C++ source codes used in analysis and initial tests on metafounder. To CAPES and CNPq for providing scholarship during all Master and Ph.D graduate period.

REFERENCES

AGUILAR, I.; MISZTAL, I. Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. **Journal of Dairy Science**, v. 91, n. 4, p. 1669-1672, 2008.

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 score¹. **Journal of Dairy Science**, v. 93, n. 2, p. 743-752, 2010.

AGUILAR, I.; MISZTAL, I.; LEGARRA, A.; TSURUTA, S. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. **Journal of Animal Breeding and Genetics**, v. 128, n. 6, p. 422-428, 2011.

ALEXANDER, D. H.; LANGE, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. **BMC bioinformatics**, v. 12, n. 1, p. 246, 2011.

CARDOSO, F. F.; TEMPELMAN, R. J. Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population. **Journal of Animal Science**, v. 82, n. 6, p. 1589-1601, 2004.

CHRISTENSEN, O. F. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. **Genetics Selection Evolution**, v. 44, n. 1, p. 37, 2012.

COLLEAU, J. J. An indirect approach to the extensive calculation of relationship coefficients. **Genetics Selection Evolution**, v. 34, n. 4, p. 409, 2002.

DECKER, J. E.; MCKAY, S. D.; ROLF, M. M.; KIM, J.; ALCALÁ, A. M.; SONSTEGARD, T. S.; HANOTTE, O.; GÖTHERSTRÖM, A.; SEABURY, C. M.; PRAHARANI, L. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. **PLoS genetics**, v. 10, n. 3, p. e1004254, 2014.

EMIK, L. O.; TERRILL, C. E. Systematic procedures for calculating inbreeding coefficients. **Journal of Heredity**, v. 40, n. 2, p. 51-55, 1949.

FERNANDO, R. L.; DEKKERS, J. C. M.; GARRICK, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. **Genetics Selection Evolution**, v. 46, n. 1, p. 50, 2014.

GARCIA-BACCINO, C. A.; LEGARRA, A.; CHRISTENSEN, O. F.; MISZTAL, I.; POCRNIC, I.; VITEZICA, Z. G.; CANTET, R. J. C. Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. **Genetics Selection Evolution**, v. 49, n. 1, p. 34, 2017.

HARRIS, B.; JOHNSON, D. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. **Journal of Dairy Science**, v. 93, n. 3, p. 1243-1252, 2010.

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

JUNQUEIRA, V.; CARDOSO, F.; OLIVEIRA, M.; SOLLERO, B.; SILVA, F.; LOPES, P. Use of molecular markers to improve relationship information in the genetic evaluation of beef cattle tick resistance under pedigree-based models. **Journal of Animal Breeding and Genetics**, v. 134, n. 1, p. 14-26, 2017.

KARIGL, G. A recursive algorithm for the calculation of identity coefficients. **Annals of human genetics**, v. 45, n. 3, p. 299-305, 1981.

LEGARRA, A.; CHRISTENSEN, O. F.; VITEZICA, Z. G.; AGUILAR, I.; MISZTAL, I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. **Genetics**, v. 200, n. 2, p. 455-468, 2015.

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, 2001.

MISZTAL, I.; TSURUTA, S.; STRABEL, T.; AUVRAY, B.; DRUET, T.; LEE, D. H. BLUPF90 and related programs. In: PROCEEDINGS OF THE 7TH WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 2002, **Anais.**, 2002. p.

MISZTAL, I.; VITEZICA, Z. G.; LEGARRA, A.; AGUILAR, I.; SWAN, A. A. Unknown-parent groups in single-step genomic evaluation. **Journal of Animal Breeding and Genetics**, v. 130, n. 4, p. 252-258, 2013.

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

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

WEI, M.; VAN DER WERF, J. H. Maximizing genetic response in crossbreds using both purebred and crossbred information. **Animal Science**, v. 59, n. 3, p. 401-413, 1994.

XIANG, T.; CHRISTENSEN, O. F.; LEGARRA, A. Genomic evaluation for crossbred performance in a single-step approach with metafounders. **Journal of animal science**, v. 95, n. 4, p. 1472-1480, 2017.

CHAPTER 3

IS SINGLE-STEP GENOMIC REML WITH ALGORITHM FOR PROVEN AND YOUNG MORE EFFICIENT WHEN LESS GENERATIONS OF DATA ARE PRESENT?

Abstract: Restricted maximum likelihood (REML) is a popular method for parameter estimation. Because it uses the mixed model equations, it is resistant to selection bias and efficient implementations are currently available. When genomic information is available, two versions of REML may be applicable. When only genotyped animals have phenotypes, genomic REML can be applied with a genomic relationship matrix. When only a fraction of animals is genotyped, a single-step REML is applicable. In general, it is of interest to include many genotyped animals in parameter estimation and into evaluations, to account for genomic selection or pre-selection. The aim of this study was to investigate to what extent generations truncation affects estimates for a simulated population under selection. The use of less generations reduced the ability of pedigree-based model in estimating the benchmark heritability (0.30). The decrease in heritabilities based on genomic information was less than using only pedigree relationships. Genomic models provided greater correlations than pedigree-based model; on average 25 points. Single-step genomic models do not require a deeper pedigree relationship to estimate reliable variance components and breeding values. The use of APY algorithm does not affect the estimation of variance components. An extra of 2 ungenotyped generations are sufficient to compute reliable variance components; as well as breeding values and accuracies.

