UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

STRATEGIES TO IMPLEMENT GENOMIC SELECTION IN MONTANA COMPOSITE BEEF CATTLE

Sabrina Kluska
Zootecnista

2021

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

STRATEGIES TO IMPLEMENT GENOMIC SELECTION IN MONTANA COMPOSITE BEEF CATTLE

Sabrina Kluska

Orientador: Dr. Fernando Baldi

Coorientadora: Dra. Daniela Lourenço

Coorientador: Dr. José Bento Sterman Ferraz

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Campus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Genética e Melhoramento Animal

2021

K66s

Kluska, Sabrina

Strategies to implement genomic selection in Montana composite beef cattle / Sabrina Kluska. -- Jaboticabal, 2021

113 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Fernando Sebastián Baldi Rey

Coorientadora: Daniela Andressa Lino Lourenço

1. Genética e Melhoramento Animal. 2. Seleção Genômica. 3. Bovinos compostos. 4. single-step GBLUP. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: STRATEGIES TO IMPLEMENT GENOMIC SELECTION IN MONTANA COMPOSITE BEEF CATTLE

AUTORA: SABRINA KLUSKA
ORIENTADOR: FERNANDO SEBASTIAN BALDI REY
COORIENTADORA: DANIELA ANDRESSA LINO LOURENÇO
COORIENTADOR: JOSÉ BENTO STERMAN FERRAZ

Aprovada como parte das exigências para obtenção do Título de Doutora em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Prof.Dr. FERNANDO SEBASTIAN BALDI REY (Participaçao Virtual)
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Pós-doutorando RAFAEL ESPIGOLAN (Participaçao Virtual)
FZEA/USP / Pirassununga/SP

Prof. Dr. DANISIO PRADO MUNARI (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / Unesp - Jaboticabal

Dr. RAFAEL MEDEIROS DE OLIVEIRA SILVA (Participaçao Virtual)
ZOETIS Brasil / Sertãozinho/SP

Prof. Dr. HENRIQUE NUNES DE OLIVEIRA (Participaçao Virtual)
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

Jaboticabal, 06 de janeiro de 2021

DADOS CURRICULARES DO AUTOR

SABRINA KLUSKA – solteira, nascida em 9 de outubro de 1993, na cidade de Corbélia – PR, filha de Julio Kluska e Marli Kluska. Iniciou o curso de Graduação em Zootecnia pela Universidade Tecnológica Federal do Paraná, campus Dois Vizinhos, no ano de 2011 obtendo o título de Zootecnista em março de 2015. No mesmo ano, ingressou no Programa de Pós-Graduação em Zootecnia pela mesma Universidade, sob orientação do Prof. Dr. Elias Nunes Martins, obtendo o título de Mestre em fevereiro de 2017. Em março de 2017, ingressou no curso de doutorado no Programa de Pós-Graduação em Genética e Melhoramento Animal pela Universidade Estadual Paulista "Júlio de Mesquita Filho" sob orientação do Prof. Dr. Fernando Baldi. Inicialmente como bolsista CAPES e posteriormente FAPESP (Processo FAPESP 2017/21573-0). No ano de 2019, realizou estágio de pesquisa no exterior (Processo BEPE FAPESP 2019/05516-1) na Universidade da Georgia (UGA), nos Estados Unidos, sob orientação da Profa. Dra. Daniela Lourenco. Obteve o título de Doutora em Genética e Melhoramento Animal em 06 de janeiro de 2021.

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou, sobre aquilo que todo mundo vê."

(Arthur Schopenhauer)

Dedico este trabalho aos meus pais Julio e Marli

AGRADECIMENTOS

A Deus.

A Faculdade de Ciências Agrárias e Veterinárias - Campus de Jaboticabal.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -Brasil (CAPES).

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pela concessão das bolsas de doutorado (processo 2017/21573-0) e doutorado sanduiche (processo 2019/05516-1), as quais me possibilitaram grande crescimento profissional.

Ao programa de melhoramento Montana Composto Tropical® pelo fornecimento do banco de dados.

Ao meu orientador Prof. Dr. Fernando Baldi pelo tempo, confiança, paciência e conselhos. Exemplo de profissional que me instigou a buscar a cada dia os meus objetivos.

A minha coorientadora, Dr. Daniela Andressa Lino Lourenço, por todos os ensinamentos e oportunidades que me proporcionou durante o estágio sanduíche na Universidade da Georgia.

A Universidade da Georgia, por ter me recebido de portas abertas.

Aos membros da banca do Exame Geral de Qualificação, Prof. Dr. Roberto Carvalheiro, Prof. Dr. Henrique Nunes de Oliveira, Pr. Danisio Prado Munari e Dr. Rafael Espigolan, pelas sugestões que contribuíram e acrescentaram a este trabalho.

Aos membros da banca de defesa da tese, Prof. Dr. Henrique Nunes de Oliveira, Pr. Danisio Prado Munari, Dr. Rafael Espigolan e Dr. Rafael Medeiros de Oliveira, pelas sugestões que contribuíram e acrescentaram a este trabalho.

A todos meus amigos do PPGMA, Sabrina, Bianca, Mariana, André, Marcos, Tonussi, Fabricia, Juan, Elisa, pela amizade e os cafezinhos durante estes anos do Doutorado.

Aos meus pais por todo apoio em todas as minhas decisões, pelos inúmeros conselhos e por serem meus exemplos de simplicidade e honestidade. Os quais, mesmo nos momentos difíceis me apararam.

Ao Altevir pelo amor, companheirismo e apoio. Por suportar meus defeitos, e a distância neste tempo.

A todos os meus familiares, em especial a minha irmã Fernanda, meus sobrinhos Gustavo e Augusto, e minhas cunhadas Naza e Kellyn, por acreditarem em min e tornarem meus dias mais alegres.

Muito Obrigada!

SUMMARY

ESTRATÉGIAS PARA IMPLEMENTAR SELEÇÃO GENÔMICA EM BOVINOS COMPOSTOS MONTANA

RESUMO - A busca por novas características e métodos capazes de incrementar o ganho genético em programas de melhoramento é constante. Com o advento da seleção genômica um grande progresso genético tem sido observado. Entretanto, a seleção genômica em animais cruzados, compostos ou populações multirraciais ainda é pouco difundida. Desta forma, os objetivos do presente estudo foram: *i)* Investigar o impacto do uso de diferentes matrizes de parentesco na avaliação genética de bovinos Montana; *ii)* Investigar o efeito do uso de metafundadores (MFs) e grupos de pais desconhecidos (UPG) na avaliação genética de bovinos Montana. Foram utilizados registros de 680.551 animais no pedigree, dos quais, 1899 foram genotipados com painéis de diferentes densidades e posteriormente imputados para o painel da Neogen GeneSeek® Genomic Profiler (GGP) com aproximadamente 30.000 SNPs. Informações fenotípicas de circunferência escrotal aos 12 meses de idade (SC12), ganho de peso pós desmame (PWG), peso à desmama (WW) e peso ao nascimento (BW) foram utilizadas. Quatro matrizes de parentesco distintas, e um modelo unicaracterística, foram utilizadas: 1) Matriz de parentesco aditivo baseada no pedigree ($\mathbf{A}$); 2) Matriz de parentesco genômico, construída como no ssGBLUP default ($\mathbf{G}_1$); 3) Matriz de parentesco genômico, centrada com base nas frequências alélicas dos grupos de tipo biológico ou de componentes principais ($\mathbf{G}_2$); 4) Matriz de parentesco genômico, centrada e escalada com base nas frequências alélicas dos grupos de tipo biológico ou de componentes principais ($\mathbf{G}_3$). Além disso, metafundadores e grupos de pais desconhecidos foram utilizados, os quais foram empregados em modelos multicaracterística BLUP e ssGBLUP. Foram testados modelos com quatro ou dez MFs e UPGs. UPGs foram adicionados na matriz $\mathbf{H}$ no ssGBLUP (ssGBLUP_UPG) ou apenas nas matrizes $\mathbf{A}^{-1}$ e $\mathbf{A}_{22}^{-1}$ (ssGBLUP_UPGA). Para a validação dos resultados, o método baseado em estatísticas de regressão linear (LR), e 436 animais com fenótipos omitidos, foram utilizados. Ajustes nas matrizes de parentesco genômico não foram capazes de captar maior proporção de variância genética aditiva em relação a variância fenotípica, ou seja, produzir maior herdabilidade. A adição da informação genômica, no modelo, em ambos os estudos, foi capaz de incrementar a estabilidade dos GEBVs. Contudo, a estabilidade dos GEBVs foi superior quando o modelo unicaracterística foi utilizado. Todos os parâmetros de comparação utilizados (estabilidade dos (G)EBVs, acurácia dos modelos, acurácia BIF, dispersão, viés, média dos (G)EBVs e correlação de Spearman) não indicaram nenhuma diferença significativa nas predições quando as matrizes de relacionamento genômico foram ajustadas com base em tipos biológicos ou grupos de componentes principais. As correlações de Spearman dos valores genéticos, entre os modelos baseados em pedigree e genômicos foram baixas, indicando mudanças no ranking dos animais quando a seleção é praticada com estes modelos. Entretanto, quando comparados os modelos genômicos, com $\mathbf{G}$ ajustada ou não, a correlação entre os GEBV foi alta, indicando pouca ou nenhuma mudança na classificação dos candidatos a seleção quando a seleção é realizada com base em qualquer um dos modelos genômicos. No geral a inclusão de UPGs nos modelos produziu estabilidade e dispersão similar aos

demais modelos genômicos, entretanto, tendências genéticas viesadas foram observadas quando estes foram incluídos somente nas matrizes de parentesco baseado no pedigree. A inclusão de metafundadores no modelo não foi capaz de provocar mudanças consideráveis na estabilidade dos (G)EBVs em avaliações subsequentes, exceto para PWG, e dispersão dos modelos. Entretanto, os modelos com metafundadores, quatro ou dez, produziram um viés menor do que os demais modelos genômicos, e semelhante ao BLUP baseado no pedigree. Estes resultados indicam que o uso de metafundadores pode reduzir o viés das avaliações genômicas em bovinos Montana ao mesmo nível dos modelos baseados em pedigree, com acurácia levemente menor e correlação ou estabilidade dos (G)EBV similar ao ssGBLUP default.

Palavras-chave: BLUP genômico de passo-único, grupos genéticos, matriz de parentesco genômico, metafundadores, seleção genômica

STRATEGIES TO IMPLEMENT GENOMIC SELECTION IN MONTANA COMPOSITE BEEF CATTLE

ABSTRACT - The search for new traits and methods able to increase the genetic gain in breeding programs is constant. With the advent of genomic selection, a great genetic progress has been observed, however the genomic selection in crossed animals, composite or multiracial populations is still not widespread. Thus, the aim of this study were: *i)* Investigate the impact of the use of different relationship matrices in the genetic evaluation of Montana cattle; *ii)* Investigate the effect of using metafounders (MFs) and unknown parents groups (UPG) on the genetic evaluation of Montana cattle. Records of 680,551 animals in the pedigree were available, of which 1,899 were genotyped with panels of different densities and subsequently imputed to the Neogen GeneSeek® Genomic Profiler (GGP) panel with around 30,000 SNPs. Phenotypic records of scrotal circumference at 12 months of age (SC12), postweaning weight gain (PWG), weaning weight (WW) and birth weight (BW) were available. Four distinct relationship matrices, and a single-trait model, were used: 1) Additive relationship matrix based on the pedigree ($\mathbf{A}$); 2) Genomic relationship matrix, built as in the default ssGBLUP ($\mathbf{G}1$); 3) Genomic relationship matrix, centered based on the specific-allele frequencies of the groups biological type groups or principal components ($\mathbf{G}2$); 4) Genomic relationship matrix, centered and scaled based on the specific-allele frequencies of the groups of biological type or principal components ($\mathbf{G}3$). In addition, metafounders and unknown parent groups were implemented in a multi-trait model at BLUP and ssGBLUP. Models with four or ten MFs and UPGs were tested. UPGs were added in the $\mathbf{H}$ matrix in ssGBLUP (ssGBLUP_UPG) or only in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ (ssGBLUP_UPGA) matrices. For the validation, the method based on linear regression statistics (LR), and 436 animals with omitted phenotypes, were used. Adjustments in the genomic relationship matrices were not able to capture a greater proportion of additive genetic variance in relation to phenotypic variance, that is, to produce greater heritability. The addition of genomic information, in the model, in both studies, was able to increase the stability of the (G)EBVs. However, the stability of the (G)EBVs was higher when the single-trait model was used. All the comparison parameters used (stability of (G)EBVs, accuracy of models, BIF accuracy, dispersion, bias, mean of (G)EBVs and Spearman correlation) did not indicate any significant difference in predictions when the genomic relationship matrices were adjusted based on biological types or groups of principal components. Spearman's correlations of genetic values between pedigree-based and genomic models were low, indicating changes in the ranking of animals when selection is practiced with these models. However, when comparing the genomic models, with G adjusted or not, the correlation between the GEBVs was high, indicating little or no change in the classification of candidates for selection when the selection is made based on any of the genomic models. Overall, the inclusion of UPGs in the models produced a stability and dispersion similar to the other genomic models, however, biased genetic tendencies were observed when UPGs were considered only in the pedigree-based matrices. The inclusion of metafounders in the model was not able to produce considerable changes in the stability of (G)EBVs in two subsequent evaluations, except for PWG and dispersion of the models. However, the models with metafounders, either four or ten, produced a lower bias than the other genomic models, and similar to the BLUP based on the

pedigree. These results indicate that the use of metafounders can reduce the bias of genomic predictions in Montana cattle to the same level as the pedigree-based models, with slightly less accuracy and correlation or stability of (G)EBV similar to the ssGBLUP default.

Keywords: genetic groups, genomic selection, genomic relationship matrix, metafounders, single-step GBLUP

CHAPTER 1 – General Considerations

Introduction

The development of mixed model equations (MME) proposed by Henderson (Henderson, 1949;Henderson, 1950) allowed the use of animal model to predict the breeding values in animal breeding programs. Mixed model equations allow to consider simultaneously fixed and random effects in the model. In the animal model, one of mixed models, the inverse of additive relationship matrix across animals in the pedigree ($\mathbf{A^{-1}}$) is used as a random effect in the model. This kind of mixed model is called animal model because it is used to evaluate an animal (Mrode, 2014). In addition to the animal effect, others random and fixed effects can be taken into account in the animal model to perform the genetic evaluation.

First traits evaluated in breeding programs in Brazil were growth traits. At the beginning of 2000s, works were focused towards the search of new traits as female fertility and carcass traits, methods to estimate variance components and models to perform the genetic evaluation (Lôbo et al., 2010). With the advances of molecular studies, the first works were focused on identification of QTLs and candidate genes.

The marker assisted selection (MAS) may be used as a way to consider the markers information to predict the breeding values. MAS consists of two steps: first the significant markers that affect the trait of interest and are in linkage disequilibrium with the QTL are identified; second those markers are taken into account in the model used to predict the breeding values. However, for quantitative traits a large number of

QTLs may be associated with the trait of interest, so that each QTL will explain just a small proportion of additive genetic variance.

Although the idea of using information coming from DNA to improve the rate of genetic gain in animal breeding has been started around 1960s (Smith, 1967), just upon 2 developments the genomic selection revolution has begun. First, was the sequencing of the bovine genome, and the discovery of many thousands of markers in the DNA. Second was the demonstration of it was possible to predict breeding values in an accurate way using only dense marker data, which was termed as genomic selection (Meuwissen et al., 2001).

In contrast to the MAS, genomic selection needs a large number of markers spread throughout the genome, so that at least one marker will be in a linkage disequilibrium with the QTL that affects the trait of interest. Genomic selection refers to a process of selection based on genomic breeding values (GEBV). The GEBV are predicted through the sum of markers effects spread throughout the genome. Thus, it is possible to estimate GEBVs for animals without own records and progenies (Meuwissen et al., 2001). It is important to point out that before the genomic selection the genetic evaluation of animals without own records and progenies was possible, but with genomic information these values are estimated more accurately, since we have another source of information beyond the parent average.

Different methodologies to estimate the SNP markers effects in genomic selection have been studied. The multi-step method is based on deregressed EBVs or others pseudo phenotypes as DYD (daughter yield deviations) to estimate SNP effects, afterwards these effects are used to predict the direct genomic values (DGV) of selection candidates. This approach allows kept the evaluation with BLUP, however,

just animals with genotypes can be evaluated. In this sense, when the number of genotyped animals and reliable pseudo phenotypes are limited this methodology becomes hard to implement. In cases in which phenotypes and genotypes are available the single step genomic BLUP (ssGBLUP) may be a good alternative. This method allows to evaluate at the same time animals with and without genotypes without the requirement of using pseudo phenotypes simplifying the genomic evaluation (Aguilar et al., 2010).

Nowadays, the genetic evaluation of composite, multibreed or crossbred beef cattle is performed in the same way as purebred, except that for these animals usually the effects of heterozygosis and breed proportion are taken into the account in the model (Lund et al., 2014;Cardoso et al., 2015;Piccoli et al., 2017). In addition, especially for chicken and swine some authors have used phenotypes from animals of different breeds or from purebred and crossbreed as different traits, added to the use of multi-trait models in the genetic evaluation (Hidalgo et al., 2015;Xiang et al., 2016;Duenk et al., 2019).

The genome of crossbreed or composite animals can be understood as a mosaic of genomic regions inherited of parental breeds; thus, SNP alleles may capture different effects depending on the breed of origin (Sevillano et al., 2019). Since the genomic relationship matrix used in the genomic evaluation take into account the allele frequencies to be centered and scaled it is important that the population under evaluation has homogeneous allele frequencies (Lourenco et al., 2016). The genomic relationship matrix ($\mathbf{G}$) is centered to fix the mean values of the effects of alleles to zero. Additionally, this matrix is scaled based on allele frequencies to be analogous to the additive relationship matrix based on pedigree ($\mathbf{A}$) (VanRaden, 2008).

In the traditional ssGBLUP approach the mean allele frequencies of genotyped animals are used to center and scale $\mathbf{G}$ matrix. However, for multibreed or composite population, it is not always the best way, because allele frequencies across genotyped animals are heterogeneous (Lourenco et al., 2016). Studies aiming to adjust the allele frequencies in $\mathbf{G}$ matrix for crossbred populations in chicken, swine and dairy cattle had been proposed (Simeone et al., 2012;Makgahlela et al., 2013b;Lourenco et al., 2014). However, as far we know, studies in composite beef cattle aiming to verify the impact of using specific allele frequency across-breed/group to center and scale $\mathbf{G}$ matrix were not performed.