Keywords: accuracy, APY, ssGBLUP, YAMS

INTRODUCTION

Restricted maximum likelihood (REML) described by PATTERSON and THOMPSON (1971) is a popular method for parameter estimation. Because it uses the mixed model equations (HENDERSON, 1975), it is resistant to selection bias and efficient implementations are currently available. With the Average Information (AI) algorithm, often convergence is achieved in a few rounds. With traces obtained by sparse matrix factorization and inversion (MEYER, 1997), computing variance components is feasible even with large models. Specialized variations exist for parameters with singular co-variance matrices (MEYER and KIRKPATRICK, 2010).

When genomic information is available, two versions of REML may be applicable. When only genotyped animals have phenotypes, genomic REML (GREML) can be applied with a genomic relationship matrix (**G**). In general, such a matrix is dense and the cost of dense matrix operations would limit computations depending on the models. When only a fraction of animals is genotyped, a single-step REML is applicable (ssGREML). In the latter, the combined relationship matrix (**H**) has dense blocks due to the genomic information, limiting efficiency of sparse matrix operations. Lately, MASUDA et al. (2015) developed a sparse matrix package YAMS that identifies dense blocks and computes them efficiently. For ssGREML, with genomic computation, such a package resulted in up to 100 times speedup, allowing 4 trait models with 20,000 genotyped animals (MASUDA et al., 2016).

In general, it is of interest to include many genotyped animals in parameter estimation and into evaluations, to account for genomic selection or pre-selection (PATRY and DUCROCQ, 2011). For instance, best reliability in

dairy was obtained using 50% of the heritability computed with a non-genomic REML (Shogo Tsuruta, personal communication). Over 1 million Holsteins have been genotyped. However, the cost of dense matrix operations with **G** in REML using YAMS is quadratic for memory and cubic for operations, and limits computations to around 50,000 animals.

The genomic information has a limited dimensionality due to the limited effective population size (STAM, 1980; VANRADEN, 2008). Such dimensionality varied from about 4,000 for pigs and chickens to 15,000 for Holsteins (POCRNIC et al., 2016b). Assuming limited dimensionality, the inverse of **G** – as needed by REML – can be sparsely constructed using the APY algorithm (MISZTAL, 2016), with close to linear memory and computing requirements. Subsequently, the inverses for over 500,000 animals can be computed and stored. However, the inverse of the **H** matrix also includes the inverse of a pedigree-based relationship matrix for genotyped animals (Aguilar et al., 2010). Such a matrix can be dense with long pedigree but is sparser with shorter pedigree. In large populations, such a matrix could not be efficiently stored but had to be accommodated indirectly (STRANDÉN and MÄNTYSAARI, 2014; MASUDA et al., 2017).

The first purpose of this study was to find whether costs of ssGREML can be reduced using the APY algorithm with truncated pedigree. The second purpose was to investigate to what extent such truncation affects estimates for a population under selection.

MATERIAL AND METHODS

Animal care and use committee approval was not needed because data used in this study were simulated.

Data simulation

To evaluate the effectiveness of the proposed approach for estimating variance components using genomic information, data were simulated using the QMSim software (SARGOLZAEI and SCHENKEL, 2009). The simulator was used to generate historical (undergoing drift and mutation) and recent (undergoing selection) populations. In total, 30 chromosomes of equal length (100 cM) were simulated. Biallelic markers (49,980) were equally distributed at random along the chromosomes with equal frequency in the first generation of the historical population. Potentially, 5,000 quantitative trait loci (QTL) affected the phenotype; QTL allele effects were sampled from a Gamma distribution with a shape parameter of 0.4. The mutation rate of the markers (recurrent mutation) and QTL was assumed to be equal to 2.5×10^{-5} per locus per generation (SOLBERG et al., 2008).

The historical population consisted of 1,000 generations with a constant size of 1,600 individuals. Then, more 20 generations were generated decreasing to 800 individuals, mimicking a bottleneck event, in generation zero. Next, three populations were sequentially generated to create the desired linkage disequilibrium (LD, $r^2 \approx 0.30$). The first recent population (P1) consisted of 400 males and 400 females randomly sampled from the last historical generation. These individuals were randomly mated over only one

generation producing a offspring size of equal proportion of males and females. The next population, P2, consisted of an expansion population in which animals were randomly mated over 8 generations; each female produced 5 offspring. In the last population (P3) 12 males and 2,000 females were selected (i.e., effective population size (N_e) ≈ 50) from last generation of P2 based on highest phenotypes. Individuals in P3 were mated along 10 generations to produce 2,000 offspring per generation, following positive assortative (matings among best males and females based on EBV) designs. In P3 each female produced only one progeny. Moreover, it was considered a sire replacement rate of 0.60 and a dam replacement rate of 0.20. Genomic information was available for 6,000 animals from generations 8 through 10.

The simulated trait had phenotypic variance and mean of 1.0 and heritability of 0.30. Phenotypes were reconstructed as

$$\mathbf{y} = \mu + \mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is the phenotype, μ is the overall mean, $\mathbf{u}$ is the weighted sum of QTL effects (i.e., additive genetic effect or animal effect), $\mathbf{e}$ is the residual term.

The simulated population was replicated 5 times and analyses were performed for all replicates.

Variance components

Variance components were estimated using average information (AI) REML algorithm as implemented in the AIREMLF90 software (Misztal et al., 2002), which was modified to incorporate the YAMS package (MASUDA et al., 2014; MASUDA et al., 2015). The incorporation of YAMS was essential for this kind of task when using genomic information. The package applies supernodal

methods using multi-core optimized libraries (parallel computing). The most computational expensive part of the variance component estimation is obtaining the inverse of the coefficient matrix used in traces. To do that, efficient algorithms are used to invert large matrices, which are based on three steps *(i)* ordering, *(ii)* factorization (i.e., symbolic and numerical) and *(iii)* inversion. Ordering is not a mandatory step, but it can save a large amount of memory after factorization. For example, a zero element in the original matrix could become a nonzero element in the factorized matrix. This is called *fill-in* effect and can be minimized by ordering using appropriate techniques. In the next step, the coefficient matrix (LHS of the mixed model equations) is factorized into two triangular matrices by LU decomposition. Finally, the Takahashi algorithm can be used for inversion. The supernodal methods are expected to provide faster inversions, because it finds and processes dense blocks in sparse matrices (MASUDA et al., 2014). Note that inversion is only required to estimate variance components or to compute prediction error variance (PEV obtained by inverting the diagonal elements of LHS). If the objective is to solve the system of equations, iterative methods as the preconditioned conjugate gradient (LIDAUER et al., 1999; TSURUTA et al., 2001) can be efficiently applied.

Different models were used for the estimation of variance components. The difference on the models relies on which relationship matrices were used. The first model was designed to consider pedigree information to construct the genetic relationships among individuals (**A** matrix) as proposed by HENDERSON (1976).

The second model was based on the single-step method where the inverse of the realized relationship matrix ($\mathbf{H}^{-1}$) is used in the mixed model equations instead of $\mathbf{A}^{-1}$. Single-step genomic BLUP (ssGBLUP) is used for breeding value estimation, whereas ssGREML is used for variance components estimation. The inversion of **H** is calculated as following (AGUILAR et al., 2010):

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

where $\mathbf{A}$ is the usual pedigree-based relationship matrix (i.e., the same as in model **A**), $\mathbf{A}_{22}^{-1}$ is the inverse of pedigree-based matrix for genotyped animals computed by the algorithm described in COLLEAU (2002). The genomic relationship matrix ($\mathbf{G}$) was computed as in VANRADEN (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_j(1-p_j)}$$

where **Z** is the matrix of gene content centered for current allele frequencies, and p_j is the allele frequency of SNP *j*.

The τ and ω are scaling factors for $\mathbf{G}_{\mathrm{APY}}^{-1}$ and $\mathbf{A}_{22}^{-1}$, respectively. In the past, these factors have been used to improve predictions by adjusting the differences between pedigree and genomic relationships. Instead of using scaling factors different from unit, inbreeding coefficients were considered when constructing the inverse of **A**. This would provide a better equivalence

between genomic and pedigree-based relationship matrices driving to a more similar genetic base. The $\mathbf{G}_{\mathrm{APY}}^{-1}$ is the inverse of the genomic relationship matrix obtained by using the algorithm for proven and young (APY) (MISZTAL et al., 2014; MISZTAL, 2016). In summary, consider that genotyped animals are arbitrarily divided into core (c) and noncore (*n*). In this algorithm, breeding values of noncore ($\mathbf{u}_n$) can be described as linear combinations of breeding values of core ($\boldsymbol{u}_c$) animals:

$$\mathbf{u}_n = \mathbf{P}_n \boldsymbol{u}_c + \boldsymbol{\Phi}_n$$

where $\mathbf{P}_n = \boldsymbol{Z}_n(\mathbf{Z'}_c\mathbf{Z}_c + \mathbf{I}\alpha)^{-\mathbf{1}}\mathbf{Z'}_c$ is a matrix that relates breeding values of noncore and core animals and $\boldsymbol{\Phi}_n$ is the Mendelian error term which has non-diagonal variance but can be approximated to diagonal. In cases where the number of core animals is large enough, breeding values of noncore animals depend only on breeding values of core animals. The inverse of $\mathbf{G}_{\mathrm{APY}}$ is constructed as following:

$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{I} - \mathbf{P'}_{cc} & -\mathbf{P}_{\boldsymbol{cn}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{M}_{cc}^{-\mathbf{1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}^{-\mathbf{1}} \end{bmatrix} \begin{bmatrix} \mathbf{I} - \mathbf{P}_{cc} & \mathbf{0} \\ -\mathbf{P}_{\boldsymbol{nc}} & \mathbf{I} \end{bmatrix}$$

If $\mathbf{G}_{cc}^{-1} = (\mathbf{I} - \mathbf{P'}_{cc})\mathbf{M}_{cc}^{-1}(\mathbf{I} - \mathbf{P}_{cc})$ is known, the complete inverse can be simplified to

$$\mathbf{G}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-\mathbf{1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{P}_{\boldsymbol{cn}} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-\mathbf{1}} [-\mathbf{P}_{\boldsymbol{nc}} \quad \mathbf{I}]$$

Because $\mathbf{G}_{\mathbf{APY}}$ is conditioned only on genotypic information of core animals, the matrix is sparser than the full $\mathbf{G}$ regularly used in ssGBLUP (Figure

1). Note that the covariance between two noncore animals is null, but variances are stored in the matrix.

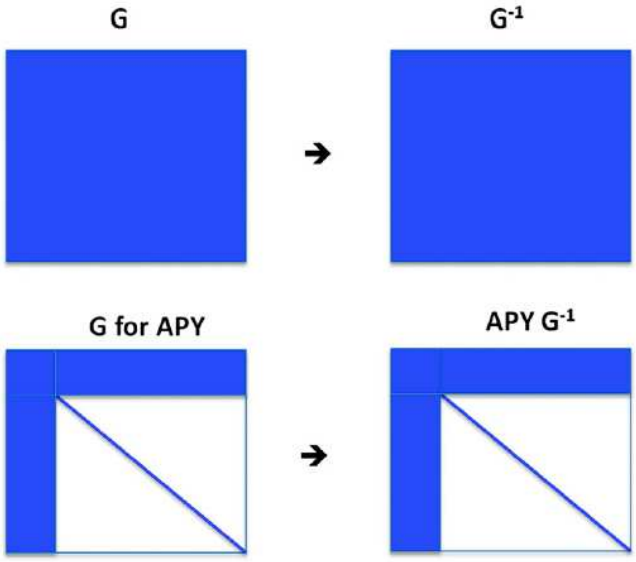

Figure 1. Scheme representing sparsity pattern differences between regular genomic ($\mathbf{G_F}$) and APY genomic ($\mathbf{G_{APY}}$) relationship matrices.