The use of genetic groups was proposed by Quaas (1988) and Westell et al. (1988) aiming to model missing pedigrees and genetic differences among animal groups in genetic evaluation, and can also be called unknown parent groups (UPG). Additionally, UPGs can be accounted for in mixed model equations to model breed differences in a multibreed evaluation (Legarra et al., 2007). The UPGs may be treated as fixed or random effects in the model. Currently there are three different approaches used to account for UPGs in the ssGBLUP: 1) UPGs are accounted for through the additive relationship matrix based on pedigree, in which $\mathbf{A^{-1}}$ is replaced by $\mathbf{A}^{*}$ using the QP transformation (Quaas and Polak transformation); 2) UPGs contributions are added to $\mathbf{A^{-1}}$ e $\mathbf{A_{22}}$ matrices; 3) UPGs contributions are added to genomic and pedigree-based relationship matrices ($\mathbf{G}$, $\mathbf{A^{-1}}$ e $\mathbf{A_{22}}$) (Misztal et al., 2013).

In theory, the genomic relationship matrix is not affected by missing pedigrees. However, for crossbreed populations, in which UPGs indicates both missing pedigrees and breed differences, both relationship matrices ($\mathbf{A}$ and $\mathbf{G}$) should be adjusted to become compatibles, and then reduce the bias for genomic predictions in ssGBLUP

(Misztal et al., 2013). Several papers have shown that the inclusion of UPGs in the model, if these groups has an enough number of records and animals, may be able to reduce the bias of GEBV (Tsuruta et al., 2014;Bradford et al., 2019;Tsuruta et al., 2019).

Accounting for UPGs in the model is equivalent to assuming that base populations are not related and inbreeding, however, it is not always true. Usually, the relationship within and across base populations exists (Legarra et al., 2015). Aiming solve this issue, Legarra et al. (2015) proposed the concept of metafounders. Metafounders can be understood as pseudo individuals added to the pedigree in a similar way as UPGs, however the relationship within and across metafounders are taken into account. In this sense, each base population may be assigned to a one metafounder (Legarra et al., 2015). The main advantage due to the inclusion of metafounders in the model is the compatibility between genomic and pedigree-based relationship matrices, which leads to the decreasing of bias (Garcia-Baccino et al., 2017).

Metafounders can be understood as random UPGs, in which relationship across groups and inbreeding are taken into account. The inclusion of metafounders into the model may be performed in a very simple way so that just changes on $\mathbf{A}$ matrix is made. In this way, $\mathbf{A}$ matrix is adjusted to be compatible with $\mathbf{G}$ contrary to the tuning methods used till now (Legarra et al., 2015). Some simulation studies showed the advantages of using metafounders, such as the similarity of accuracy or predictive ability of models, better convergence rate and decrease of GEBVs' bias (Garcia-Baccino et al., 2017;van Grevenhof et al., 2018;Bradford et al., 2019). However, there is the need to apply metafounders to real databases in order to evaluate the benefits

and limitations of this tool in different population structures, number of genotyped animals, among other factors.

Objective

General Objective

The general objective of this study is to implement new methodologies into the single step genomic BLUP evaluation able to increasing the accuracy of genomic predictions in Montana composite beef cattle.

Specific Objective

- Evaluate the impact of using different adjustments in the genomic relationship matrix on accuracy, bias, GEBVs averages and ranking of animals evaluated with different relationship matrices.
- Verify the impact of the use of unknown parent groups (UPG) and metafounders on genomic predictions of Montana cattle.

Literature Review

Montana composite beef cattle

Montana animals are composed by a set of different breeds (at least three), which makes animals' genetic composition heterogeneous. The breeds used to produce these animals are grouped by the Montana® breeding program according to their similarity. This system of identifications of animals' breed compositions is called NABC, because these animals arise from four different biological types (*i.e*., breed groups) termed N, A, B and C. In group N, are animals of some *B. taurus indicus* adapted to the tropical conditions (heat tolerance, parasite resistance and food with poor quality); in group A, animals of some breeds *B. taurus taurus* known by their fertility and adaptative traits under tropical conditions; in group B, *B. taurus taurus* of British origin sexually precocious, with high growth rate and carcass quality; and in group C, *B. taurus taurus* from continental Europe with high growth rate and good carcass quality (Ferraz et al., 1999).

These animals can be classified from the NABC system into sixteenth of breed proportions. In this sense, a balanced animal has the same proportion of all biological types NABC (N = 4, A = 4, B = 4 and C = 4), always summed up to 16 in the total composition (Santana et al., 2013). However, animals' composition may vary due to the region in which these are raised and also depending on the breeder.

The Montana breeding program is recent, beginning in 1994, and its main goal is to produce a crossbred animal, favored by the effects of heterosis and breed complementarity. The traits evaluated by this breeding program are related mainly to growth and animal's conformation, such as weights in the birth, at weaning, that is adjusted to the 205 days of age, at 14 months, that is adjusted to the 420 days of age; and conformation traits such as muscle and navel score; and scrotal circumferences.

Currently, reproductive traits of females are being gradually included in the breed evaluation.

Genome wide association studies (GWAS) for traits of growth, carcass and meat quality as well as studies with runs of homozygosis in Montana composite beef cattle have been developed (Grigoletto et al., 2019;Grigoletto et al., 2020;Peripolli et al., 2020). In addition, works with genomic selection for this population are in development.

Genomic selection in composite, crossbred and multibreed beef cattle populations

The success of animal breeding programs depends directly on the early identification of the best animals. This may be reached if the decision-making process is performed in an accurate way. The accuracy of selection decisions is related to the choice of good models to evaluate those animals. The search for tools to increase the accuracy and the robustness of genetic evaluation should be constant, especially if this is still in not well developed for the breed or for the traits target of improvement.

Genomic selection allows to predict GEBVs of animals without pedigree and own phenotypic records, that is, just through the sum of the single nucleotide polymorphisms (SNP) effects with higher accuracy (Calus, 2010). This procedure allowed to reduce the generation interval in animal breeding schemes, causing the increase in genetic gain by year, since the GEBV can be estimated as early as the animal is genotyped (Stock and Reents, 2013). In this sense, the predictive ability of models used to estimate the GEBVs is very important to have genetic gain over time.

In addition predictive ability of models is important to evaluate the success of genomic selection in breeding programs (Boddhireddy et al., 2014).

In Holstein and Angus breeds, the genomic selection is already implemented gradually in breeding programs. Also, the accuracy of GEBV is higher than EBVs with pedigree-based evaluation of young animals for those populations (Boddhireddy et al., 2014;García-Ruiz et al., 2016). In Brasil, ANCP's Nelore Brasil breeding program had implemented the genomic evaluation in 2014 for several traits, such as growth, reproductive and carcass. Currently, other breeding programs are using information from markers in the genetic evaluation. Furthermore, for species as poultry and swine several commercial companies are already using genomic information (Eenennaam et al., 2014).

In composite populations, the genomic selection is still limited (Berry et al., 2016), since the accuracy of genomic predictions is related to the effective size of population, number of animals with genotypes and phenotypes and finally the relationship between the reference populations and selection candidates. In contrast, in a study with Bradford beef cattle and tick resistance, Cardoso et al. (2015) showed higher accuracy by using genomic models when compared to pedigree-based.

Differences in allele frequencies between SNP and QTL, and differences in the allele frequencies of QTL between breeds, are factors that can influence the accuracy of genomic evaluations in multiracial populations. In extreme cases, QTLs can segregate in only one of the breeds, or even the effect of QTL present in both breeds can be different (Wientjes et al., 2015). In order to consider interactions between marker effects and breed of origin in the Nordic Red dairy cattle, Makgahlela et al. (2013a) used a multi-trait random regression model, in which the same trait in different

breed was treated as a different trait. However, those authors found small increase in the prediction accuracy with that model.

The increasing in the predictive ability of models is expected when the number of markers spread through the genomic is increased. However, in studies with real and simulate data sets, in a multiracial population, models with sequence data and panels of high density showed similar accuracies for the real population (van den Berg et al., 2017).

The use of a reference population composed of several breeds has been proposed in studies with multiracial populations. In a simulation study, de Roos et al. (2009) showed genomic predictions more accurate if phenotypes of all populations are combined for a training set. In addition, those authors emphasized that when the populations under evaluation are very divergent it is important to increase the markers' density to have the same accuracy. Pryce et al. (2011) showed that when genetic values are predicted for a breed that doesn't have individuals in the reference population, the best way is to use a multiracial reference population to increase the accuracy of selection.

In a study with Gelbvieh beef cattle breed, without markers' information, Legarra et al. (2007) employed a multiracial model, with direct and maternal effects of heterosis, and direct and maternal effects of interaction between founder breed and generation group by using UPGs. In a Bayesian approach, the authors used different weights for the *priori* information present in the literature, and estimated from the data. According to the authors, it is possible to combine phenotypic records, pedigree information and priori estimates from the literature to perform a multiracial genetic evaluation for animals with different breed combinations.

Adjusts in the genomic relationship matrix

The development of ssGBLUP method allowed combining genomic ($\mathbf{G}$) and pedigree ($\mathbf{A}$) relationship matrices in the genetic evaluation through the joint matrix $\mathbf{H}$. This method has a lower computational cost and higher accuracy than models without genomic information (Legarra et al., 2014). One of the advantages of the ssGBLUP method is the easy inversion of $\mathbf{H}$ matrix, and the possibility of grouping genotyped and non-genotyped animals, since the process of genotyping in all population is not feasible due to the costs of this. In addition, ssGBLUP does not need pseudo phenotypes and adequately weighs the information from genotyped animals, avoiding double counting, and reducing the bias (Legarra et al., 2014).

The genomic relationship matrix is built based on markers information, through the $\mathbf{MM}'$ matrix of order *m* x *m*, where *m* is the number of genotyped animals (Legarra and Misztal, 2008). Given the frequency of the second allele at locus i be p_i; and given that $\mathbf{P}$ has the allele frequency expressed as the difference of 0.5 multiplied by two, thus the column i of $\mathbf{P}$ is $2(p_i - 0.5)$. The subtraction of $\mathbf{P}$ from $\mathbf{M}$ results in $\mathbf{Z}$, which sets the mean values of allele effects to zero. This procedure described above is known as genomic relationship matrix centralization, and is performed to force the mean of genetic values to be zero. In addition, allele frequencies are used to scale $\mathbf{G}$ matrix to be analogous to the pedigree-based relationship matrix $(\mathbf{A})$. In this sense, the genomic relationship matrix can be obtained as follows: $\mathbf{G} = \mathbf{ZZ}'/2\sum \mathrm{p_i}(1 - \mathrm{p_i})$ (VanRaden, 2008).

In the ssGBLUP, the $\mathbf{G}$ matrix is centered and scaled based on the mean allele frequencies across populations; however, this is appropriate just if population allele

frequencies are homogeneous, which is not found in crossbred or composite populations (Lourenco et al., 2016). For multiracial population, the $\mathbf{G}$ matrix should take into account specific allele frequencies for the population under evaluation, since these populations tend to have alleles of segregation in different frequencies, which can affect the GEBVs. In cases where the breed composition is known, the scale of $\mathbf{G}$ matrix can be more easily perform (VanRaden, 2008;Aguilar et al., 2010). However, for population in which the breed composition is approximated it is still unclear how the GEBV are affected by the scale of the genomic relationship matrix.

For multiracial, crossbred or composite populations the $\mathbf{G}$ matrix used to build $\mathbf{H}$ in ssGBLUP should be scaled based on population-specific allele frequencies, or even, one can assume constant allele frequency (0.5) for all markers (Simeone et al., 2012). Since the $\mathbf{H}$ matrix includes $\mathbf{G} - \mathbf{A}$ (difference between genomic and pedigree-based relationship), any inflation or deflation in the $\mathbf{G}$ matrix should cause errors in the weighting of pedigree and genomic information, causing bias in the evaluation (Forni et al., 2011). In multiracial populations, the use of ssGBLUP is possible when $\mathbf{G}$ matrix takes into account information from all breeds that make up the population (Hidalgo et al., 2020).

The influence of the $\mathbf{G}$ matrix scale in the accuracy of ssGBLUP method is most relevant when the number of animals' phenotype is limited and the trait is under strong selection pressure (Vitezica et al., 2011). Additionally, the impact of considering the specific allele frequencies for each breed in the $\mathbf{G}$ matrix seems to be more important for young animals than for proved bulls, since for young animals the genotype may be the only source of information available for the estimation of breeding values (Hidalgo et al., 2020).

According to Lourenco et al. (2016), taken into account different allele frequencies for purebred and crossbred animals is beneficial for genomic selection, especially if the correlations across allele frequencies of purebred used to produce crossbred animals are low. The genomic evaluation in multiracial populations adjusting $\mathbf{G}$ matrix for specific-breed allele frequencies was also proposed in other species, such as chicken (Simeone et al., 2012), swine (Lourenco et al., 2016) and dairy cattle (Makgahlela et al., 2013b). However, to the best of our knowledge there are no studies in the literature with composite beef cattle. Additionally, to adjust the $\mathbf{G}$ matrix to specific-founder breed allele frequency should provide greater consistencies between the relationship of genotyped and non-genotyped individuals (Makgahlela et al., 2013b).

Given the above, it is important to point out that this methodology was not initially developed for multiracial populations, together with the shortage of works with ssGBLUP in composite beef cattle.

Genetic groups and metafounders

The genetic groups were initially proposed by Quaas (1988) aiming to model missing pedigrees in the genetic evaluation. Genetic groups, unknown parent groups (UPG), or phantom parent groups, allow assuming different genetic values for unknown parents in the pedigree. The UPGs can be formed based on different criteria of missing parents, such as sex, year of birth, country of origin, selection design, breed composition, among others (Quaas, 1988;Legarra et al., 2007). Although developed to

model missing pedigrees, UPG can also account for breed differences in a multiracial evaluation (Legarra et al., 2007).

Reduction of bias in genetic trends by using UPGs were reported for works with and without genomic information (Theron et al., 2002;Tsuruta et al., 2019). However, some studies have reported the opposite, *i.e.,* increase in bias with the use of UPG (Phocas and Laloë, 2004). This is related to the poor definition of groups, small number of animals per group and also to the small number of phenotypes available for each group and trait (Tsuruta et al., 2019). In this sense, UPGs should be removed from the model when they cause a reduction in the accuracy or the increase in the genetic trend bias of the evaluations via BLUP (Phocas and Laloë, 2004).

In a simulation study, Bradford et al. (2019) pointed out that a large amount of information is necessary to obtain good estimates for UPG solutions. In addition, the authors showed that the best approach to model UPGs is likely trait-dependent, so the best definition of UPG in dairy cattle may not be the best in pigs and so on.

According to Tsuruta et al. (2014), when the inclusion of UPG in the model introduces bias, the redefinition of the groups, aiming to have more robust groups (with greater number of animals, phenotypes and better connected), can increase the reliability of GEBV estimates and produce correct genetic trends. In that study, the authors showed that the redefinition of UPGs can lead to an increase of the GEBV for both national animals and bulls imported from other countries, as well as greater correlations of ranking between GEBV of animals.

In some cases $\mathbf{G}$ and $\mathbf{A}$ matrices may be incompatible, resulting in biased genomic predictions by using the ssGBLUP (Misztal and Legarra, 2017). This occurs because all animals present in the pedigree are not always genotyped (Legarra et al.,

2014;Legarra et al., 2015). Usually, just a small proportion of pedigree's animals are genotyped (young and important animals), due to the high cost of this tool. To make the $\mathbf{G}$ and $\mathbf{A}$ matrices compatible, it is necessary that genomic and pedigree-based relationship matrices refer to the same base population (Legarra et al., 2015). However, the reference for pedigree-based relationship is the founders of pedigree, while the refence for the genomic relationship matrix is often the group of genotyped animals; this is the origin of incompatibility between these matrices (Powell et al., 2010).

In animal breeding programs it is assumed that base population are infinite and unrelated, but it is inconsistent, especially in populations with small effective size. Christensen (2012) showed that individuals from the base populations are related each other. To solve this problem, Legarra et al. (2015) proposed the concept of metafounders, making it possible to consider the relationship within and across base populations. Metafounders (MF) are pseudo individuals added to the pedigree as founders, that can be consider as father and mother simultaneously (Legarra et al., 2015).

The use of metafounders is indicated in two cases; 1) combine genomic and pedigree information among individuals, as in the ssGBLUP method; 2) when there are several base populations simultaneously (Legarra et al., 2015). In contrast to the other methods, when MF are added to the pedigree, changes occur in the pedigree-based relationship matrix ($\mathbf{A}$), so that $\mathbf{G}$ and $\mathbf{A}$ matrices becomes compatible. In the MF approach it is possible to consider several base populations, through several MF probably related each other (Legarra et al., 2015).

Ancestral population and base population are depicted in the Figure 1. Ancestral population is the population from which founders of the pedigree are drawn and base population, the set of these pedigree founders (Legarra et al., 2015). Usually, individuals in the base population are assumed to be from a large, unrelated and a population under random mating, so that individuals should be unrelated. However, as mentioned above, it is not totally true.

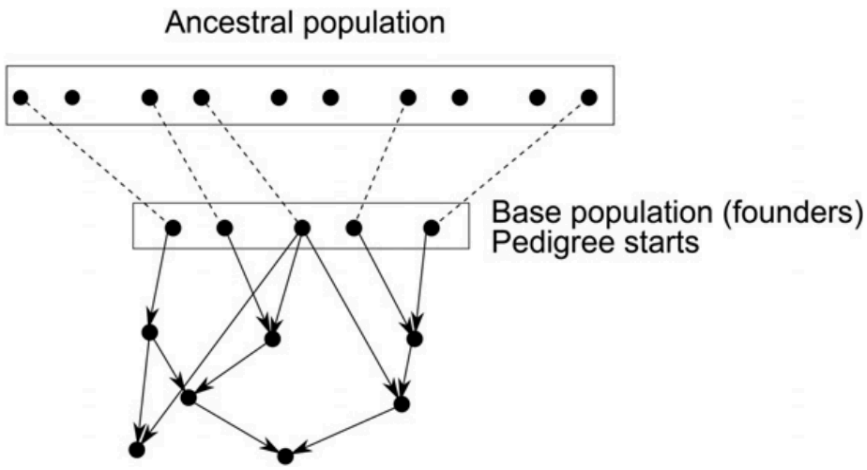

Figure 1. Ancestral and base population and pedigree (Legarra et al., 2015).

Erroneously, genetic groups were considered unrelated (Kennedy, 1991). Just like phantom parent groups or UPGs, the MF are formed according to the year of birth, sex, country, breed composition, among others (Legarra et al., 2015). According to Legarra et al. (2015) founders should be related, so that this does not occur, they should come from an infinite population, which is not always achievable.

Within and across base population relationships can be depicted in Figure 2. In this case, there are two MF (individual 1 and 2) which represents a finite-size pool of gametes, from which individuals 3 to 7 (base population) are coming from. In addition,

these 2 MF are likely related each other, in a positive or negative direction. If base populations overlapped, this value will be positive; on the other hand, if base populations have diverged due to selection in opposite directions, this value will be negative. The coefficient γ_1 and γ_2 represent the self-relationship in the MF, which can be understood as inbreeding as well. This values of relationships within and across metafounders either can come from knowledge of the history of population or can be inferred from pedigree and genomic data (Legarra et al., 2015).

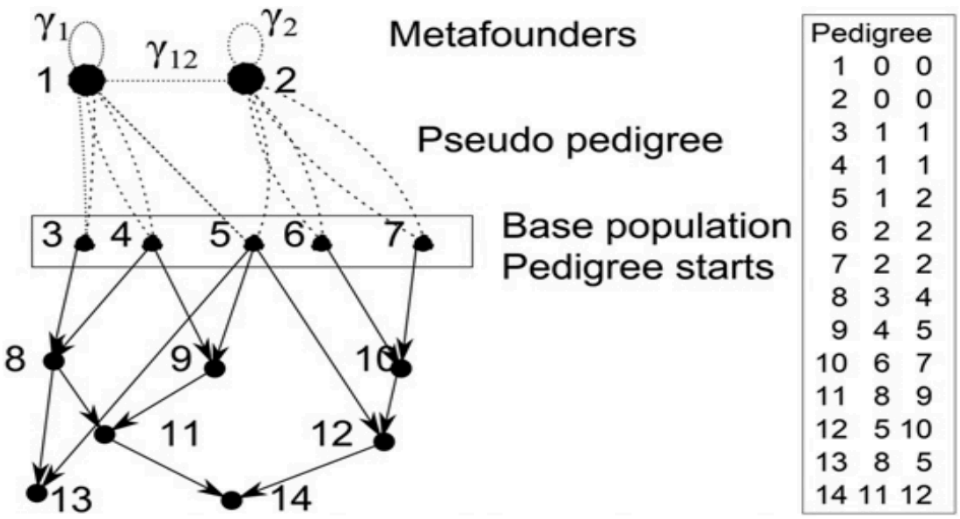

Figure 2. Base population with two related metafounders 1 and 2 metafounders and self-relationship coefficients γ_1 and γ_2 ; and across relationship γ_{12} (Legarra et al., 2015).

Some studies with MF, mainly with simulated data, have indicated the advantages of using those. Garcia-Baccino et al. (2017) simulated a purebred

population of dairy cattle in the QMSim software. These authors found genomic predictions less biased without loss of accuracy in the ssGBLUP method with MF. Also with a simulated data set, mimicking a three-way swine crossing, van Grevenhof et al. (2018) reported that ssGBLUP with metafounders may be the method of choice to implement genomic selection in the evaluation of crossing performance in animal breeding schemes. The authors showed that the use of MF improves the convergence of analysis without affecting the accuracy of predictions.

Few studies have reported the use of MF in real populations. In a study with real population, with swine, Xiang et al. (2017) reported that using MF in the ssGBLUP can works as well as to consider the breed of origin in genomic predictions. The MF can be also used to align $\mathbf{G}$ and $\mathbf{A}$ relationship matrices resulting in genomic predictions less biased. Meyer et al. (2018) showed that the use of MF is a simple and effective method to avoid bias in genetic trends for simulated and real Australian sheep data sets.

Studies with MF in beef cattle composite populations were not found in the literature, as well as studies with several MF. It is important to point out that in Montana's breeding program the base population are assumed to be unrelated and missing pedigrees are not adjusted, that is the genetic groups are not taking into account in the genetic evaluation. In addition, genomic selection still not implemented. In this sense, studies addressing these issues are of great importance for the development of the Montana breeding program.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T.J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93**,** 743-752.

Berry, D.P., Garcia, J.F., and Garrick, D.J. (2016). Development and implementation of genomic predictions in beef cattle. *Animal Frontiers* 6**,** 32-38.

Boddhireddy, P., Kelly, M.J., Northcutt, S., Prayaga, K.C., Rumph, J., and Denise, S. (2014). Genomic predictions in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation1. *Journal of Animal Science* 92**,** 485-497.

Bradford, H.L., Masuda, Y., Vanraden, P.M., Legarra, A., and Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. *Journal of Dairy Science* 102**,** 2336-2346.

Calus, M.P.L. (2010). Genomic breeding value prediction: methods and procedures. *Animal* 4**,** 157-164.

Cardoso, F.F., Gomes, C.C.G., Sollero, B.P., Oliveira, M.M., Roso, V.M., Piccoli, M.L., Higa, R.H., Yokoo, M.J., Caetano, A.R., and Aguilar, I. (2015). Genomic prediction for tick resistance in Braford and Hereford cattle1. *Journal of Animal Science* 93**,** 2693-2705.

Christensen, O.F. (2012). Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genetics Selection Evolution* 44**,** 37.

De Roos, A.P.W., Hayes, B.J., and Goddard, M.E. (2009). Reliability of genomic predictions across multiple populations. *Genetics* 183**,** 1545-1553.

Duenk, P., Calus, M.P.L., Wientjes, Y.C.J., Breen, V.P., Henshall, J.M., Hawken, R., and Bijma, P. (2019). Estimating the purebred-crossbred genetic correlation of body weight in broiler chickens with pedigree or genomic relationships. *Genetics Selection Evolution* 51**,** 6.

Eenennaam, A.L.V., Weigel, K.A., Young, A.E., Cleveland, M.A., and Dekkers, J.C.M. (2014). Applied Animal Genomics: Results from the Field. *Annual Review of Animal Biosciences* 2**,** 105-139.

Ferraz, J.B.S., Eler, J.P., and Golden, B.L. (1999). Análise genética do composto Montana Tropical. *Revista Brasileira de Reprodução Animal* 23**,** 111-113.

Forni, S., Aguilar, I., and Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* 43**,** 1.

Garcia-Baccino, C.A., Legarra, A., Christensen, O.F., Misztal, I., Pocrnic, I., Vitezica, Z.G., and Cantet, R.J.C. (2017). Metafounders are related to F (st) fixation indices and reduce bias in single-step genomic evaluations. *Genetics, selection, evolution : GSE* 49**,** 34-34.

García-Ruiz, A., Cole, J.B., Vanraden, P.M., Wiggans, G.R., Ruiz-López, F.J., and Van Tassell, C.P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic

selection. *Proceedings of the National Academy of Sciences* 113**,** E3995-E4004.

Grigoletto, L., Brito, L.F., Mattos, E.C., Eler, J.P., Bussiman, F.O., Silva, B.D.C.A., Da Silva, R.P., Carvalho, F.E., Berton, M.P., Baldi, F., and Ferraz, J.B.S. (2019). Genome-wide associations and detection of candidate genes for direct and maternal genetic effects influencing growth traits in the Montana Tropical® Composite population. *Livestock Science* 229**,** 64-76.

Grigoletto, L., Ferraz, J.B.S., Oliveira, H.R., Eler, J.P., Bussiman, F.O., Abreu Silva, B.C., Baldi, F., and Brito, L.F. (2020). Genetic Architecture of Carcass and Meat Quality Traits in Montana Tropical® Composite Beef Cattle. *Frontiers in Genetics* 11.

Henderson, C.R. (1949). Estimation of changes in herd environment. *Journal of Dairy Science* 32**,** 706.

Henderson, C.R. (Year). "Estimation of genetic parameters", in: *Biometrics*: International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210), 186-187.

Hidalgo, A.M., Bastiaansen, J.W.M., Lopes, M.S., Harlizius, B., Groenen, M.a.M., and De Koning, D.-J. (2015). Accuracy of Predicted Genomic Breeding Values in Purebred and Crossbred Pigs. *G3: Genes|Genomes|Genetics* 5**,** 1575-1583.

Hidalgo, J., Tsuruta, S., Lourenco, D., Masuda, Y., Huang, Y., Gray, K.A., and Misztal, I. (2020). Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. *Journal of Animal Science* 98.

Kennedy, B.W. (1991). C. R. Henderson: the unfinished legacy. *J Dairy Sci* 74**,** 4067-4081.

Legarra, A., Bertrand, J.K., Strabel, T., Sapp, R.L., Sánchez, J.P., and Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics* 124**,** 286-295.

Legarra, A., Christensen, O.F., Aguilar, I., and Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livestock Science* 166**,** 54-65.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., and Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. *Genetics* 200**,** 455-468.

Legarra, A., and Misztal, I. (2008). Technical Note: Computing Strategies in Genome-Wide Selection. *Journal of Dairy Science* 91**,** 360-366.

Lôbo, R.B., Bittnecourt, T.C.B.D.S.C.D., and Pinto, L.F.B. (2010). Progresso científico em melhoramento animal no Brasil na primeira década do século XXI. *Revista Brasileira de Zootecnia* 39**,** 223-235.

Lourenco, D.a.L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T.J., Forni, S., and Weller, J.I. (2014). Are evaluations on young genotyped animals benefiting from the past generations? *Journal of Dairy Science* 97**,** 3930-3942.

Lourenco, D.a.L., Tsuruta, S., Fragomeni, B.O., Chen, C.Y., Herring, W.O., and Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices1. *Journal of Animal Science* 94**,** 909-919.

Lund, M.S., Su, G., Janss, L., Guldbrandtsen, B., and Brøndum, R.F. (2014). Genomic evaluation of cattle in a multi-breed context. *Livestock Science* 166**,** 101-110.

Makgahlela, M.L., Mäntysaari, E.A., Strandén, I., Koivula, M., Nielsen, U.S., Sillanpää, M.J., and Juga, J. (2013a). Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *Journal of Animal Breeding and Genetics* 130**,** 10-19.

Makgahlela, M.L., Strandén, I., Nielsen, U.S., Sillanpää, M.J., and Mäntysaari, E.A. (2013b). The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. *Journal of Dairy Science* 96**,** 5364-5375.

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

Meyer, K., Tier, B., and Swan, A. (2018). Estimates of genetic trend for single-step genomic evaluations. *Genetics Selection Evolution* 50**,** 39.

Misztal, I., and Legarra, A. (2017). Invited review: efficient computation strategies in genomic selection. *animal* 11**,** 731-736.

Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I., and Swan, A.A. (2013). Unknown-parent groups in single-step genomic evaluation. *Journal of Animal Breeding and Genetics* 130**,** 252-258.

Mrode, R.A. (2014). *Linear models for the prediction of animal breeding values.* Wallingford, Oxfordshire, UK: CABI.

Peripolli, E., Stafuzza, N.B., Amorim, S.T., De Lemos, M.V.A., Grigoletto, L., Kluska, S., Ferraz, J.B.S., Eler, J.P., Mattos, E.C., and Baldi, F. (2020). Genome-wide scan for runs of homozygosity in the composite Montana Tropical® beef cattle. *Journal of Animal Breeding and Genetics* 137**,** 155-165.

Phocas, F., and Laloë, D. (2004). Should genetic groups be fitted in BLUP evaluation? Practical answer for the French AI beef sire evaluation. *Genetics, selection, evolution : GSE* 36**,** 325-345.

Piccoli, M.L., Brito, L.F., Braccini, J., Cardoso, F.F., Sargolzaei, M., and Schenkel, F.S. (2017). Genomic predictions for economically important traits in Brazilian Braford and Hereford beef cattle using true and imputed genotypes. *BMC genetics* 18**,** 2-2.

Powell, J.E., Visscher, P.M., and Goddard, M.E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics* 11**,** 800-805.

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M.E., and Hayes, B.J. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science* 94**,** 2625-2630.

Quaas, R.L. (1988). Additive Genetic Model with Groups and Relationships. *Journal of Dairy Science* 71**,** 1338-1345.

Santana, M.L., Eler, J.P., Cardoso, F.F., Albuquerque, L.G., and Ferraz, J.B.S. (2013). Phenotypic plasticity of composite beef cattle performance using reaction norms model with unknown covariate. *animal* 7**,** 202-210.

Sevillano, C.A., Bovenhuis, H., and Calus, M.P.L. (2019). Genomic Evaluation for a Crossbreeding System Implementing Breed-of-Origin for Targeted Markers. *Frontiers in Genetics* 10.

Simeone, R., Misztal, I., Aguilar, I., and Vitezica, Z.G. (2012). Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *Journal of Animal Breeding and Genetics* 129**,** 3-10.

Smith, C. (1967). Improvement of metric traits through specific genetic loci. *Anim. prod* 9**,** 3.

Stock, K., and Reents, R. (2013). Genomic Selection: Status in Different Species and Challenges for Breeding. *Reproduction in Domestic Animals* 48**,** 2-10.

Theron, H., Kanfer, F., and Rautenbach, L. (2002). The effect of phantom parent groups on genetic trend estimation. *South African Journal of Animal Science* 32**,** 130-135.

Tsuruta, S., Lourenco, D.a.L., Masuda, Y., Misztal, I., and Lawlor, T.J. (2019). Controlling bias in genomic breeding values for young genotyped bulls. *Journal of Dairy Science* 102**,** 9956-9970.

Tsuruta, S., Misztal, I., Lourenco, D.a.L., and Lawlor, T.J. (2014). Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. *Journal of Dairy Science* 97**,** 5814-5821.

Van Den Berg, I., Bowman, P.J., Macleod, I.M., Hayes, B.J., Wang, T., Bolormaa, S., and Goddard, M.E. (2017). Multi-breed genomic prediction using Bayes R with sequence data and dropping variants with a small effect. *Genetics Selection Evolution* 49**,** 70.

Van Grevenhof, E.M., Vandenplas, J., and Calus, M.P.L. (2018). Genomic prediction for crossbred performance using metafounders1. *Journal of Animal Science* 97**,** 548-558.

Vanraden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91**,** 4414-4423.

Vitezica, Z.G., Aguilar, I., Misztal, I., and Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research* 93**,** 357-366.

Westell, R.A., Quaas, R.L., and Van Vleck, L.D. (1988). Genetic Groups in an Animal Model. *Journal of Dairy Science* 71**,** 1310-1318.

Wientjes, Y.C.J., Calus, M.P.L., Goddard, M.E., and Hayes, B.J. (2015). Impact of QTL properties on the accuracy of multi-breed genomic prediction. *Genetics Selection Evolution* 47**,** 42.

Xiang, T., Christensen, O.F., and Legarra, A. (2017). Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. *Journal of Animal Science* 95**,** 1472-1480.

Xiang, T., Nielsen, B., Su, G., Legarra, A., and Christensen, O.F. (2016). Application of single-step genomic evaluation for crossbred performance in pig1. *Journal of Animal Science* 94**,** 936-948.

CHAPTER 2 - Genomic evaluation for composite beef cattle using adjusted realized genomic relationship matrix

Abstract - The aim of this study was to compare genomic predictions in a composite beef cattle population using different relationship matrices. A total of 680,551 animals in the pedigree and phenotypic records of scrotal circumference at 12 months of age (SC12), postweaning weight gain (PWG), weaning weight (WW) and birth weight (BW), were available. GeneSeek® Genomic Profiler (GGP) genotypes were available for 1,899 animals. A single-trait model was used to estimate variance components and breeding values. Analysis were carried out with four different relationship matrices: 1) we had used additive genetic relationship matrix, based on pedigree ($\mathbf{A}$); 2) Genomic relationship matrix was building as a default in ssGBLUP method ($\mathbf{G}1$); 3) Genomic relationship matrix was centered based on specific allele frequency of biological types or principal component (PC) groups ($\mathbf{G}2$); 4) Genomic relationship matrix was centered and scaled based on specific allele frequency of biological types or PC's groups ($\mathbf{G}3$). Groups used to adjust $\mathbf{G}$ matrix were based on the biological types and principal components. Validation was performed on 436 genotyped animals that had their phenotypes omitted. The linear regression (LR) validation method was used to validate results. Using group-specific allele frequencies was not able to capture more proportion of phenotypic variance explained by genetic additive variance, that is heritability. Overall, the accuracy of models (ranged from 0.42 to 0.69 across trait and models) and stability of (G)EBV in two subsequent evaluations (ranged from 0.73 to 0.92 across trait and models) was increased when genomic information was taken into account, however, differences across genomic models were not noted. For PWG and WW traits accuracy of BLUP models was higher than genomic models. Dispersion of genomic models was higher than pedigree-based, except for WW. Also, the bias was higher for genomic models than pedigree-based, indicating that including genomic information may introduce some bias into the genetic evaluation. However, slight differences across genomic models were observed for dispersion and bias. The median of (G)EBVs were higher for genomic models (did not differ across genomic models) which is likely due to the bias introduced by genomic information. Spearman's correlation among breeding values from pedigree-based and genomic models was moderate (~0.70) indicating change on ranking. However, correlations among GEBV from genomic models were higher (>0.90) and it implies no or

small changes in the classification of selection candidates when any of those models are used. Therefore, the best performance of ssGBLUP in this population is using this method as a default (no group-specific adjustments), however, in a scenario where the correlation between across-group and specific-groups allele frequencies is low more investigation is necessary.

Keywords: Allele frequencies, composite beef cattle, $\mathbf{G}$ adjustments, genomic relationship matrices, single-step genomic BLUP

Introduction

The substitution of pedigree-based relationship by the genomic relationship matrix was initially proposed by Nejati-Javaremi et al. (1997). In 2001, Meuwissen et al. (2001) proposed the use of genomic selection to increase the genetic gain in breeding programs. Later, VanRaden (2008) proposed an efficient method to create genomic relationship matrix aiming to increase reliability of estimated breeding values and to estimate thousands of marker effects simultaneously. The main advantage of genomic selection is the higher accuracy of estimated breeding values of young animals, which usually leads to shorter generation interval due to the higher contributions from young genetically superior animals (Matthews et al., 2019).