The number of animals in the core group was chosen as the number of largest eigenvalues of $\mathbf{G}$ explaining 98% of variation (POCRNIC et al., 2016a). For computational reasons, the single value decomposition of **Z** was calculated instead of the eigenvalue decomposition of $\mathbf{G}$. Across the replicates, approximately 2,700 animals were randomly chosen as core.

Scenarios

Using $\mathbf{G}_{\mathrm{APY}}^{-1}$ helps to reduce computing time for genomic predictions (FRAGOMENI et al.; MASUDA et al., 2016) because of its sparsity; however, in the single-step approach, the combined $\mathbf{H}^{-1}$ contains also $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$,

which are rather dense. The APY method was earlier applied to the construction of $\mathbf{A}_{22}^{-1}$ without success (Breno Fragomeni, personal communication). Although the sparsity of $\mathbf{A}_{22}^{-1}$ may not be a requirement for genomic predictions, it becomes essential for reducing computing time in variance components estimation. To increase the sparsity in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$, the reduction of generations was attempted. A total of eight different scenarios were designed differing on the amount of generations in pedigree used for variance components estimation. Reduction in the generations of phenotypes was also used to follow the incompleteness of pedigree and avoid bias. The scenarios were designed to mimic a real situation where usually the true founder population is unknown. Figure 2 shows the structure of pedigree, genotype, and phenotype files. Note that only three genotyped generations (6,000 youngest animals) were kept in the genotypic file for all analyses. For validation purposes, phenotypes for animals in the tenth generation (2,000 youngest animals) were removed from the dataset. Subsequent scenarios were constructed by removing one generation of phenotypes and pedigree, a time, from the oldest to the youngest animals.

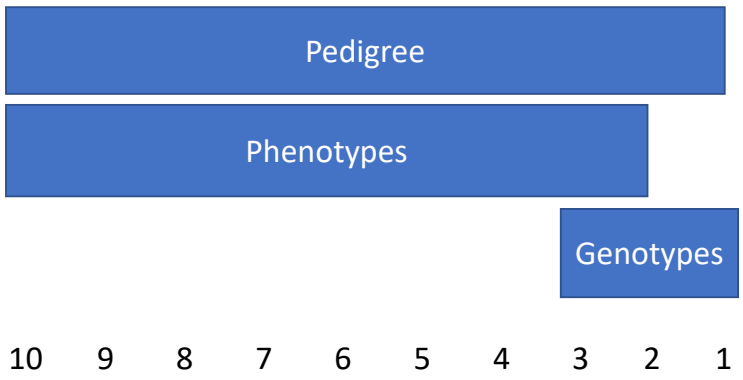

Figure 2. Structure of pedigree, phenotypic and genotypic data simulated with 10 generations. The genotypic file remained the same for all scenario and was composed of 6,000 youngest individuals. A total of 2,000 individuals from the most recent generation was used as validation set.

Inflation and accuracy of breeding values

To evaluate the impact of increasing sparsity of $\mathbf{A}_{22}^{-1}$ on genomic predictions, breeding values were also estimated in all scenarios. The regression coefficient of TBV on EBV was used as a measure of inflation of the prediction method; a value of one denotes EBV are not inflated. Validation accuracy (*r*) was computed as the correlation between TBV and (G)EBV from the animals in the tenth generation, that had their phenotypes removed from the analysis.

In addition, prediction error variance (PEV) was used to calculate breeding value accuracy (ρ) usually computed on breeding programs as $\rho = \sqrt{1 - \mathrm{PEV}/\sigma_u^2}$; here σ_u^2 is the additive genetic variance estimated for each replicate and scenario. All the results were calculated as the mean of the 5 replicates of each scenario.

RESULTS AND DISCUSSION

Some previous studies focused on the implementation of APY for large scale genetic evaluations (MASUDA et al., 2016), breeding values accuracy (FRAGOMENI et al., 2015; LOURENCO et al., 2015), efficiency on real and simulated populations with different effective sample size (POCRNIC et al., 2016a; POCRNIC et al., 2016b) and the impact of different core definitions (BRADFORD et al., 2017). In this study, we have addressed the impact of removing pedigree and phenotypic data on the feasibility of variance components estimation and on breeding value predictions using APY. Variance components were estimated using the AIREML method modified to incorporate the YAMS package for sparse matrix calculations (MASUDA et al., 2014).