First genomic methods were called GBLUP and can be summarized an estimation of breeding values with mixed model equations and relationship coefficients deduced from SNP information that are now known as GEBV's (Georges et al., 2019). The GBLUP method allows to estimate GEBVs just for genotyped animals and requires accurate methods to get pseudo phenotypes, which can create bias in genetic evaluation. In this sense, Aguilar et al. (2010) proposed a single-step GBLUP method (ssGBLUP), which allows to eliminate extra steps in GBLUP and combining all available phenotypic, pedigree

and genomic information. Nowadays, ssGBLUP has been widely used for genetic evaluation of most livestock species (Chen et al., 2011;Tsuruta et al., 2013;Lourenco et al., 2015).

In the ssGBLUP method, genomic ($\mathbf{G}$) and pedigree relationship are combined in a joint matrix ($\mathbf{H}$). Traditionally, $\mathbf{G}$ matrix is centered and scaled based on mean allele frequency of genotyped animals. Center and scale of $\mathbf{G}$ matrix are carried out to sets mean values of the allele effects to 0 and try to make $\mathbf{G}$ analogous to that numerator relationship matrix ($\mathbf{A}$), respectively (VanRaden, 2008). Nevertheless, using the across-breed allele frequencies to center and scale $\mathbf{G}$ is suitable just in homogeneous populations(Lourenco et al., 2016). For composite or multi-breed populations a best way to center and scale $\mathbf{G}$ matrix may be using specific-breed allele frequencies.

It is important to point out that any inflation or deflation in the $\mathbf{G}$ matrix can create bias in genetic evaluation due to the wrong weightings between pedigree and genomic information since $\mathbf{H}$ takes to account $\mathbf{G} - \mathbf{A}$ (Forni et al., 2011). Additionally, some authors had related incompatibilities between $\mathbf{G}$ and $\mathbf{A}$ matrices in ssGBLUP as a bias source (Simeone et al., 2012;Makgahlela et al., 2013;Lourenco et al., 2016). The incompatibility among these matrices may be related to small or incomplete pedigrees, genotypes with low quality or wrong, and even to population structure issues omitted in genetic evaluation (Misztal et al., 2013).

According with Lourenco et al. (2016), not taking different allele frequencies for different breeds into account can create a fake relationship between animals, once they can be related through $\mathbf{G}$ when they are not related through $\mathbf{A}$. In broilers, Simeone et al. (2012) showed that the allele frequencies have little impact on rank within the lines, but large impact on ranking across lines. In pigs, Forni et al. (2011) pointed out that the best way to perform evaluations was using current allele frequencies, and scaling $\mathbf{G}$ matrix to

be similar to the numerator relationship matrix for genotyped animals. In this sense, the aim of our study was to compare genomic predictions for Montana composite beef cattle using different adjusted relationship matrix and ssGBLUP method.

Material and Methods

Animal Care and Use Committee approval was not obtained for this study because data set was provided by existing database.

Data

The dataset was provided by Montana Composto Tropical® - CFM Leachman Pecuária Ltda. breeding program. A total of 680,551 purebred, intermediate crossbred and composite animals were available in the pedigree. Genotypes were available for 1899 animals genotyped for 30 (*n* = 1,436), 35 (*n* = 484), and 770 (*n* = 16) thousand SNPs. Subsequently, all these genotypes were imputed to GeneSeek® Genomic Profile BeadChip from Neogen containing over 30,105 markers using FImpute software (Sargolzaei et al., 2014) with an imputation accuracy of 92%. For quality control of genotypes, markers and animals with call rate lower than 0.9, minor allele frequency (MAF) lower than 0.05, departures from Hardy-Weinberg equilibrium (difference between expected and observed frequency of heterozygous) greater than 0.15, with redundant position in the genome and located in non-autosomal chromosomes were removed. After quality control 27,427 SNP and 1899 animals were kept.

Phenotypic records of scrotal circumference at 12 months of age (SC12), postweaning weight gain (PWG; calculated as the difference between weight at 420 days of age and weaning weight), weaning weight (WW; adjusted to 205 days of age) and birth

weight (BW) were available. In the data quality control, phenotypic records deviating from the mean of contemporary groups $\pm$ 3 standard deviations and contemporary groups (GC) with less than five records were removed. The GC were composed by farm, year and season of birth, sex and management group. However, for postweaning traits (SC12 and PWG), the management groups were not taken into account in the GC, but considered as fixed effects. Descriptive statistics of the records used in the analyses are presented in the Table 1.

Table 1. Number of phenotypic records (N), number of contemporary groups (N_{GC}), mean and standard deviation (SD) for scrotal circumference at 12 months of age (SC12), postweaning weight gain (PWG), weaning weight (WW), and birth weight (BW) traits.

Traits	N	N_{GC}	Mean	SD
SC12 (cm)	49,541	958	28.67	3.99
PWG (Kg)	96,994	1,764	60.55	34.26
WW (Kg)	325,014	5,793	190.17	32.33
BW (Kg)	264,961	5,549	33.48	4.87

Structure of genotyped population using principal components

The Montana composite beef cattle are composed by at least three different breeds. Due to the large number of breeds used in this beef cattle, the traditional concept of breed becomes difficult. In this sense, the Montana® breeding program grouped the different breeds used to get these animals according to their likeness of type, function, physiology and aspects of growth and reproduction, that is called biological type. The system of breed proportion identification NABC splits breeds in four groups (biological types) called N, A, B and C. Where: N is *Bos taurus indicus* cattle breeds; A is *Bos taurus*

taurus cattle breeds adapted to tropical conditions; B is *Bos taurus taurus* British breeds and C is European Continental breeds (Ferraz et al., 1999).

With this in mind, it is possible to classify Montana animals based in their biological type or, even, in groups of biological types. These animals can be classified into sixteenths of the breed proportion from the NABC system proposed by breed association. All animals have proportions of biological types summing up to 16 in their composition (i.e., 4:4:4:4, N=4, A=4, B=4, C=4, represents an animal that has the same proportion of all biological types).

In order to verify the population structure of genotyped animals a principal component analysis of the genomic relationship matrix was performed. Population structure can be accounted for in the genomic evaluation by using different allele frequency in $\mathbf{G}$ matrix (Lourenco et al., 2016). In a principal component plot while a single population cluster has an oval shape, clusters indicate lines, breeds or animals genetically different. We clustered genotyped animals based on two approaches. First, we used groups of biological types as proposed by the Montana breeding program. Five groups of biological types were found among genotyped animals (4444; 4624; 4804; 4840 e 4822).

In the Figure 1, the plot of principal components was colored based on the biological type groups.

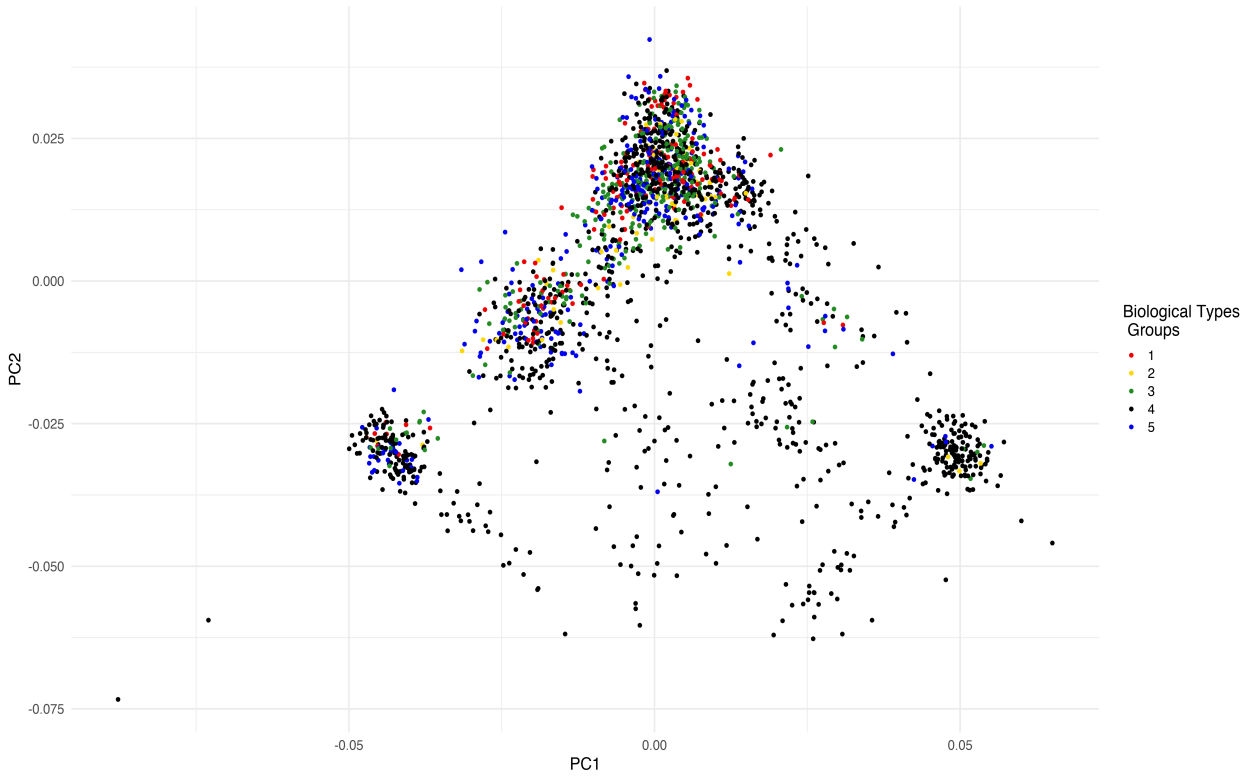

Figure 1. Projection of genomic relationship matrix ($\mathbf{G}$) into two principal components (PC) colored according to the Biological Type Groups.

In an alternative way, we clustered the groups based on a principal component analysis. The packages *NbClust* e *Cluster* in the software R were used to appoint the number of groups based on the Euclidean distance and to shape the groups, respectively. A total of five groups based on principal components (PC) may be depicted from Figure 2.

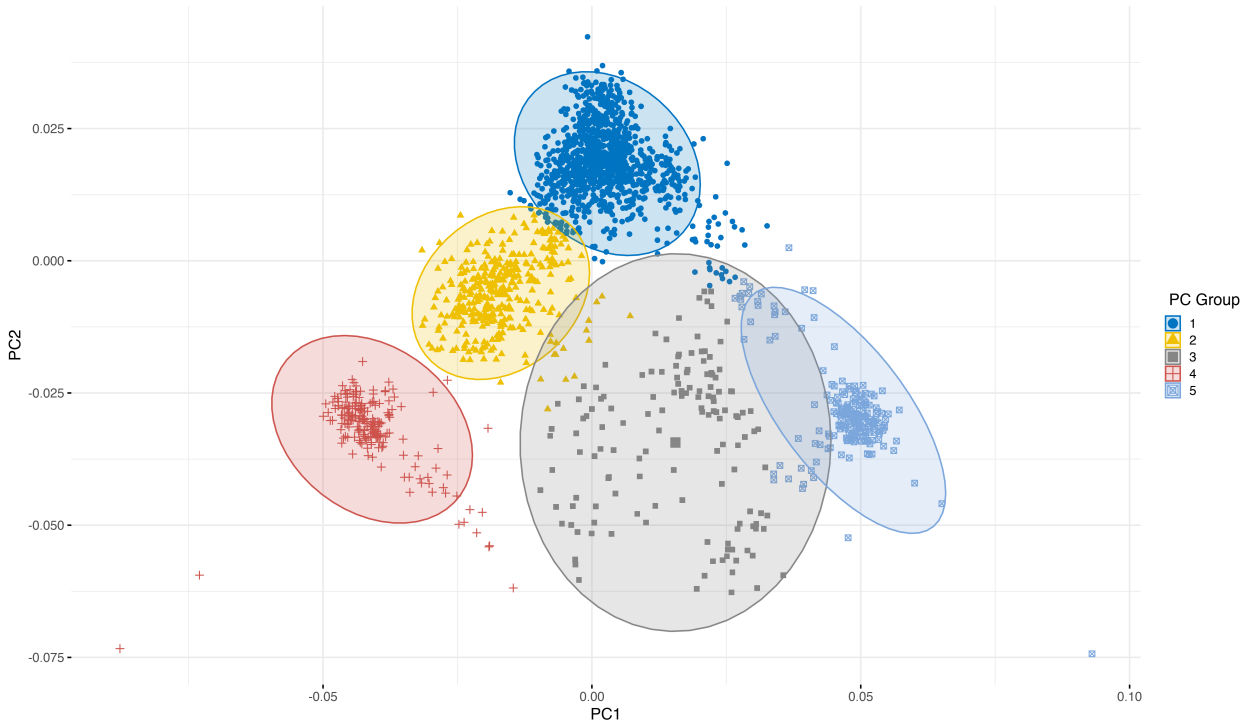

Figure 2. Projection of genomic relationship matrix ($\mathbf{G}$) into two principal components (PC) colored according to the PC Groups.

In the Table 2 are the number of genotyped animals in each group, which were grouped based on biological type groups or principal components analysis.

Table 2. Number of genotyped animals in each group of biological type or principal components.

Group	**1**	**2**	**3**	**4**	**5**
Biological Type	172	60	294	1,065	308
Principal Components	160	350	1,012	199	178

Model

The animal model as follows was adjusted:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1 g} + \mathbf{Z_2 m} + \mathbf{Wc} + \mathbf{e}\ ,$$

In which, $\mathbf{y}$ is the vector of phenotypic records, $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$ and $\mathbf{W}$ are the incidence matrices for $\mathbf{b}$, $\mathbf{g}$, $\mathbf{m}$ and $\mathbf{c}$ vectors, respectively. Where, $\mathbf{b}$ is the vector of fixed effects, $\mathbf{g}$ is the vector of direct genetic additive effects, $\mathbf{m}$ is the vector of maternal genetic additive effects (for WW and BW traits), $\mathbf{c}$ is the vector of maternal permanent environmental effects (for WW and BW traits), and $\mathbf{e}$ is the vector of residual errors associated with each record. As fixed effects, contemporary groups, class of dam's age (for WW and BW) and the management group at weaning for SC12 and PWG were taken into account. The additive genetic effects of biological type composition for individuals (A, B or C), non-additive effects of maternal total heterozygosis ($\mathrm{H_M}$), direct heterozygosis ($\mathrm{N_xA}$, $\mathrm{N_xB}$, $\mathrm{N_xC}$, $\mathrm{A_xB}$, $\mathrm{A_xC}$ and $\mathrm{B_xC}$) and age of measurement (for SC12) were assumed as covariables. To avoid multicollinearity, direct additive effects of the biological type N were not include in the model, *i.e.,* the effects of A, B and C were estimated as deviations from the additive effects of N.

Analysis

Single-trait pedigree-based and genomic evaluations were performed for SC12, PWG, WW and BW traits. For all genomic analysis the ssGBLUP approach was used. In the ssGBLUP method the additive relationship based on pedigree ($\mathbf{A}$) and genomic relationship ($\mathbf{G}$) are combined into a joint matrix $\mathbf{H}$ (Aguilar et al., 2010), as follows:

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

Genomic relationship matrix ($\mathbf{G}$) was built according to VanRaden (2008) as shown below

$$\mathbf{G} = \mathbf{Z}'\mathbf{Z}$$

In which,

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P}) / \left[2 \sum_{j=1}^{n} p_j (1 - p_j) \right]^{1/2},$$

Where, $\mathbf{M}$ is the SNP markers matrix with *n* SNPs for each genotyped animal and, $\mathbf{P}$ is a matrix of two times the allele frequency of the second allele p at locus j (p_j). It is important to point out that both numerator and denominator of the above equation take into account observed allele frequencies across genotyped animals with the denominator being a scaling factor for $\mathbf{G}$. However, for crossbred, multibreed or composite animals, that is heterogeneous populations, likely this approach will produce false relationships. This implies that breeds with different origins are related through $\mathbf{G}$ when they are not through $\mathbf{A}$ (based in recent data), that is, it creates segments identical by state (Lourenco et al., 2016).

In this sense, $\mathbf{G}$ matrix was built in three different ways, hereinafter called $\mathbf{G1}$, $\mathbf{G2}$ and $\mathbf{G3}$. In $\mathbf{G1}$, $\mathbf{G}$ matrix was built as default in ssGBLUP, in which the allele frequencies across genotyped animals are used. In $\mathbf{G2}$, $\mathbf{G}$ matrix was centered based on specific allele frequencies of biological types or PC groups described before. The SNP genotype was centered by 2 times the specific allele frequency of each group of biological type or PC.

In which, each element of $\mathbf{P_K}$ is $[p_{jk}]$, and p_{jk} is the allele frequency of the second allele at locus *j*, being specific for each group of biological type or PC. Thus, $\mathbf{G2}$ is as follows:

$$\mathbf{G2} = \mathbf{Z'Z}, \text{ with } \mathbf{Z} = (\mathbf{M} - \mathbf{P_K})/\left[2\sum_{j=1}^{n} p_j(1 - p_j)\right]^{1/2},$$

For $\mathbf{G3}$, $\mathbf{G}$ matrix was centered and scaled based on group-specific allele frequencies (biological types or PC groups). In this scenario, the numerator of $\mathbf{G3}$ is the same as in $\mathbf{G2}$. However, now the denominator is a scale factor specific for each group. In which, p_{jk} is the frequency of the second allele at locus *j*, being specific for each group. As described below:

$$\mathbf{G3} = \mathbf{Z'Z}, \text{ with } \mathbf{Z} = (\mathbf{M} - \mathbf{P_K})/\left[2\sum_{j=1}^{n} p_{jk}(1 - p_{jk})\right]^{1/2},$$

Variance components and genetic parameters were estimated using restricted maximum likelihood method and software AIREMLF90 (Misztal et al., 2014). Standard deviations of estimates were obtained by resampling of covariances matrices from the asymptotic multivariate normal distribution (Houle and Meyer, 2015). The estimated breeding values (EBV) and genomic estimated breeding values (GEBVs) were obtained from BLUPF90 program.

Validation

Validation was performed to evaluate models' performance. The method of choice for validation was based on linear regression (LR) proposed by Legarra and Reverter (2018). Validation group was formed by 436 young genotyped animals with their

phenotypes removed from evaluation, it was called reduced data. Hereafter, the *r* subscript is used for the reduced data. For complete data, 1899 genotyped animals were used, and for indicating this group we will use the subscript *c*. Reduced and complete data were kept the same for all traits. This validation model was chosen because it is a best way to evaluate maternal models and traits with a low heritability (Legarra and Reverter, 2018).

The stability of (G)EBV in two subsequent evaluations was measured as the correlation between the breeding values of validation animals coming from reduced and complete data, as follows:

$$r = (\mathrm{G})\mathrm{EBV}_r, (\mathrm{G})\mathrm{EBV}_c \ ,$$

The accuracy of (G)EBVs from the reduced data set was estimated as follows:

$$\widehat{acc_{LR}} = \sqrt{\frac{cov(\widehat{u_c},\widehat{u_r})}{(1-\bar{F})\sigma_u^2}} \ ,$$

Where, $\bar{F}$ is the average pedigree inbreeding for the validation animals and σ_u^2 is the additive genetic variance, $\hat{u}_c$ and $\hat{u}_r$ are the vectors of (G)EBV from complete and reduced data, respectively.

Dispersion of (G)EBV (inflation or deflation of (G)EBVs) was measured as the slope of the regression of complete on reduced breeding values for validation animals, as

$$(\mathrm{G})\mathrm{EBV}_\mathrm{c} = \mathrm{b}_0 + \mathrm{b}_1 \times (\mathrm{G})\mathrm{EBV}_\mathrm{r} \ ,$$

Finally, the bias of models was taken as the difference between mean of estimated and true breeding values. As true breeding values are unknown in other studies than simulation, the estimated breeding values for a validation group from complete data was assumed as a true breeding value; and estimated breeding values for a validation group from reduced data as an estimated breeding value. Thus, the bias was obtained as follows:

$$\mu_{cr} = \overline{\hat{u}_r} - \overline{\hat{u}_c} \ ,$$

Additionally, the evaluations' accuracy, performed according to the Beef Improvement Federation (BIF), mean of (G)EBVs and, the spearman correlation across models were evaluated. The animals' accuracy was performed in the BLUPF90 software and it was corrected by inbreeding.

Results and Discussion

Variance components and genetic parameters for SC12, PWG, WW and BW traits are presented in Table 3. The direct heritabilities estimates were moderate for SC12, WW and BW traits (~0.20). While, low estimates of heritability were found for PWG trait (0.09). These results indicate that environmental and genetic non-additive effects are more important in the phenotypic expression of the last trait. In this way, higher gains in SC12, WW and BW can be achieved through direct selection than in PWG trait. While, the improvement in environmental conditions can be more important to phenotypic expression of PWG trait.

Overall, the heritabilities estimates across models did not change but for SC12 trait slight changes were noted. However, these changes are not enough to assert that different relationship matrices were able to capture more proportion of phenotypic variance explained by genetic additive variance, that is heritability for that trait. Additionally, the additive and residual variances, maternal heritabilities and maternal permanent environmental effects, as well as their standard errors were very similar across all relationship matrices.

Overall, genetic additive variances were a little bit higher in genomic models than in pedigree-based, while the residual variances were a bit lower, but heritability estimates did not change for PWG, WW and BW. In agreement with our results, working with a swine population, Forni et al. (2011) also reported slight changes in the variances (genetic and residual), when combining full data set and different relationships matrices. In contrast, those authors found an increase in the genetic additive variance for a subset of genotyped animals. In this sense, we can confirm that small changes are expected when genomic relationships are a small fraction of all relationships in the dataset.

Higher heritabilities' estimates for SC12 and PWG (0.43 and 0.14, respectively) were reported by Biegelmeyer et al. (2017) for Hereford and Braford beef cattle. Also, in a composite beef cattle population Santana et al. (2013) found higher mean heritability for PWG trait (0.18). These authors reported a variation from 0.01 to 0.43 for that trait in a reaction norm study. The large variation in heritability for the same trait found by Santana et al. (2013) can also explains the variability across studies.

Similarly, the heritability estimates for WW and BW reported in this study were lower than those reported by others authors. Schiermiester et al. (2015) in a study with a multi-breed population found heritabilities of 0.22 and 0.42 for WW and BW, respectively. In Nelore cattle, Kluska et al. (2018) reported direct heritability of 0.22 for WW. Maternal

heritability and permanent environmental effects were bigger for WW than for BW. However, these values were of low magnitude for both traits (Table 3).

For maternal heritability and permanent environmental effects similar results were showed in the literature with others breeds. In a population of Brahman beef cattle Kamprasert et al. (2019) reported maternal heritabilities of 0.05 and 0.02 to birth weight and weight at 200 days of age, respectively. In the same way, Kluska et al. (2018) found slight changes in the maternal heritability and maternal permanent environmental effects for WW when compared to this study. For crossbred population higher values of maternal heritability and maternal permanent environmental effects were found for WW (0.17 and ~0.23, respectively) (Schiermiester et al., 2015).

Overall, the estimates of heritability showed in this study for all traits and relationship matrices were slight lower than other studies. It can be related to several factors as the population structure, model used in variance components estimation and the decreasing of genetic variance in population under selection over time (Hidalgo et al., 2020).

Table 3. Estimates of additive (σ_a^2) and residual (σ_e^2) variances, direct heritability (h_d^2), maternal heritability (h_m^2) and maternal permanent environmental effects (c^2) (± standard deviation) for all traits based on different relationship matrices.

	SC12		
	σ_a^2	σ_e^2	h^2
BLUP	1.78±0.14	6.81±0.12	0.21±0.01
G1	1.92±0.14	6.70±0.12	0.22±0.02
G2_BIO	1.95±0.14	6.67±0.12	0.23±0.02
G2_PC	1.95±0.14	6.67±0.12	0.23±0.02
G3_BIO	1.95±0.14	6.67±0.12	0.23±0.02
G3_ PC	1.95±0.14	6.67±0.12	0.23±0.02

	PWG				
	σ_a^2	σ_e^2	h^2		
BLUP	25.46±2.13	258.64±2.06	0.09±0.01		
G1	26.50±2.13	257.79±2.05	0.09±0.01		
G2_BIO	26.25±2.12	257.95±2.04	0.09±0.01		
G2_PC	26.32±2.12	257.94±2.05	0.09±0.01		
G3_BIO	26.24±2.12	257.96±2.04	0.09±0.01		
G3_PC	26.35±2.13	257.92±2.04	0.09±0.01		
	WW				
	σ_a^2	σ_e^2	h_d^2	h_m^2	c^2
BLUP	92.68±3.43	320.49±2.33	0.18±0.01	0.06±0.00	0.13±0.00
G1	91.84±3.42	320.82±2.33	0.18±0.01	0.06±0.00	0.13±0.00
G2_BIO	91.66±3.42	320.92±2.33	0.18±0.01	0.06±0.00	0.13±0.00
G2_PC	91.74±3.42	320.87±2.32	0.18±0.01	0.06±0.00	0.13±0.00
G3_BIO	91.66±3.42	320.92±2.33	0.18±0.01	0.06±0.00	0.13±0.00
G3_PC	91.78±3.42	320.84±2.32	0.18±0.01	0.06±0.00	0.13±0.00
	BW				
	σ_a^2	σ_e^2	h_d^2	h_m^2	c^2
BLUP	2.72±0.11	12.10±0.06	0.17±0.01	0.03±0.00	0.04±0.00
G1	2.75±0.11	12.08±0.08	0.17±0.01	0.03±0.00	0.04±0.00
G2_BIO	2.76±0.11	12.07±0.08	0.17±0.01	0.03±0.00	0.04±0.00
G2_PC	2.78±0.11	12.06±0.08	0.17±0.01	0.03±0.00	0.04±0.00
G3_BIO	2.76±0.11	12.07±0.08	0.17±0.01	0.03±0.00	0.04±0.00
G3_PC	2.78±0.11	12.06±0.08	0.17±0.01	0.03±0.01	0.04±0.00

BLUP: pedigree-based matrix ($\mathbf{A}$); **G1**: the across-groups genomic relationship matrix; **G2_BIO**: the genomic relationship matrix centered by specific-biological type groups allele frequencies; **G2_PC**: the genomic relationship matrix centered by specific-PC groups allele frequencies; **G3_BIO**: the genomic relationship matrix centered and scaled by specific-biological type group allele frequencies; **G3_PC**: the genomic relationship matrix centered and scaled by specific-PC groups allele frequencies.

There are two kinds of (G)EBV precision. First is measured by relating the PEV (prediction error variance) of (G)EBV to the base population genetic additive variance, which quantifies the magnitude of PEV in relation to the base additive genetic variance. This measure is used to reflect the extend which (G)EBV will change when more information becomes available and it is call evaluations' accuracy. On the other hand, the other measure of accuracy is the correlation between true and estimated breeding values. This may be understood as the response to selection that can be achieved when individuals are selected from those (G)EBV, since the response to selection depends on the accuracy. These accuracies are presents in the Table 4 and Figure 3, respectively.

In the Figure 3 are presented the accuracy of (G)EBV from reduced data based on different relationship matrices. The accuracy of (G)EBVs across models ranged from 0.47 to 0.54 for SC12, 0.40 to 0.47 to PWG, 0.42 to 0.45 for WW and 0.55 to 0.69 for BW trait. Higher accuracies were observed for BW followed by SC12. Accuracies for PWG and WW traits were very similar. An increasing in accuracy was observed when genomic information was added to the models for traits SC12 and BW (~14.9 and 23.6 percent, respectively). While, the accuracy of PWG and WW traits has decreased (till ~14.9 and 6.7 percent, respectively, for models with lower accuracies) when the genomic information was taken into account.

The (G)EBV accuracy across all models based on genomic relationship matrices either $\mathbf{G1}$, $\mathbf{G2}$ or $\mathbf{G3}$ with biological types or PC groups was very similar (slight changes, till 0.04 across models). However, overall, when the biological types were taken into account to building $\mathbf{G}$ matrix the accuracy was higher than when PC groups were used, but it was very similar to the across-groups relationship. These results indicate that there is no gain in accuracy by using biological types groups to center and scale the genomic relationship matrix in this population.

Lourenco et al. (2016) showed that accuracies for traditional BLUP decreased for traits under selection when compared to simulated scenarios with no selection and selection based on higher phenotypic values; and with a real population. These authors pointed out that when the simulated population was undergoing selection based on EBV, accuracies of GEBV using adjusted relationship matrices were even less than the accuracy for traditional BLUP. Our results are in agreement with this, since we found higher accuracies for traits with no selection (BW) than for those under selection (WW) and similar heritability (Figure 3). We also found higher accuracy for EBV from traditional BLUP than those from all genomic models in traits under selection (PWG and WW).

Accounting for group-specific allele frequencies did not cause changes in $\mathbf{G}$ able to produce improvement as reported previously by other authors in pigs and dairy cattle (Makgahlela et al., 2013;Lourenco et al., 2016). It is important to point out that when the parental generation is genotyped adjusting $\mathbf{G}$ matrix has little or no impact in GEBV because genomic information is just a small amount of information. In contrast, for young animals, genomic information and parental average are the only sources of information. Nevertheless, even in young animals we didn't find impact on GEBV when $\mathbf{G}$ was adjusted based on biological types groups. Population accuracy is important to maximize genetic progress and compare predictive ability of models (Legarra and Reverter, 2018). While, individuals' accuracy is a measure of the risk when chose a particular animal for breeding.

The accuracy based on PEV is important to selling sires and when semen or embryos are marketed. Accuracy for all traits in all models was very low (0.11 to 0.18) because validation animals are young animals without own phenotypes and just few of those had progeny information. The increasing in evaluations' accuracy by adding genomic information was around 20 percent in relation to pedigree-based model for all traits.

Table 4. Average of evaluations' BIF accuracy for validation animals and standard deviations (between brackets), based on different relationship matrix.

	SC12	PWG	WW	BW
BLUP	0.11 (0.05)	0.11 (0.05)	0.15 (0.05)	0.14 (0.06)
G1	0.14 (0.04)	0.13 (0.04)	0.18 (0.04)	0.17 (0.04)
G2_BIO	0.14 (0.03)	0.13 (0.03)	0.17 (0.03)	0.17 (0.04)
G2_PC	0.13 (0.03)	0.12 (0.03)	0.16 (0.03)	0.16 (0.04)
G3_BIO	0.14 (0.03)	0.13 (0.03)	0.17 (0.03)	0.17 (0.04)
G3_PC	0.13 (0.03)	0.12 (0.03)	0.16 (0.03)	0.16 (0.04)

BIF: Beef Improvement Federation methodology to calcule the accuracy; **BLUP**: pedigree-based matrix ($\mathbf{A}$); **G1**: the across-groups genomic relationship matrix; **G2_BIO**: the genomic relationship matrix centered by specific-biological type groups allele frequencies; **G2_PC**: the genomic relationship matrix centered by specific-PC groups allele frequencies; **G3_BIO**: the genomic relationship matrix centered and scaled by specific-biological type group allele frequencies; **G3_PC**: the genomic relationship matrix centered and scaled by specific-PC groups allele frequencies.

Slights or no changes in evaluations' accuracy from $\mathbf{G1}$ model were observed when the adjusted $\mathbf{G}$ matrices were used to built $\mathbf{H}$. However, similarly to the accuracy of models, but now for all traits, we observed that when PC groups were used to center or to center and scale $\mathbf{G}$, the average of evaluations' accuracy was a little bit lower (0.01 to 0.02) than other genomic models. These results indicate that changes in GEBV of those animals (validation animals) should be big when more information becomes available even with genomic data.

Evaluations' accuracy is related to several factors as trait heritability, number of phenotypic records in the individual and its relatives and number of records on any traits that are correlated genetically (Van Eenennaam, 2011). As heritability cannot be directly influenced the increasing in the number of close relatives, since these shares many of the

same genes, is the easiest way to increase the animals' accuracy. Another way is the addition of genomic information into genetic evaluation; however, the amount of increasing is limited by the genetic variance explained by markers (Van Eenennaam, 2011).

The reason that markers information is able to increase the accuracy of GEBV is that they account the mendelian sampling. In contrast to our results, Forni et al. (2011) found an increase in accuracy from genomic information for genotyped animals using different **G** matrices, however the increasing was higher for females due to their lower initial accuracy. It is worth pointing out that the increase in accuracy by using different **G** matrices was not observed for all animals (genotyped and non-genotyped).

The stability of GEBVs is presented in Table 5. Stability ranged from 0.79 to 0.84 for SC12, 0.75 to 0.82 for PWG, 0.73 to 0.76 for WW and 0.87 to 0.92 for BW. Overall, the stability increased for all traits but PWG when the genomic information was taken into account. Additionally, for PWG when PC groups were used to center or center and scale **G** matrix the stability was even lower than other genomic methods. Similarly, the p stability of all models based on PC groups was lower than across-groups genomic relationships and specific-biological type groups allele frequencies.

Table 5. stability of (G)EBVs from two subsequent evaluations based on different relationship matrices.

	SC12	PWG	WW	BW
BLUP	0.79	0.82	0.73	0.87
G1	0.84	0.79	0.76	0.92
G2_BIO	0.84	0.80	0.74	0.92

G2_PC	0.83	0.75	0.73	0.91
G3_BIO	0.84	0.80	0.74	0.92
G3_PC	0.83	0.75	0.73	0.91

BLUP: pedigree-based matrix ($\mathbf{A}$); **G1**: the across-groups genomic relationship matrix; **G2_BIO**: the genomic relationship matrix centered by specific-biological type groups allele frequencies; **G2_PC**: the genomic relationship matrix centered by specific-PC groups allele frequencies; **G3_BIO**: the genomic relationship matrix centered and scaled by specific-biological type group allele frequencies; **G3_PC**: the genomic relationship matrix centered and scaled by specific-PC groups allele frequencies.

The stability was the same for $\mathbf{G2}$ or $\mathbf{G3}$ within traits and groups (biological types or PC) used to adjust $\mathbf{G}$. The stability of models based on biological types using either $\mathbf{G2}$ or $\mathbf{G3}$ was the same as $\mathbf{G1}$ for SC12 and BW. Higher values of stability in across-group genomic relationship matrix than models with biological types were found for WW, whereas for PWG trait $\mathbf{G1}$ showed a lower stability than the models based on biological types.

Similar to our study, in a study with broilers, Chen et al. (2011) found similar values of stability for all $\mathbf{G}$ used, however, these authors called stability predictive ability. The similarity across all $\mathbf{G}$ matrices tested in our study may be related to the correlation of allele frequencies between groups (Figure 4). The correlations of allele frequencies among biological type groups were high (~0.90) while for PC groups those were lower, ranging from 0.35 to 0.87. However, the correlations of allele frequencies either within biological type groups (BTG) or principal component groups (PCG) with the average across-group did not differ considerably. Likely this is the main reason of the similarity across genomic models.

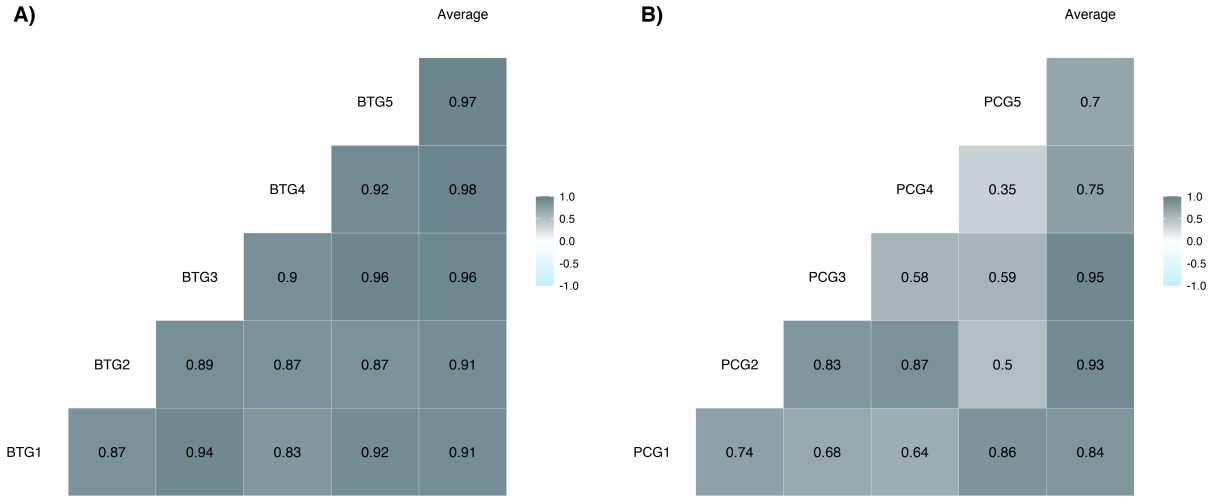

Figure 4. Correlations of allele frequencies among five biological types groups (BTG) (**A**) or principal components groups (PCG) (**B**); and the average allele frequency.