Heritability estimates and computing performance

Heritabilities using different number of generations under pedigree-based and genomic-based variance components estimation are shown in Figure 3. Because the simulation involved some level of selection, the expected heritability is lower than the initial value of 0.3. Therefore, the scenario with 10 generations of pedigree was used as a benchmark. The use of less generations reduced the ability of pedigree-based model in estimating the benchmark heritability (0.30). Small fluctuations were observed when retaining only 3 to 5 generations of pedigree and phenotypes. In those scenarios, the drop-in heritability was almost nonexistent. Heritability estimates based on the pedigree relationship matrix followed an expected

trend based on quantitative theory for populations under selection (BULMER, 1971). The use of full relationship matrix accounts for change in genetic mean, variance, genetic drift, and selection (KENNEDY et al., 1988). However, the lack of pedigree relationships or its incompleteness affects the ability to trace back gene frequencies and consequently the establishment of covariance between genotypic values. In this study, we might have two different sources of genetic variance changes. The first source is related to the lack of relationship knowledge because generations were removed in different simulated scenarios. Unknown relationships (i.e., wrong base population definition) affects the variance of Mendelian sampling in different intensities depending on how many parents are known, consequently affecting heritability. If both parents are unknown, Mendelian sampling is equal to σ_u^2 and if only one parent is known, it equals to $(0.75 - 0.25 \times \mathrm{F}_p)\sigma_u^2$ (HENDERSON, 1976). Under mixed models approach, breeding values are estimated as function of parent breeding values and Mendelian sampling. Thus, all animals with unknown relationship are treated as a sample from a base population with average breeding value 0 and common variance σ_u^2. The second source of change in genetic variance is due to the presence of selection over generations, which affects the distribution of sire and dam breeding values. Unfortunately, it is impossible to identify each factor separately in this study because the scenarios were not drawn to that purpose.

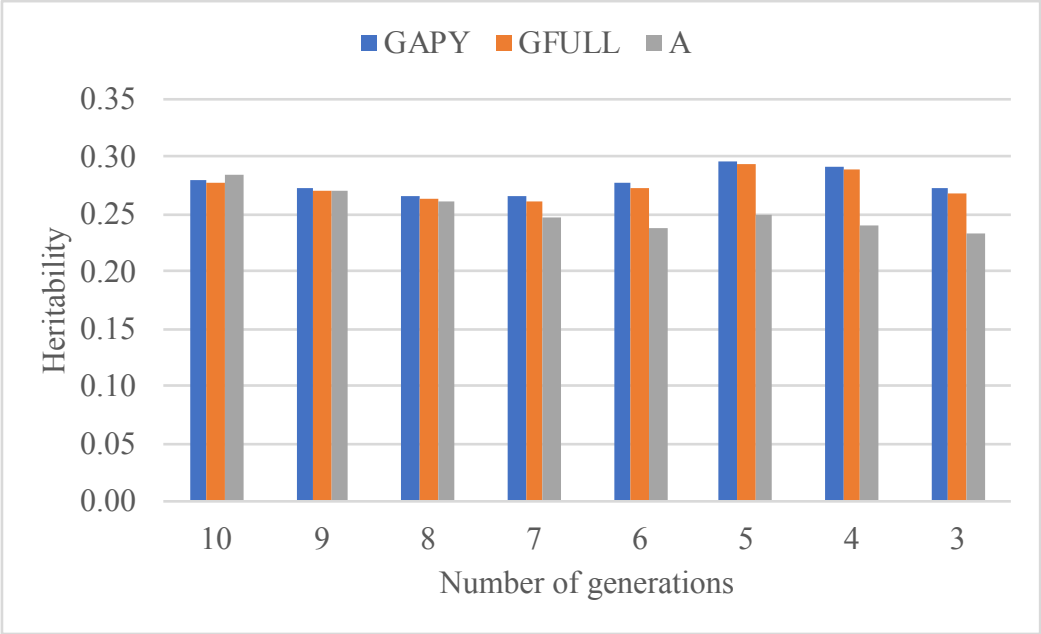

Figure 3. Average heritability estimated using a pedigree (A) and genomic (GF and GAPY) animal model across five replicates. Here A is the traditional numerator relationship matrix, GF is the single-step method using a regular genomic inversion and GAPY is the single-step model using a APY inversion.

The decrease in heritabilities based on genomic information using $\mathbf{G}_{\mathrm{APY}}^{-1}$ was less than using only pedigree relationships (Figure 3). The trend observed for heritability was also observed for additive genetic variance (results not shown), meaning the changes were due to this component instead of residual variance. Although pedigrees were more limited after that, the combination of pedigree and genomic information did not allow further decrease in heritability. The ability to estimate the Mendelian sampling term combined to the compatibility between pedigree and genomic relationships may be the possible factors. Therefore, when APY is used for estimation of variance components, removing generations of phenotypes and pedigree may reduce computing time without being harmful for the estimation of variance components.

Computing resources

Nowadays, a lot of effort is placed on developing faster and computationally feasible methods for an unlimited number of genotyped individuals. Iteration on data (IOD) by using PCG would be the best choice when the objective is estimating only breeding values. However, the estimation of variance components requires the trace of the coefficient matrix inversion. The possibility to conjugate APY (genomic) inversion, reduced pedigree, and YAMS (i.e., dense blocks operation) can be computationally beneficial.

Inflation

The degree of inflation from the breeding values is indicated by the coefficient of regression (b_1) of TBV on (G)EBV (Figure 4). The optimal method for prediction of genetic merit of young animals would have a regression coefficient close to 1. Results showed that strong simulated selection provided divergent trend between genomic and pedigree-based model. Whereas an inflation is shown for pedigree-based models, we observed a deflation for genomic models. It seems the deflations was, at some extent, inversely proportional to the decay in heritability. As previously stated, relationship matrix can account for selection, drift, and assortative mating; however, it does not account for wrong base population. If it is assumed a wrong base population, then all subsequent breeding values may be inflated because genetic flow does not have the ability to account for changes in genetic means. This topic is especially important under single-step genomic evaluations. Single-step is a blending method of genotyped and ungenotyped individuals

are accommodated into $\mathbf{H}^{-1}$ by the term $(\mathbf{G} - \mathbf{A}_{22})^{-1}$. If both matrices rely on different genetic base reference (as usually expected under livestock conditions), some inflation/deflation may appear. Differences may be due to several causes, for example due to different base population or mismatches on pedigree information (JUNQUEIRA et al., 2016). This is not a recent topic (POLLAK and QUAAS, 1983; WESTELL et al., 1988; MISZTAL et al., 2010; VITEZICA et al., 2011). More recently, LEGARRA et al. (2015) suggested the adoption of metafounders to accommodate different conditional means from one or several relationships between founders.

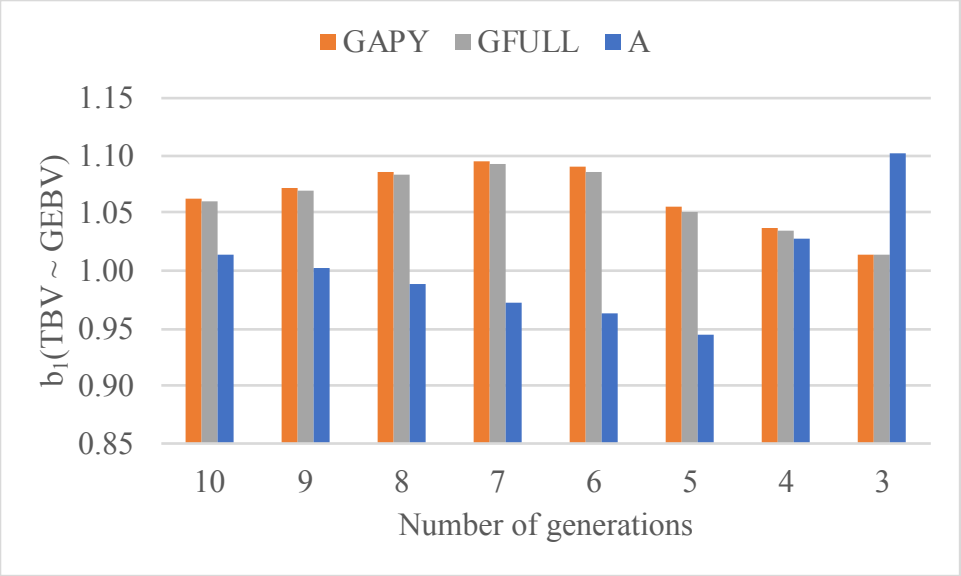

Figure 4. Regression coefficient of true (simulated) breeding value on estimated breeding value ((G)EBV) using a pedigree (**A**) and genomic ($\mathbf{G}_F$ and $\mathbf{G}_{APY}$) animal model across five replicates. Here **A** is the traditional numerator relationship matrix, $\mathbf{G}_F$ is the single-step method using a regular genomic inversion and $\mathbf{G}_{APY}$ is the single-step model using the APY inversion.

Validation and breeding value accuracy

The correlation (*r*) between true and (G)EBV for young animals (generation 10) was adopted as a measure of prediction accuracy (Figure 5).

As expected, genomic models provided greater correlations than pedigree-based model; on average 25 points. In general, this is an important parameter because response to selection is proportional to accuracy (i.e., explicitly affects genetic gain on breeding schemes). Average correlation was not reduced when generations were removed from the pedigree. This would be an indicative that older generations are not effectively contributing to accuracy of breeding values on young validation animals (LOURENCO et al., 2014). It is important to recover that contributions between genetic ties decay by half each generation and distant ancestors might have a small influence on the recent genetic background. So, the inclusion of older ancestors may not be informative especially in breeding populations (i.e., under selection).

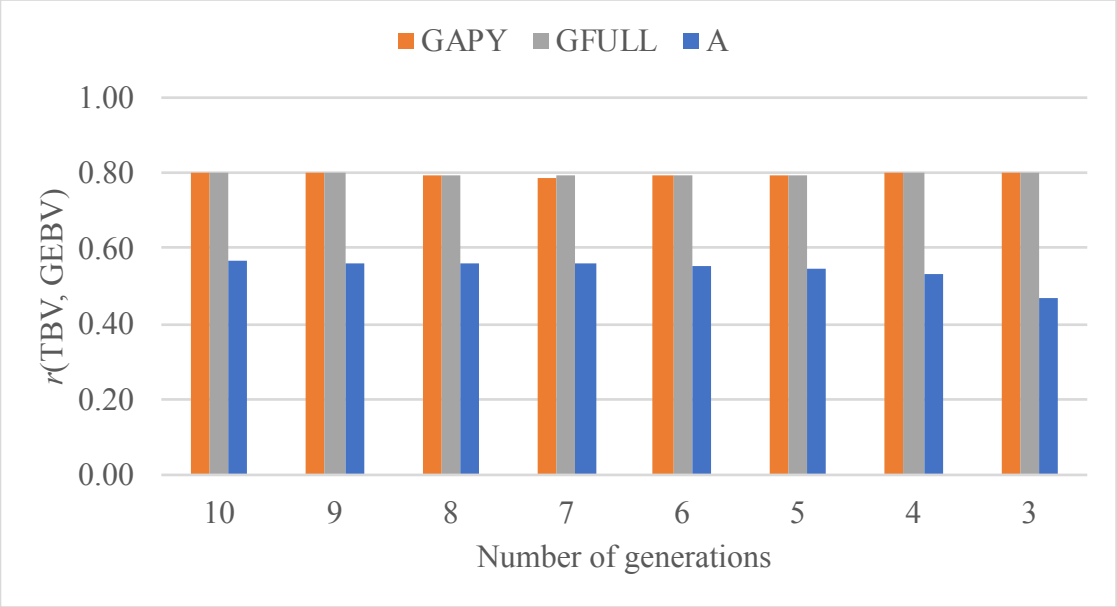

Figure 5. Correlation (i.e., accuracy) between true (simulated) breeding value and estimated breeding value (GEBV) using a pedigree (**A**) and genomic ($\mathbf{G}_F$ and $\mathbf{G}_{APY}$) animal model across five replicates. Here **A** is the traditional numerator relationship matrix, $\mathbf{G}_F$ is the single-step method using a regular genomic inversion and $\mathbf{G}_{APY}$ is the single-step model using the APY inversion.