The low correlations among groups or breeds used to center and scale $\mathbf{G}$ matrix may justify the need for breed/group-specific allele frequencies (Lourenco et al., 2016). However, even in a scenario with low correlations (PC groups) among the groups we did not find big advantages in using group-specific allele frequencies. Similar results were also reported by Lourenco et al. (2016) and it is likely due to the high correlation between groups and across-breed allele frequencies reported in both studies. Correlations between allele frequencies were not reported by other authors (Chen et al., 2011;Simeone et al., 2012;Makgahlela et al., 2014).

Dispersion of models is a feature used to compare them. It is desirable regression coefficient (b1) close to 1 to ensure that predictions are in the same scale. The slope or dispersion is presented in Table 6. Dispersion ranged from 1.06 to 1.14 for SC12, 0.94 to 0.98 for PWG, 0.82 to 1.0 for WW and 1.03 to 1.08 for BW. For all traits but WW, the additive relationship matrix ($\mathbf{A}$) yielded b1 closer to 1 when compared to genomic models, indicating that to add genomic information in the models may increase the inflation.

Table 6. Regression coefficient of (G)EBV using complete data on (G)EBV using reduced data for validation animals when (G)EBV were estimated using different relationship matrices.

	SC12	PWG	WW	BW
BLUP	1.06	0.98	0.82	1.03
G1	1.16	0.96	1.00	1.08
G2_BIO	1.14	0.96	0.98	1.07
G2_PC	1.13	0.94	0.96	1.08
G3_BIO	1.14	0.96	0.97	1.07
G3_PC	1.13	0.94	0.96	1.08

BLUP: pedigree-based matrix ($\mathbf{A}$); **G1**: the across-groups genomic relationship matrix; **G2_BIO**: the genomic relationship matrix centered by specific-biological type groups allele frequencies; **G2_PC**: the genomic relationship matrix centered by specific-PC groups allele frequencies; **G3_BIO**: the genomic relationship matrix centered and scaled by specific-biological type group allele frequencies; **G3_PC**: the genomic relationship matrix centered and scaled by specific-PC groups allele frequencies.

Dispersion estimations were quite similar among genomic models. The models with $\mathbf{G}$ adjusted by PC groups had a slightly higher dispersion than other models, except for SC12 trait, in which $\mathbf{G1}$ showed the greatest slope. In the same way, for BW trait the dispersion of models with $\mathbf{G}$ adjusted by PC groups was as high as in the $\mathbf{G1}$. While for PWG the models with $\mathbf{G}$ adjusted by biological type groups and $\mathbf{G1}$ had similar and lower dispersion than models using PC groups. Across genomic models $\mathbf{G1}$ showed no over/under dispersion for WW trait.

Using adjusted $\mathbf{G}$ instead of across-groups allele frequencies yielded changes from up to 3 unites in b1. Similar results were also found by Makgahlela et al. (2014) in a dairy

cattle population when observed allele frequencies were used. According to the same authors, using base population allele frequency this difference is increased. The expectation of b1 may not be 1.0 when genotyped animals have been selected based on their EBVs, which is, those are not a representative sample of their comparison group, therefore the expectation of b1 is reduced (Mäntysaari et al., 2009;Legarra and Reverter, 2018). According with Mäntysaari et al. (2009) inflation of EBV (b1<1.0) is a frequent phenom observed in dairy cattle.

It is important to point out that inflation does not change the ranking of animals in the same generation, however, if the selection is performed across generation it becomes important since young animals can be considered genetically superior to proven candidates when these are not (Piccoli et al., 2018). Vitezica et al. (2011) showed increasing in inflation with values lower than 1.0, when strong selection is performed.

The bias of models was measured as the difference between (G)EBVs from reduced data and (G)EBV from complete data. Results of bias are presented in the Table 7. For all models the bias was negative, indicating that (G)EBV from reduced data was lower than from complete. These results denote that (G)EBVs of young animal would be underestimated since the estimation of (G)EBVs from reduced is lower than from complete data. Bias ranged from -0.04 to -0.12 for SC12, -0.31 to -0.55 for PWG, -1.40 to -1.96 for WW and -0.05 to 0.09 for BW. For all traits, pedigree-based relationship BLUP was the less biased model.

Across genomic models, slight changes on bias were observed. For PWG and WW traits, the models in which PC groups were used to adjust $\mathbf{G}$ matrix showed lower bias among genomic models, while for SC12 and BW use $\mathbf{G}$ adjusted by biological type groups yielded lower bias. However, it is important to point out that although there is change these

are not significant except for PWG in which changes are at least by 7 points (models based on PC were ~13% less biased than $\mathbf{G1}$).

Table 7. Bias of (G)EBVs estimated with different relationship matrices.

	SC12	PWG	WW	BW
BLUP	-0.04	-0.31	-1.40	-0.05
G1	-0.11	-0.55	-1.95	-0.09
G2_BIO	-0.10	-0.53	-1.96	-0.08
G2_PC	-0.12	-0.45	-1.93	-0.09
G3_BIO	-0.10	-0.52	-1.96	-0.08
G3_PC	-0.12	-0.45	-1.92	-0.09

BLUP: pedigree-based matrix ($\mathbf{A}$); **G1**: the across-groups genomic relationship matrix; **G2_BIO**: the genomic relationship matrix centered by specific-biological type groups allele frequencies; **G2_PC**: the genomic relationship matrix centered by specific-PC groups allele frequencies; **G3_BIO**: the genomic relationship matrix centered and scaled by specific-biological type group allele frequencies; **G3_PC**: the genomic relationship matrix centered and scaled by specific-PC groups allele frequencies.

Unbiased predictions are essential to get accurate estimation of genetic trends. If predictions are biased upwards the genetic trend will be overestimated, benefiting young animals. Vitezica et al. (2011) showed that even under selection pedigree-based BLUP models can predict EBV correctly. However, we found some bias even in pedigree-based BLUP, but it was always lower than genomic models. On the other hand, similar to our results, these authors showed underestimation of TBV/(G)EBV in the last generation with ssGBLUP method.

We further investigated the (G)EBVs with all matrices tested by summarizing those in the validation dataset. Boxplots of (G)EBVs based on different relationship matrices are presented in Figure 5. For each trait, six boxplots are displayed for each combination of

trait and relationship matrices used. Overall, the distribution of (G)EBVs across models was very similar, however the median of BLUP model seem be shifted downwards for all traits. Likely, the superiority of (G)EBVs' median, or at least part of that, is due to the bias introduced by genomic information. Within genomic models, similar or lower variability of (G)EBVs for validation animals was observed using PC groups to adjust $\mathbf{G}$. However, these models seem to produce a larger number of extreme values of (G)EBVs (outliers) than other genomic models (Figure 5).

We also investigate the rank correlation across validation animals aiming verify the similarity in the selection candidates across models. Spearman's correlation among (G)EBVs is presented in Figure 6. Spearman's correlations ranged from 0.68 to 1.0 for SC12, 0.61 to 1.0 for PWG, 0.44 to 1.0 for WW and 0.75 to 1.0 for BW. Overall, correlations among pedigree-based and genomic models were lower than across genomic models. Correlations among pedigree-based and genomic models ranged across traits, ranging from 0.44 to 0.79, while correlations among genomic models were higher than 0.89.

A correlation of 1.0 was observed between $\mathbf{G2}$ and $\mathbf{G3}$ within the same strategy used to adjust $\mathbf{G}$ (PC or biological types). These results indicate that there is no difference in the rank of selection candidates if $\mathbf{G}$ matrix is only centered or centered and scaled based on specific-group allele frequencies. Additionally, correlations among genomic models show that minor changes in animals' rank should be observed by using any of those genomic models, however, rank would change considerably if the evaluation is performed with pedigree-based or genomic models.

Conclusion

Take genomic information into account provides higher stability of (G)EBVs in two subsequent evaluations. Account to specific-group allele frequencies to build $\mathbf{G}$ matrix

using either biological types or PC groups may not provide benefits in the genetic evaluation of Montana composite beef cattle since change in accuracy, stability, dispersion and bias were small. These results are likely due to the high correlation between across-group and specific-groups allele frequencies.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T.J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93**,** 743-752.

Biegelmeyer, P., Gomes, C., Roso, V., Laurino, D., and Cardoso, F. (2017). Tick resistance genetic parameters and its correlations with production traits in Hereford and Braford cattle. *Livestock Science* 202.

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., and Muir, W.M. (2011). Effect of different genomic relationship matrices on accuracy and scale. *Journal of Animal Science* 89**,** 2673-2679.

Ferraz, J.B.S., Eler, J.P., and Golden, B.L. (1999). Análise genética do composto Montana Tropical. *Revista Brasileira de Reprodução Animal* 23**,** 111-113.

Forni, S., Aguilar, I., and Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* 43**,** 1.

Georges, M., Charlier, C., and Hayes, B. (2019). Harnessing genomic information for livestock improvement. *Nature Reviews Genetics* 20**,** 135-156.

Hidalgo, J., Tsuruta, S., Lourenco, D., Masuda, Y., Huang, Y., Gray, K.A., and Misztal, I. (2020). Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. *Journal of Animal Science* 98.

Houle, D., and Meyer, K. (2015). Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. *Journal of Evolutionary Biology* 28**,** 1542-1549.

Kamprasert, N., Duijvesteijn, N., and Van Der Werf, J.H.J. (2019). Estimation of genetic parameters for BW and body measurements in Brahman cattle. *Animal* 13**,** 1576-1582.

Kluska, S., Olivieri, B.F., Bonamy, M., Chiaia, H.L.J., Feitosa, F.L.B., Berton, M.P., Peripolli, E., Lemos, M.V.A., Tonussi, R.L., Lôbo, R.B., Magnabosco, C.D.U., Di Croce, F., Osterstock, J., Pereira, A.S.C., Munari, D.P., Bezerra, L.A., Lopes, F.B., and Baldi, F. (2018). Estimates of genetic parameters for growth, reproductive, and carcass traits in Nelore cattle using the single step genomic BLUP procedure. *Livestock Science* 216**,** 203-209.

Legarra, A., and Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution* 50**,** 53.

Lourenco, D.a.L., Tsuruta, S., Fragomeni, B.O., Chen, C.Y., Herring, W.O., and Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased

predictor using adjusted realized relationship matrices1. *Journal of Animal Science* 94**,** 909-919.

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., and Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus1. *Journal of Animal Science* 93**,** 2653-2662.

Makgahlela, M.L., Strandén, I., Nielsen, U.S., Sillanpää, M.J., and Mäntysaari, E.A. (2013). The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. *Journal of Dairy Science* 96**,** 5364-5375.

Makgahlela, M.L., Strandén, I., Nielsen, U.S., Sillanpää, M.J., and Mäntysaari, E.A. (2014). Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. *Journal of Dairy Science* 97**,** 1117-1127.

Mäntysaari, E., Liu, Z., and Vanraden, P. (2009). Interbull Validation Test for Genomic Evaluations. *Interbull Bull* 41.

Matthews, D., Kearney, J.F., Cromie, A.R., Hely, F.S., and Amer, P.R. (2019). Genetic benefits of genomic selection breeding programmes considering foreign sire contributions. *Genetics Selection Evolution* 51**,** 40.

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

Misztal, I., Aggrey, S.E., and Muir, W.M. (2013). Experiences with a single-step genome evaluation. *Poult Sci* 92**,** 2530-2534.

Misztal, I., Tsuruta, S., Lourenco, D.a.L., Masuda, Y., Aguilar, I., Legarra, A., and Vitezica, Z.G. (2014). *Manual for BLUPF90 family of programs* [Online]. Available: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf [Accessed].

Nejati-Javaremi, A., Smith, C., and Gibson, J.P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J Anim Sci* 75**,** 1738-1745.

Piccoli, M., Brito, L., Braccini Neto, J., Brito, F., Cardoso, F., Cobuci, J., Sargolzaei, M., and Schenkel, F. (2018). A comprehensive comparison between single- and two-step GBLUP methods in a simulated beef cattle population. *Canadian Journal of Animal Science* 98.

Santana, M.L., Eler, J.P., Cardoso, F.F., Albuquerque, L.G., and Ferraz, J.B.S. (2013). Phenotypic plasticity of composite beef cattle performance using reaction norms model with unknown covariate. *animal* 7**,** 202-210.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC genomics* 15**,** 478-478.

Schiermiester, L., Thallman, R., Kuehn, L., Kachman, S., and Spangler, M. (2015). Estimation of Breed-specific Heterosis Effects for Birth, Weaning and Yearling Weight in Cattle. *Journal of animal science* 93**,** 46-52.

Simeone, R., Misztal, I., Aguilar, I., and Vitezica, Z.G. (2012). Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *Journal of Animal Breeding and Genetics* 129**,** 3-10.

Tsuruta, S., Misztal, I., and Lawlor, T.J. (2013). Short communication: Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *Journal of Dairy Science* 96**,** 3332-3335.

Van Eenennaam, A.L. (2011). Improving EPD accuracy by combining EPD information with DNA test results. *Proceedings, Applied Reproductive Strategies in Beef Cattle*, 31.38-31.39.

Vanraden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91, 4414-4423.

Vitezica, Z.G., Aguilar, I., Misztal, I., and Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research* 93, 357-366.

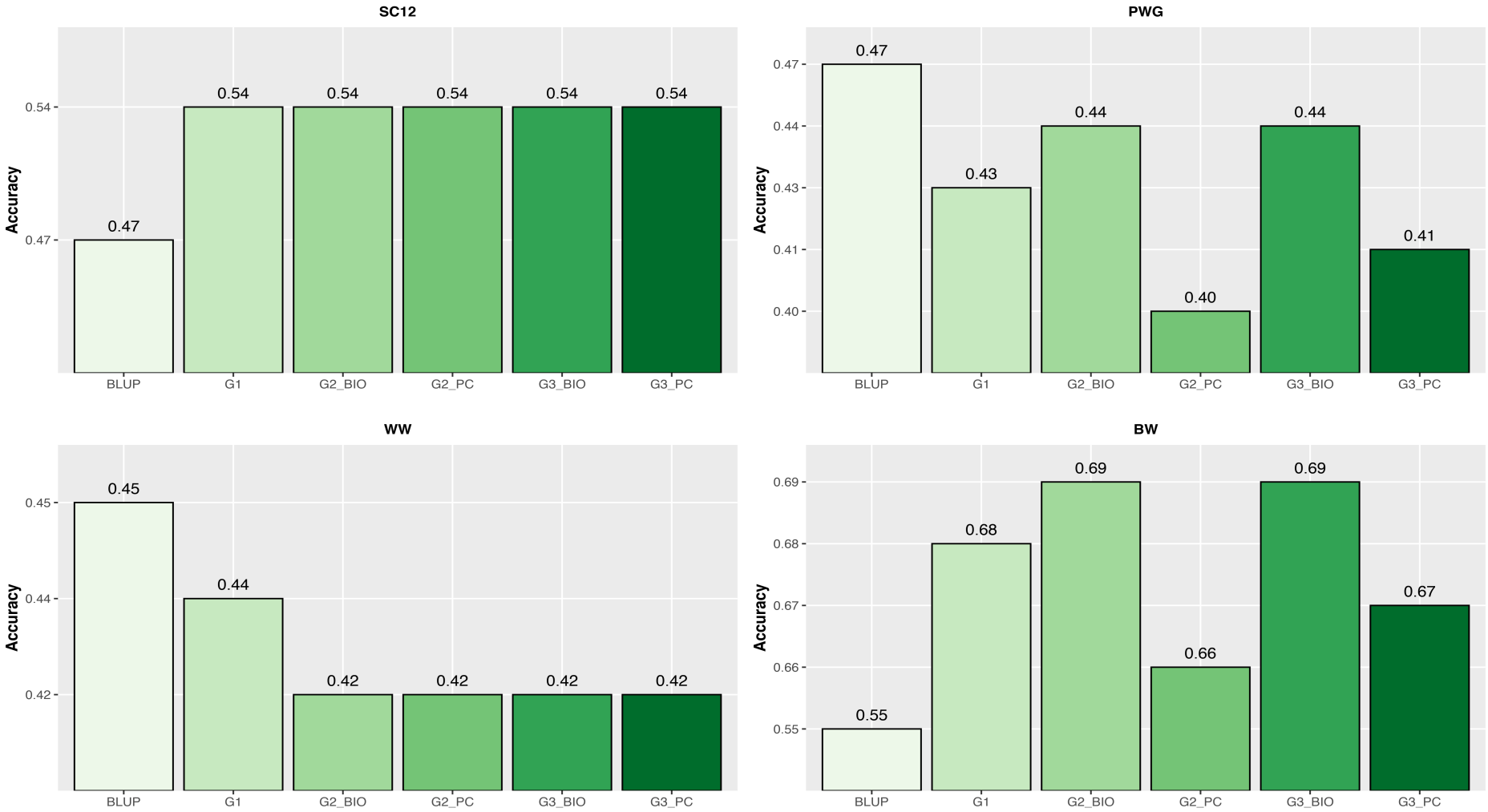

Figure 3. Accuracy of (G)EBVs from reduced data based on different relationship matrices.

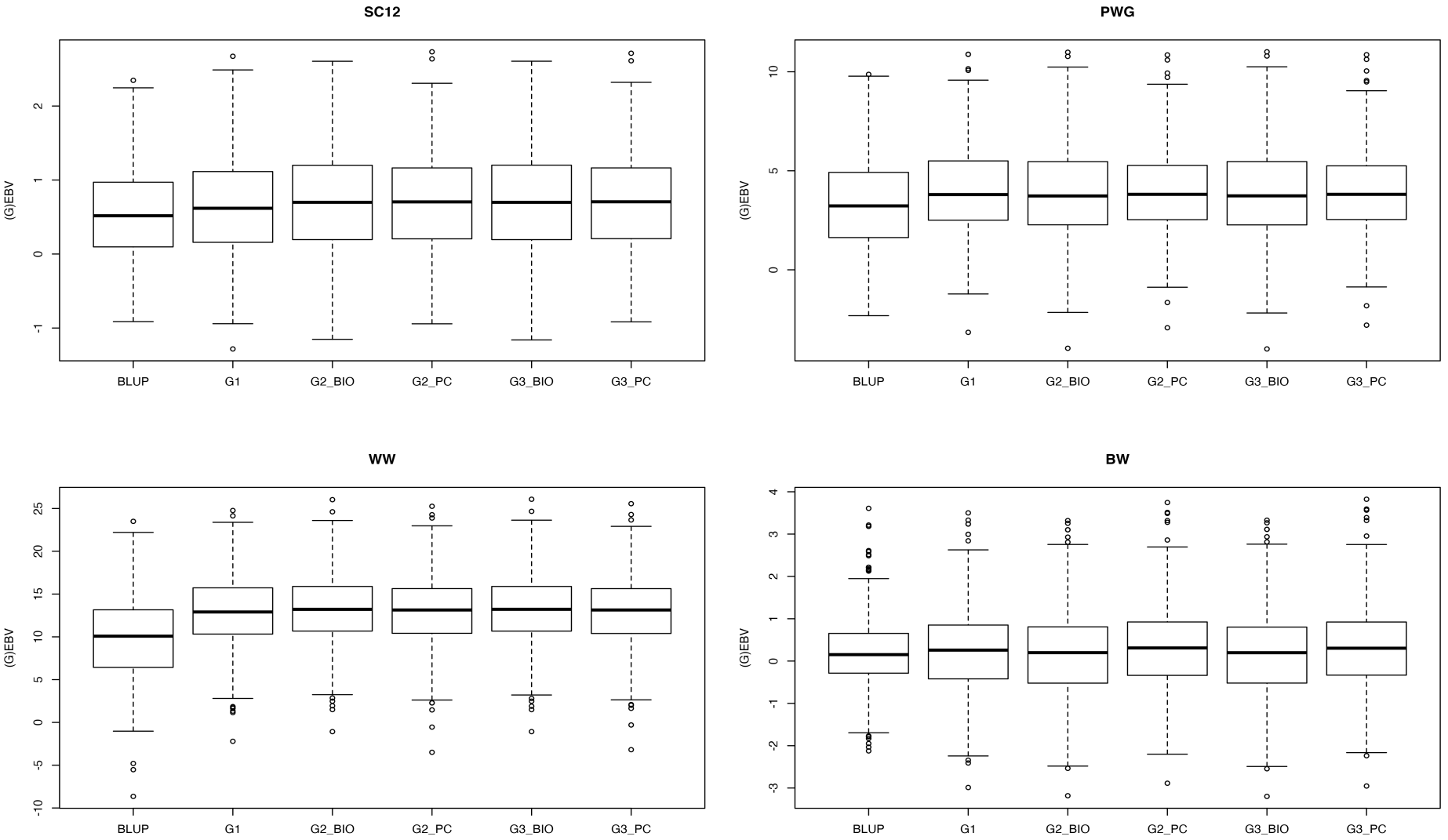

Figure 5. Boxplot of (G)EBVs for validation animals based on different relationship matrices. The box represents the first and third quartiles, the line is the median, the whiskers extend to the furthest point up to 1.5 times the interquartile range, and circles are for outliers beyond that distance.

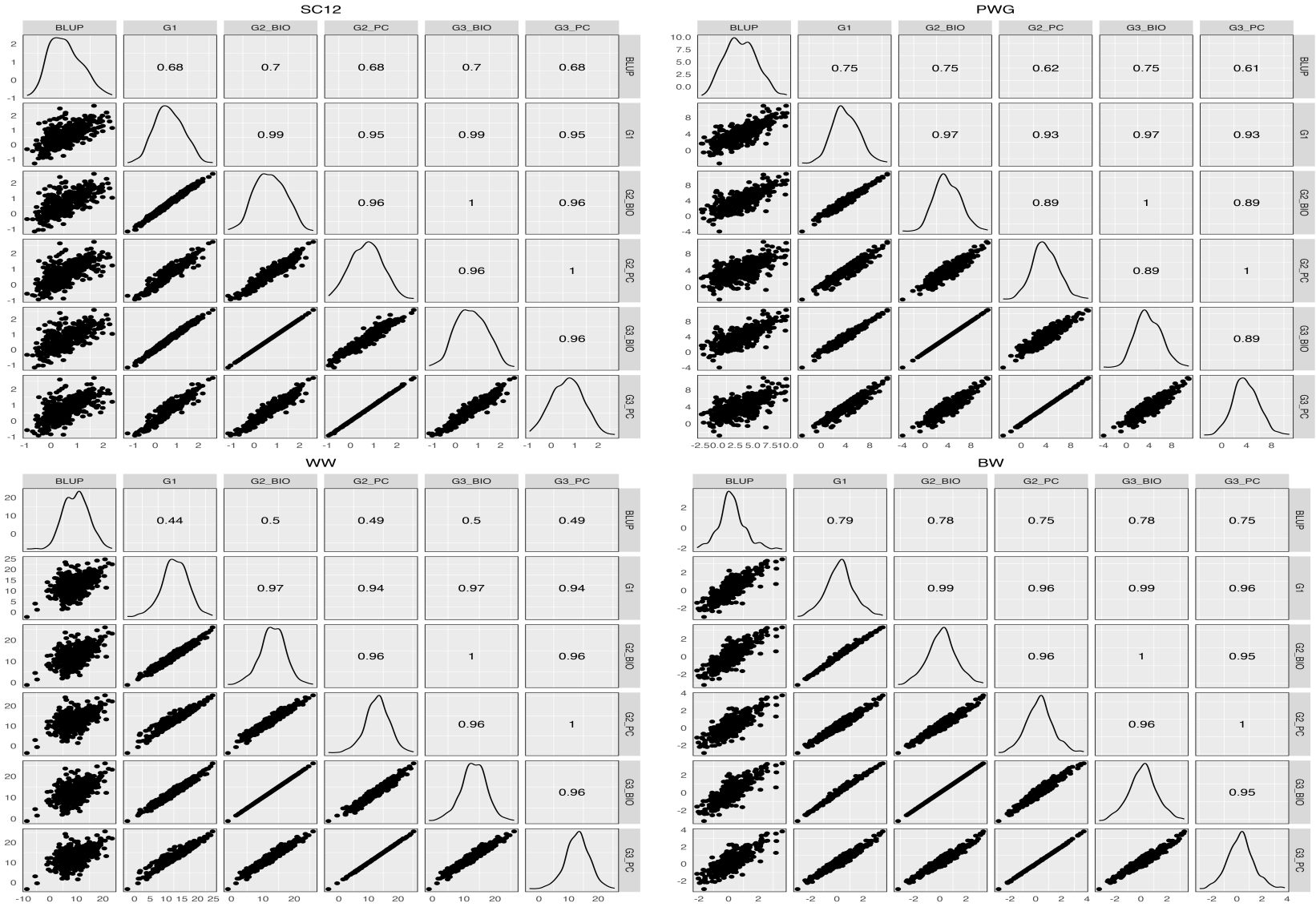

Figure 6. Spearman correlation of (G)EBVs across models with different relationship matrices.

CHAPTER 3 - Metafounders may reduce bias in composite cattle genomic predictions

ABSTRACT - Metafounders are pseudo-individuals that act as proxies for the animals in the base populations. When metafounders are used, individuals from different breeds can be related though the pedigree, improving the compatibility between genomic and pedigree relationships. The aim of this study was to investigate the use of metafounders and unknown parent groups in the genomic evaluation of a composite beef cattle population. Phenotypes were available for scrotal circumference at 12 months of age (SC12), postweaning gain (PWG), weaning weight (WW), and birth weight (BW). The pedigree included 680,551 animals, of which 1899 were genotyped for or imputed to around 30,000 SNPs. Evaluations were performed based on pedigree (BLUP), pedigree with unknown parent groups (BLUP_UPG), pedigree with metafounders (BLUP_MF), single-step genomic BLUP (ssGBLUP), ssGBLUP with unknown parent groups for genomic and pedigree relationship matrices (ssGBLUP_UPG) or only for the pedigree relationship matrices (ssGBLUP_UPGA), and ssGBLUP with metafounders (ssGBLUP_MF). Each evaluation considered either four or 10 groups that were assigned based on breed of founders and intermediate crosses. To evaluate the model performance, we used a validation method based on linear regression statistics to obtain accuracy, stability of (G)EBV, dispersion, and bias. Overall, relationships within and among metafounders were stronger in the scenario with 10 metafounders. Accuracy was greater for models with genomic information compared to BLUP. Also, the stability of (G)EBVs was greater when genomic information was taken into account. Overall, pedigree-based methods had no inflation/deflation (slight differences from 1.0), except for WW that had regression coefficients from 0.77 to 0.80. The level of inflation/deflation in genomic models was small and trait-dependent. Compared to the regular ssGBLUP, ssGBLUP_MF and ssGBLUP_UPGA helped the regression coefficients to approach one. In general, the genomic models with metafounders seemed to be slightly more stable than models with UPG. This is because the results were more similar with different number of groups. Additionally, models with metafounders showed better genetic trends than ssGBLUP_UPG or

ssGBLUP_UPGA. Metafounders can help to reduce bias in genomic evaluations of composite beef cattle populations without reducing the stability of GEBVs and providing adequate genetic trends.

Keywords: Genomic selection, inflation, Montana cattle, single-step genomic BLUP, unknown-parent groups

Introduction

Single-step genomic BLUP (ssGBLUP) has been widely used for genetic evaluation in several domestic species, as dairy and beef cattle, swine and chicken (Chen et al., 2011;Tsuruta et al., 2013;Lourenco et al., 2015b;Song et al., 2017). The main advantage of this method is the combination of genotyped and non-genotyped animals in the same analysis, possibly providing less biased and more accurate predictions than multistep methods (Chen et al., 2011;Legarra et al., 2014). However, the realized relationship matrix ($\mathbf{H}$) used in ssGBLUP was developed under assumptions that may not hold in practice and may result in biased genomic EBV (GEBV), especially when pedigree information is missing for genotyped animals (Misztal et al., 2013). In such a case, incompatibilities between the genomic ($\mathbf{G}$) and the pedigree ($\mathbf{A}$) relationship are observed (Misztal et al., 2013).

Incompatibilities are also related to the differences in the assumptions of base population for each source of information. While the base population for $\mathbf{A}$ is assumed to be the founders of the pedigree, the base for $\mathbf{G}$ is frequently the current genotyped population, given that $\mathbf{G}$ is most often centered by current allele frequencies (Vitezica et al., 2011). Several approaches have been proposed to solve the incompatibility between $\mathbf{G}$ and $\mathbf{A}$ in ssGBLUP, namely truncation of

pedigree to the most recent generations (Lourenco et al., 2014), scaling parameters for $\mathbf{G}$ and $\mathbf{A}$ (Aguilar et al., 2010), and different ways to construct $\mathbf{G}$ (Chen et al., 2011;Simeone et al., 2012).

The incompatibility between $\mathbf{G}$ and $\mathbf{A}$ may be intensified in crossbred or multibreed populations because the allele frequencies used to center and scale $\mathbf{G}$ are usually based on the average across genotyped animals in the population (Lourenco et al., 2016). An additional problem arises in composite breeds that are formed based on two or more component breeds, and sometimes their combinations, causing the base population to be heterogeneous. Correctly modeling the differences in the base population can aid for less biased genomic predictions, which consequently helps to take the right selection decisions in such populations (Macedo et al., 2020).

In practice, pedigrees used in genetic evaluations may trace back to several base populations that are assumed to be unrelated because this information is beyond pedigree recording. However, those base animals may be related in $\mathbf{G}$ because of its identity-by-state nature. If this is the case, $\mathbf{G}$ and $\mathbf{A}$ will have unbalanced information causing GEBV to be biased (Legarra et al., 2015). Additional to the missing information at the “beginning” of the pedigree, animals in different generation may have missing pedigree information (Tsuruta et al., 2019). If not correctly modeled, founders and animals with missing pedigree will have their breeding values regressed toward zero, which is not realistic given populations are under selection (Legarra et al., 2015). Quaas (1988) and Westell et al. (1988) proposed the use of unknown parent groups (UPG) to overcome the problems related to missing pedigree information. The UPG allows modeling differences in genetic merit across classes of missing parents as year of birth,

country, sex, breed, and so on. Additionally, UPG may be used to account for differences among breeds (Legarra et al., 2007). However, the UPG approach still assumes that the base populations are unrelated, which is not often true.

To solve this issue, Legarra et al. (2015) recently proposed the concept of metafounders, which are pseudo-individuals that act as proxies for the animals in the base populations. When metafounders are used, base individuals can be related though the pedigree, improving the compatibility between $\mathbf{G}$ and $\mathbf{A}$. Metafounders can be also understood as a generalization of UPG or genetic groups, but related within and across base populations (Legarra et al., 2015). According to Garcia-Baccino et al. (2017), the inclusion of metafounders in the model reduces the bias of genomic predictions without loss of accuracy. The assumption that pedigree founders are fully unrelated is voided in the metafounders approach. There are few studies evaluating the performance of ssGBLUP with metafounders in real crossbred populations (Xiang et al., 2017), but none in composite populations.

Montana is a beef cattle composite breed in which the base population is mainly composed of four different biological types, defined as the NABC system. The biological type N is based on animals of some *Bos taurus indicus* breeds; A by animals of some *Bos taurus taurus* breeds adapted to the tropics; B by British *Bos taurus taurus*, and group C by taurine of continental Europe (Ferraz et al., 2002). The current genomic evaluation system for Montana does not account for a heterogeneous base population, but fits breed proportion as covariates in the model. The objective of this study was to evaluate the use of metafounders and unknown parent groups to model the base population in genomic evaluations of a Montana composite population.

Material and Methods

Animal Care and Use Committee approval was not obtained for this study because data set was provided by existing database.

Phenotypic and genomic data

Data from Montana Composto Tropical® - *CFM Leachman Pecuária Ltda*. breeding program were available for this study that included phenotypes on scrotal circumference and growth traits. Pedigree information was available for 680.551 pure-breed, intermediate crossbred, and composite (Montana) animals. A total of 4,212 sires and 192,619 dams were the founders of this breed according to the Montana breed association. Traits included scrotal circumference at 12 months of age (SC12, cm), postweaning weight gain (PWG, Kg; calculated as the difference between weight at 420 days of age and weaning weight), weaning weight (WW, kg; body weight adjusted to 205 days of age), and birth weight (BW, Kg). Phenotypic records deviating from the mean of contemporary groups $\pm$ 3 standard deviations and contemporary groups (GC) with less than five records were removed. The GC were formed by farm, year and season of birth, sex, and management group. After data editing, a total of 49,541 phenotypic records for SC12, 96,994 for PWG, 325,014 for WW, and 264,981 for BW were available.

A total of 1899 animals were genotyped for 30k, 35k, and 770k SNPs. Subsequently, all the genotypes were imputed to the Neogem *GeneSeek*® *Genomic Profiler* (GGP) commercial panel with around 30k SNPs markers using the FImpute software (Sargolzaei et al., 2014). After removing SNP with minor

allele frequency lower than 5%, call rate lower than 90%, departures from Hardy-Weinberg equilibrium (difference between expected and observed frequency of heterozygous) greater than 0.15, and with unknown position or located on sex chromosomes, genotypes on 27,373 SNPs were retained for 1797 animals born from 1999 to 2016.

Statistical Analysis

Both UPG and metafounders (MF) were used to model the heterogeneous base population in Montana. A four-trait model was applied under pedigree BLUP (BLUP), pedigree BLUP with UPG (BLUP_UPG), pedigree BLUP with metafounders (BLUP_MF), single-step GBLUP (ssGBLUP), single-step GBLUP with UPG for $\mathbf{A}$, $\mathbf{G}$, and the pedigree relationship matrix for genotyped animals ($\mathbf{A}_{22}$) (ssGBLUP_UPG) or only in $\mathbf{A}$ and $\mathbf{A}_{22}$ (ssGBLUP_UPGA), and single-step GBLUP with metafounders (ssGBLUP_MF). The model without UPG was fitted as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{c} + \mathbf{e},$$

in which $\mathbf{y}$ is the vector of phenotypes for each trait; $\mathbf{b}$ is the vector of fixed effects, of CG, age of dam in classes for WW and BW, and management group at weaning for SC12 and PWG, and covariables of biological type composition, non-additive effects of total maternal and direct heterozygosis, and age at recording for SC12; $\mathbf{u}$ is the vector of direct additive genetic effects; $\mathbf{m}$ is the vector of maternal additive genetic effects (only for WW and BW); $\mathbf{c}$ is the vector

of maternal permanent environmental effects (only for WW and BW); and $\mathbf{e}$ is the vector of residuals; $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ are the incidence matrices for the effects in $\mathbf{b}$, $\mathbf{u}$, $\mathbf{m}$, and $\mathbf{c}$, respectively.