Since true breeding values are unknown in real situations, accuracy of breeding values (ρ) is usually computed on breeding programs and delivered as a measure of precision. To calculate accuracy of breeding value, prediction error variances are required, and therefore, the inverse of the coefficient matrix. Results for ρ are presented in Figure 6. The ability of genomic information to provide greater breeding value accuracy than pedigree, as observed here, is reported elsewhere (PUTZ et al., 2018). In the same figure, it is possible to observe a greater decrease in breeding value accuracy for the pedigree-based model compared to the genomic-based model. Again, the combination of pedigree and genomic relationships in ssGBLUP avoided further drops when less information was available in the pedigree.

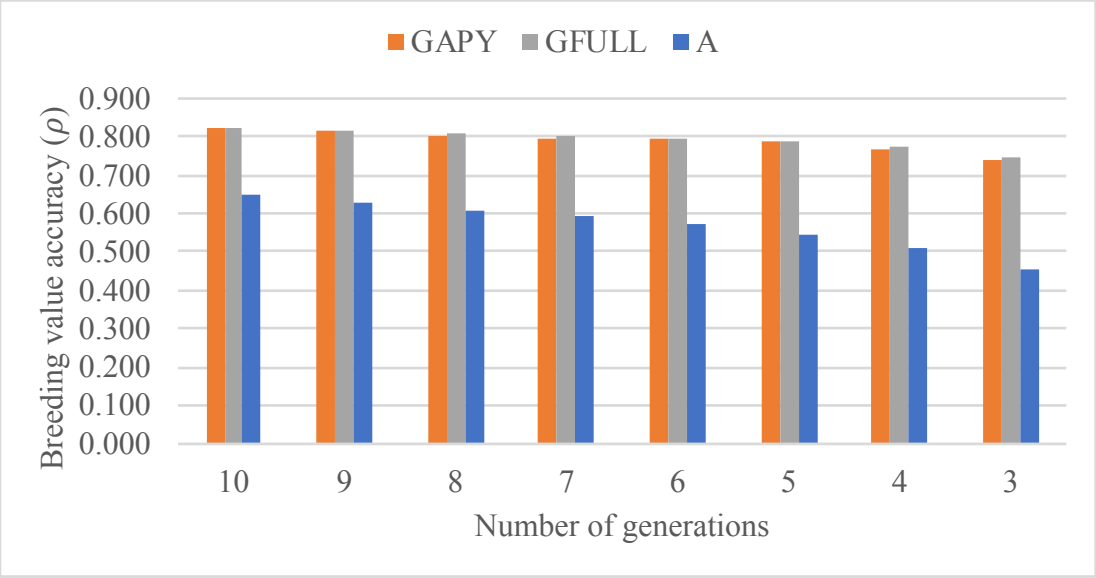

Figure 6. EBV accuracy calculated based on prediction error variance (PEV) from pedigree (**A**) and genomic ($\mathbf{G}_{APY}$) animal model across five replicates. Here **A** is the traditional numerator relationship matrix and $\mathbf{G}_{APY}$ is the single-step model using the APY inversion.

CONCLUSION

According to results presented in this study, single-step genomic models do not require a deeper pedigree relationship to estimate reliable variance components and breeding values. The use of APY algorithm does not affect the estimation of variance components. An extra of 2 ungenotyped generations are sufficient to compute reliable variance components; as well as breeding values and accuracies.

ACKNOLWDGMENTS

To Breno Fragomeni for his helpful comments, for sharing past previous with APY experiments, and for giving me suggestion when analysis did not work. To CAPES and CNPq for providing scholarship during all Master and Ph.D graduate period.

REFERENCES

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**, v. 93, n. 2, p. 743-752, 2010.

BRADFORD, H. L.; POCRNIĆ, I.; FRAGOMENI, B. O.; LOURENCO, D. A. L.; MISZTAL, I. Selection of core animals in the Algorithm for Proven and Young using a simulation model. **Journal of Animal Breeding and Genetics**, 2017.

BULMER, M. G. The effect of selection on genetic variability. **The American Naturalist**, v. 105, n. 943, p. 201-211, 1971.

COLLEAU, J. J. An indirect approach to the extensive calculation of relationship coefficients. **Genetics Selection Evolution**, v. 34, n. 4, p. 409, 2002.

FRAGOMENI, B. O.; LOURENCO, D. A. L.; TSURUTA, S.; MASUDA, Y.; AGUILAR, I.; LEGARRA, A.; LAWLOR, T. J.; MISZTAL, I. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. **Journal of Dairy Science**, v. 98, n. 6, p. 4090-4094.

FRAGOMENI, B. O.; LOURENCO, D. A. L.; TSURUTA, S.; MASUDA, Y.; AGUILAR, I.; MISZTAL, I. Use of genomic recursions and algorithm for proven and young animals for single-step genomic BLUP analyses–a simulation study. **Journal of Animal Breeding and Genetics**, v. 132, n. 5, p. 340-345, 2015.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, 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**, p. 69-83, 1976.

JUNQUEIRA, V. S.; CARDOSO, F. F.; OLIVEIRA, M. M.; SOLLERO, B. P.; SILVA, F. F.; LOPES, P. S. Use of molecular markers to improve relationship information in the genetic evaluation of beef cattle tick resistance under pedigree-based models. **Journal of Animal Breeding and Genetics**, p. 1-13, 2016.

KENNEDY, B. W.; SCHAEFFER, L. R.; SORENSEN, D. A. Genetic properties of animal models. **Journal of Dairy Science**, v. 71, p. 17-26, 1988.

LEGARRA, A.; CHRISTENSEN, O. F.; VITEZICA, Z. G.; AGUILAR, I.; MISZTAL, I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. **Genetics**, v. 200, n. 2, p. 455-468, 2015.

LIDAUER, M.; STRANDÉN, I.; MÄNTYSAARI, E. A.; PÖSÖ, J.; KETTUNEN, A. Solving large test-day models by iteration on data and preconditioned conjugate gradient. **Journal of Dairy Science**, v. 82, n. 12, p. 2788-2796, 1999.

LOURENCO, D. A. L.; MISZTAL, I.; TSURUTA, S.; AGUILAR, I.; LAWLOR, T. J.; FORNI, S.; WELLER, J. I. Are evaluations on young genotyped animals benefiting from the past generations? **Journal of Dairy Science**, v. 97, n. 6, p. 3930-3942, 2014.

LOURENCO, D. A. L.; TSURUTA, S.; FRAGOMENI, B. O.; MASUDA, Y.; AGUILAR, I.; LEGARRA, A.; BERTRAND, J. K.; AMEN, T. S.; WANG, L.; MOSER, D. W. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science**, v. 93, n. 6, p. 2653-2662, 2015.

MASUDA, Y.; AGUILAR, I.; TSURUTA, S.; MISZTAL, I. Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. **Journal of animal science**, v. 93, n. 10, p. 4670-4674, 2015.

MASUDA, Y.; BABA, T.; SUZUKI, M. Application of supernodal sparse factorization and inversion to the estimation of (co) variance components by residual maximum likelihood. **Journal of Animal Breeding and Genetics**, v. 131, n. 3, p. 227-236, 2014.

MASUDA, Y.; MISZTAL, I.; LEGARRA, A.; TSURUTA, S.; LOURENCO, D. A. L.; FRAGOMENI, B. O.; AGUILAR, I. Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. **Journal of animal science**, v. 95, n. 1, p. 49-52, 2017.

MASUDA, Y.; MISZTAL, I.; TSURUTA, S.; LEGARRA, A.; AGUILAR, I.; LOURENCO, D. A. L.; FRAGOMENI, B. O.; LAWLOR, T. J. Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals. **Journal of Dairy Science**, v. 99, n. 3, p. 1968-1974, 2016.

MEYER, K. An average information restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. **Genetics Selection Evolution**, v. 29, n. 2, p. 97, 1997.

MEYER, K.; KIRKPATRICK, M. Better estimates of genetic covariance matrices by'bending'using penalized maximum likelihood. **Genetics**, 2010.

MISZTAL, I. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. **Genetics**, v. 202, n. 2, p. 401-409, 2016.

MISZTAL, I.; AGUILAR, I.; LEGARRA, A.; LAWLOR, T. J. Choice of parameters for single-step genomic evaluation for type. **Journal of Dairy Science**, v. 93, n. Suppl 1, p. 533, 2010.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Using recursion to compute the inverse of the genomic relationship matrix. **Journal of Dairy Science**, v. 97, n. 6, p. 3943-3952, 2014.

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

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

POCRNIC, I.; LOURENCO, D. A. L.; MASUDA, Y.; LEGARRA, A.; MISZTAL, I. The dimensionality of genomic information and its effect on genomic prediction. **Genetics**, v. 203, n. 1, p. 573-581, 2016a.

POCRNIC, I.; LOURENCO, D. A. L.; MASUDA, Y.; MISZTAL, I. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. **Genetics Selection Evolution**, v. 48, n. 1, p. 82, 2016b.

POLLAK, E. J.; QUAAS, R. L. Definition of group effects in sire evaluation models. **Journal of Dairy Science**, v. 66, n. 7, p. 1503-1509 %@ 0022-0302, 1983.

PUTZ, A. M.; TIEZZI, F.; MALTECCA, C.; GRAY, K. A.; KNAUER, M. T. A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine litter size traits using Large White and simulated data. **Journal of Animal Breeding and Genetics**, v. 135, n. 1, p. 5-13, 2018.

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

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A. Genomic selection using different marker types and densities. **Journal of Animal Science**, v. 86, n. 10, p. 2447-2454, 2008.

STAM, P. The distribution of the fraction of the genome identical by descent in finite random mating populations. **Genetics Research**, v. 35, n. 2, p. 131-155, 1980.

STRANDÉN, I.; MÄNTYSAARI, E. A. Comparison of some equivalent equations to solve single-step GBLUP. In: PROCEEDINGS OF THE 10TH WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. VANCOUVER, 2014, **Anais.**, 2014. p. 22.

TSURUTA, S.; MISZTAL, I.; STRANDEN, I. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. **Journal of Animal Science**, v. 79, n. 5, p. 1166-1172, 2001.

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

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

WESTELL, R. A.; QUAAS, R. L.; VAN VLECK, L. D. Genetic groups in an animal model. **Journal of Dairy Science**, v. 71, n. 5, p. 1310-1318, 1988.