When UPG were added to pedigree-based BLUP and ssGBLUP evaluations, the $\mathbf{Z}_1\mathbf{Q}\mathbf{g}$ term was added to the model as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{c} + \mathbf{Z}_1\mathbf{Q}\mathbf{g} + \mathbf{e},$$

where $\mathbf{Q}$ is an incidence matrix relating animals in vector $\mathbf{u}$ to unknown parent groups in vector $\mathbf{g}$. Traditional EBV and Genomic EBV (GEBV) for UPG models were calculated as:

$$(\mathrm{G})\mathrm{EBV} = \mathbf{Q}\mathbf{g} + \mathbf{u},$$

The UPGs were modeled in two different ways in ssGBLUP. Firstly, UPG were applied to all the relationship matrices (genomic and pedigree-based) that compose $\mathbf{H}$ (Misztal et al., 2013); this model will be hereinafter defined as ssGBLUP_UPG. In this case, $\mathbf{H}^{-1}$ is represented by $\mathbf{H}^{*}_{UPG}$ and was constructed as:

$$\mathbf{H}^{*}_{UPG} = \mathbf{A}^{*} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \\ \mathbf{0} & -\mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q} \end{bmatrix}$$

Where $\mathbf{A}^{*}$ is the inverse of the pedigree relationship matrix with unknown parent groups, i.e., modified with the QP transformation (Quaas, 1988), and $\mathbf{G}$ was

constructed as in VanRaden (2008) with allele frequencies based on the current genotyped population. It is important to point out that relationships in $\mathbf{G}$ are identical by state, and therefore $\mathbf{G}$ is not affected by missing pedigrees (Tsuruta et al., 2019). Because of that, a second formulation used UPG only in $\mathbf{A}$ and $\mathbf{A}_{22}$; this model will be hereinafter defined as ssGBLUP_UPGA. The $\mathbf{H}^{-1}$ is represented by $\mathbf{H}^{*}_{UPGA}$ and was constructed as follows:

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

A third approach used to model the heterogeneous base population in Montana was metafounders (Legarra et al., 2015). In the MF approach, $\mathbf{A}$ is changed to be compatible to a $\mathbf{G}$ centered by 0.5 allele frequencies ($\mathbf{G}_{05}$). The $\mathbf{H}^{-1}$ with MF is represented by $\mathbf{H}^{\Gamma-1}$:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{-1} - \mathbf{A}_{22}^{\Gamma-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

Where $\mathbf{A}^{\Gamma-1}$ and $\mathbf{A}_{22}^{\Gamma-1}$ are $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ modified with $\boldsymbol{\Gamma}$, a matrix with relationships among metafounders. This matrix can be understood as a function of similarity across base populations (Legarra et al., 2015) and was computed using gammaf90 from the blupf90 software suite (Misztal et al., 2014). All the

computations in this study were done with programs from the same software suite.

The MF and UPG were assigned based on the biological type (N, A, B and C) of the animals. The Montana composite beef cattle is formed by clusters defined by the biological types according to the likeness, physiology, production and reproduction levels, with a combination of both *Bos taurus indicus* and *Bos taurus taurus* animals. The base population is mainly composed of four different biological types defined as the NABC system. The biological type N is composed of animals of some *Bos taurus indicus* breeds; A by animals of some *Bos taurus taurus* breeds adapted to the tropics; B by British *Bos taurus taurus*, and group C by taurine of continental Europe (Ferraz et al., 2002). Intermediate crosses of the base animals can also be used in different proportions to generate the composites. Therefore, two different group definitions were used for both MF and UPG. First, only four groups (GG4) were assigned that traced animals back to their N, A, B, and C founders. In this way, for animals with a missing parent and higher proportion of biological type N, the missing parent was assumed to be from the group N. The same approach assigned animals to other groups.

In the second group definition, 10 groups (GG10) were assigned that traced animals back to the main biological type and their two-way, three-way, and four-way crosses. Initially, the groups were represented by N, A, B, C, NA, AB, AC, BC, NB, NC, NAB, ABC, NAC, NBC, and NABC. As the number of genotyped animals is small in this population, some of these groups had few animals and were, therefore, merged to avoid problems when estimating group effects. Groups NA, AB, AC, and BC were merged into a single group, as well as NAB, NAC, and ABC. The breed proportion of the animal and its known parent were

taken into account for the group assignment. Thus, for an animal with a biological type NA and a known parent of biological type A, the missing parent was assigned to group N. When both parents were unknown, just the breed proportion of animal was taken into account for the group assignment. In this way, for an animal with a biological type NA one parent was assumed to be from the group N and the other from the group A, and so on. The number of animals in each group for the two group definitions is in Table 1.

Table 1. Number of animals in each group

Group definition	GG4[1]	GG10[2]
N	367,737	257,939
A	62,753	7,214
B	43,568	31,572
C	7,058	17,370
NA, AB, AC, BC	-	8,588
NB	-	25,583
NC	-	20,590
NAB, NAC, ABC	-	39,319
NBC	-	22,316
NABC	-	59,625

[1]Group four is based on the main biological types (N, A, B, and C);
[2]Group ten is based on the main biological types (N, A, B, and C) and the intermediate crosses: (NA, AB, AC, BC, NB, NC, NAB, NAC, ABC, NBC, and NABC; where NA, AB, AC, BC were combined into one group, as well as NAB, NAC, ABC).

Evaluation of model performance

The LR validation method (Legarra and Reverter, 2018) was used to evaluate model performance. The validation group was composed by 436 genotyped animals born from June to December of 2016 that had their phenotypes removed from the evaluation, as well as phenotypes of their contemporaries. This will be referred to as the reduced data, and will be represented by the subscript *r*. The total number of records in the reduced data was 49,105, 96,558, 324,578, and 264,546 for SC12, PWG, WW, and BW, respectively. The complete data will be represented by the subscript *c* and was composed by 1797 genotyped animals and, 49,541 phenotypes for SC12, 96,994 for PWG, 325,014 for WW, and 264,981 for BW. This data was used as a benchmark for validations. All the evaluation models were run with both reduced and complete datasets, and all the computations were done with programs from the blupf90 software suite (Misztal et al., 2014).

The estimators of the LR method were calculated based on (Legarra and Reverter, 2018) and Macedo et al. (2020). Accuracy of (G)EBV for validation animals was calculated as $\rho_{c,r} = \sqrt{\frac{cov(\hat{\boldsymbol{u}}_c, \hat{\boldsymbol{u}}_r)}{(1-\bar{F}).\sigma_u^2}}$, where cov is the sample covariance, $\hat{\boldsymbol{u}}$ is the vector of (G)EBV, $\bar{F}$ is the average inbreeding coefficient for validation animals and σ_u^2 is the genetic additive variance. The correlation between $\hat{\boldsymbol{u}}_c$ and $\hat{\boldsymbol{u}}_r$, *i.e.*, $cor_{c,r} = cor(\hat{\boldsymbol{u}}_c, \hat{\boldsymbol{u}}_r)$, which estimates the ratio between the accuracies obtained with complete and reduced datasets was used as a measure of consistency between subsequent evaluations. If $cor_{c,r}$ is high, it means that the changes in breeding values by adding more data are less, and the reduced data is a good predictor of the complete data. Dispersion of (G)EBV was assessed as

the deviation of the regression coefficient (b_1) from 1, where b_1 was obtained from the regression of $\widehat{\boldsymbol{u}}_c$ on $\widehat{\boldsymbol{u}}_r$:

$$\widehat{\boldsymbol{u}}_c = b_0 + b_1\widehat{\boldsymbol{u}}_r$$

Another estimator used to measure the model performance was the bias. This was calculated as $\mu_{cr} = \overline{\hat{u}_r} - \overline{\hat{u}_c}$, where μ_{cr} has an expected value of zero if the evaluation is unbiased.

Results

Relationship within and across MF ($\boldsymbol{\Gamma}$)

The relationships within MF (diagonal of $\boldsymbol{\Gamma}$ matrix) were smaller than one and between MF (off-diagonal of $\boldsymbol{\Gamma}$ matrix) were different than zero in both scenarios (GG4 and GG10). Relationships within MF in GG4 are presented in $\boldsymbol{\Gamma}_4$ below and ranged from 0.15 to 0.38, whereas relationship across MF ranged from 0.09 to 0.18.

$$\boldsymbol{\Gamma}_4 = \begin{bmatrix} 0.19 & 0.11 & 0.09 & 0.09 \\ & 0.15 & 0.13 & 0.13 \\ \vdots & & 0.24 & 0.18 \\ & \ldots & & 0.38 \end{bmatrix}.$$

Overall, relationships within MF in GG10 were greater than in GG4 and ranged from 0.15 to 0.65. However, relationships between MF one and three (biological types N and B) and between one and four (biological types N and C)

showed negative values in this scenario. Additionally, relationships between MF one and two (biological types N and A) was close to zero. The relationship across-metafounders ranged from -0.11 to 0.23 for GG10 (see $\boldsymbol{\Gamma}_{10}$ below).

$$\boldsymbol{\Gamma}_{10} = \begin{bmatrix} 0.59 & 0.02 & -0.11 & -0.08 & 0.02 & 0.09 & 0.18 & 0.08 & 0.07 & 0.10 \\ & 0.21 & 0.15 & 0.12 & 0.15 & 0.10 & 0.09 & 0.14 & 0.13 & 0.13 \\ & & 0.39 & 0.23 & 0.20 & 0.20 & 0.11 & 0.15 & 0.21 & 0.13 \\ & & & 0.65 & 0.16 & 0.15 & 0.07 & 0.13 & 0.17 & 0.13 \\ & & & & 0.48 & 0.16 & 0.13 & 0.14 & 0.16 & 0.13 \\ & & & & & 0.48 & 0.15 & 0.15 & 0.20 & 0.12 \\ & & & & & & 0.57 & 0.12 & 0.16 & 0.12 \\ & & & & & & & 0.19 & 0.14 & 0.13 \\ \vdots & & & & & & & & 0.35 & 0.13 \\ & & \ddots & & & & & & & 0.15 \end{bmatrix}.$$

Accuracy and stability of (G)EBV

Accuracies of (G)EBV for the 436 validation animals are in Table 2. Accuracies were very similar across all pedigree-based models and across genomic models for all traits. However, for genomic models and traits of SC12 and PWG, adding 10 UPG to $\mathbf{A}$ and $\mathbf{A}_{22}$ in ssGBLUP (ssGBLUP_UPGA10) showed greater changes on accuracy when compared to ssGBLUP. The accuracy of (G)EBVs ranged from 0.38 to 0.62 for SC12; 0.27 to 0.45 for PWG; 0.39 to 0.51 for WW and 0.46 to 0.58 for BW. Higher accuracies were observed when genomic information was added to the model, but PWG that did not have considerable changes. Conversely, the inclusion of MF, either four or 10 into ssGBLUP models, brought accuracies of those models downwards (0.03 to 0.07), but it is still bigger than pedigree-based.

Table 2. Accuracy of (G)EBV and correlation between (G)EBV using complete data and (G)EBV using reduced data for validation animals when (G)EBV were estimated using different models with and without genetic groups.

	Accuracy[1]				Correlation[2]			
Model	SC12	PWG	WW	BW	SC12	PWG	WW	BW
BLUP	0.40	0.31	0.39	0.46	0.70	0.80	0.61	0.77
BLUP_UPG4	0.39	0.30	0.41	0.49	0.71	0.80	0.62	0.77
BLUP_UPG10	0.38	0.27	0.40	0.48	0.71	0.77	0.63	0.77
BLUP_MF4	0.40	0.30	0.40	0.46	0.71	0.80	0.64	0.80
BLUP_MF10	0.40	0.28	0.40	0.46	0.72	0.79	0.66	0.81
ssGBLUP	0.47	0.29	0.48	0.57	0.79	0.78	0.75	0.87
ssGBLUP_UPG4	0.48	0.29	0.48	0.57	0.80	0.78	0.76	0.87
ssGBLUP_UPG10	0.47	0.29	0.47	0.58	0.79	0.78	0.75	0.87
ssGBLUP_UPGA4	0.50	0.29	0.51	0.56	0.79	0.79	0.78	0.86
ssGBLUP_UPGA10	0.62	0.45	0.49	0.58	0.83	0.89	0.75	0.88
ssGBLUP_MF4	0.44	0.31	0.43	0.51	0.80	0.84	0.73	0.86
ssGBLUP_MF10	0.43	0.29	0.41	0.50	0.80	0.84	0.74	0.87

[1]Accuracy of (G)EBV for validation animals was calculated as $\rho_{c,r} = \sqrt{\frac{cov(\hat{\boldsymbol{u}}_c,\hat{\boldsymbol{u}}_r)}{(1-\bar{F}).\sigma_u^2}}$, where cov is the sample covariance, $\hat{\boldsymbol{u}}$ is the vector of (G)EBV, *c* and *r* are subscripts to denote complete and reduced datasets, $\bar{F}$ is the average inbreeding coefficient for validation animals and σ_u^2 is the genetic additive variance;
[2]Correlation between $\hat{\boldsymbol{u}}_c$ and $\hat{\boldsymbol{u}}_r$.

Stability or correlation between (G)EBV in two subsequent evaluations ($\hat{\boldsymbol{u}}_r$ and $\hat{\boldsymbol{u}}_c$) are also in Table 2. Correlations ranged from 0.70 to 0.83 for SC12; 0.77 to 0.89 for PWG; 0.61 to 0.78 for WW and 0.77 to 0.88 for BW. Adding genomic information to the model helped to increase the stability of GEBV compared to EBV as correlations increased by at least 0.09, except for PWG (slight changes from pedigree-based models). Under genomics models, there was no overall advantage of fitting either GG4 or GG10 in ssGBLUP_UPG, but an increase of 0.0 to 0.10 with ssGBLUP_UPGA10 when compared with ssGBLUP. Additionally, fitting 10 MF to the genomic models helped increase the stability of GEBV for

PWG trait which had lower stability of GEBV than EBV with ssGBLUP default model. Therefore, fitting genetic groups had a small impact on stability of GEBV in this population.

Slope or Dispersion

The slope (b_1) of the regression of (G)EBVs with complete data on (G)EBVs with reduced data measures the degree of dispersion of (G)EBV estimated under a given model. The b_1 is in Table 3. This coefficient should be close to one to assure that there is no inflation or deflation in (G)EBV for validation animals. The regression coefficients ranged from 0.93 to 1.11 for SC12, 0.93 to 0.99 for PWG, 0.77 to 1.08 for WW, and 0.99 to 1.08 for BW. Overall, pedigree-based methods had no inflation (slight differences from 1.0) except for WW; however, the inclusion of genomic information considerably reduced inflation for WW. Most of the genomic models showed a slight deflation (b1 greater than 1), except for PWG.

Table 3. Regression coefficient of (G)EBV using complete data on (G)EBV using reduced data for validation animals when (G)EBV were estimated using different models with and without genetic groups.

	SC12	PWG	WW	BW
BLUP	0.96	0.95	0.77	0.99
BLUP_UPG4	1.01	0.96	0.77	0.99
BLUP_UPG10	1.00	0.95	0.77	0.99
BLUP_MF4	1.00	0.95	0.80	1.01

BLUP_MF10	0.98	0.95	0.79	1.02
ssGBLUP	1.09	0.93	1.03	1.07
ssGBLUP_UPG4	1.11	0.93	1.03	1.07
ssGBLUP_UPG10	1.11	0.94	1.01	1.08
ssGBLUP_UPGA4	1.04	0.96	1.08	1.04
ssGBLUP_UPGA10	0.93	0.99	0.95	1.08
ssGBLUP_MF4	1.08	0.95	0.89	1.06
ssGBLUP_MF10	1.07	0.96	0.91	1.07

For each trait, the differences among pedigree-based models were small, and the approaches had similar dispersion (traditional BLUP, BLUP with UPGs, and BLUP with MF). For genomic models and all traits, adding UPG to $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{A}_{22}$ in ssGBLUP (ssGBLUP_UPG) showed slight changes on inflation (0.01 to 0.02) when compared to ssGBLUP, independently of number of groups. When UPG were removed from $\mathbf{G}$ (ssGBLUP_UPGA), b1 became closer to 1.0 compared with the abovementioned genomic models for all traits, except WW that had a larger deflation with 4 groups but a slight inflation with 10 groups. Inflation of ssGBLUP and ssGBLUP_MF were very similar (differences of 0.01 to 0.03), except for WW which showed more inflation with MF in the model.

Bias

Table 4 shows the bias of (G)EBV that was calculated as the difference between average (G)EBV from reduced and complete datasets. This difference

has an expected value of zero if the evaluation is unbiased. The (G)EBV for all traits evaluated under all methods had a negative bias, indicating the mean (G)EBV for validation animals from the reduced data was lower than from the complete data. Overall, genomic models were more biased when compared to pedigree-based models, but models with metafounders had similar bias with and without genomic information (slight differences). When UPG were added to pedigree-based models, the bias was greater, except for BW. On the other hand, when UPG were added to $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{A}_{22}$ or to $\mathbf{A}$ and $\mathbf{A}_{22}$ slight differences in bias were observed when compared to ssGBLUP for SC12 and BW. For PWG, the model with four UPG in $\mathbf{A}$ and $\mathbf{A}_{22}$ yielded more bias than the ssGBLUP model, whereas for WW the two models with ten UPG yielded more bias than the ssGBLUP model. Bias of genomic models with metafounders was similar to pedigree-based BLUP and about 20% lower than in the other genomic models.

Table 4. Bias of (G)EBV were estimated using different models with and without genetic groups.

	SC12	PWG	WW	BW
BLUP	-0.10	-0.35	-2.09	-0.26
BLUP_UPG4	-0.15	-0.45	-2.09	-0.25
BLUP_UPG10	-0.47	-0.53	-2.44	-0.03
BLUP_MF4	-0.12	-0.34	-2.09	-0.25
BLUP_MF10	-0.13	-0.37	-2.02	-0.24
ssGBLUP	-0.15	-0.57	-2.48	-0.28
ssGBLUP_UPG4	-0.13	-0.44	-2.36	-0.28
ssGBLUP_UPG10	-0.14	-0.54	-2.56	-0.30
ssGBLUP_UPGA4	-0.13	-0.62	-2.16	-0.25
ssGBLUP_UPGA10	-0.14	-0.41	-2.71	-0.25
ssGBLUP_MF4	-0.13	-0.37	-2.11	-0.26
ssGBLUP_MF10	-0.11	-0.36	-2.01	-0.25

Increasing the number of groups had different effect on each trait. Without genomic information, increasing the group numbers from four to ten led to higher bias for all traits but BW. On the other hand, when more UPG were added to all matrices in $\mathbf{H}$, the bias increased for all traits and the magnitude depended on the trait (i.e., more bias was observed for PWG and WW). In the models with UPG in $\mathbf{A}$ and $\mathbf{A}_{22}$, almost no change in bias was observed for SC12 and BW, whereas bias decreased for PWG and increased for WW. A Slight decrease in bias was observed when the number of metafounders increased from four to ten.

Correlation of (G)EBVs among models

The correlations among different models to predict (G)EBV of young animals are presented in Figures 1 to 4. The Pearson correlation coefficient was assessed to measure the degree of similarity of (G)EBV between models. Overall, all correlations among models were positive and high, except for ssGBLUP_UPGA10 models. For SC12, the correlations ranged from 0.52 to 1.0 (Figure 1). Higher correlations were observed within pedigree-based (>0.98) and within genomic models (>0.91) excluding ssGBLUP_UPGA10 that had correlations around 0.70. The correlations between pedigree-based and genomic models were lower (from 0.52 to 0.75).

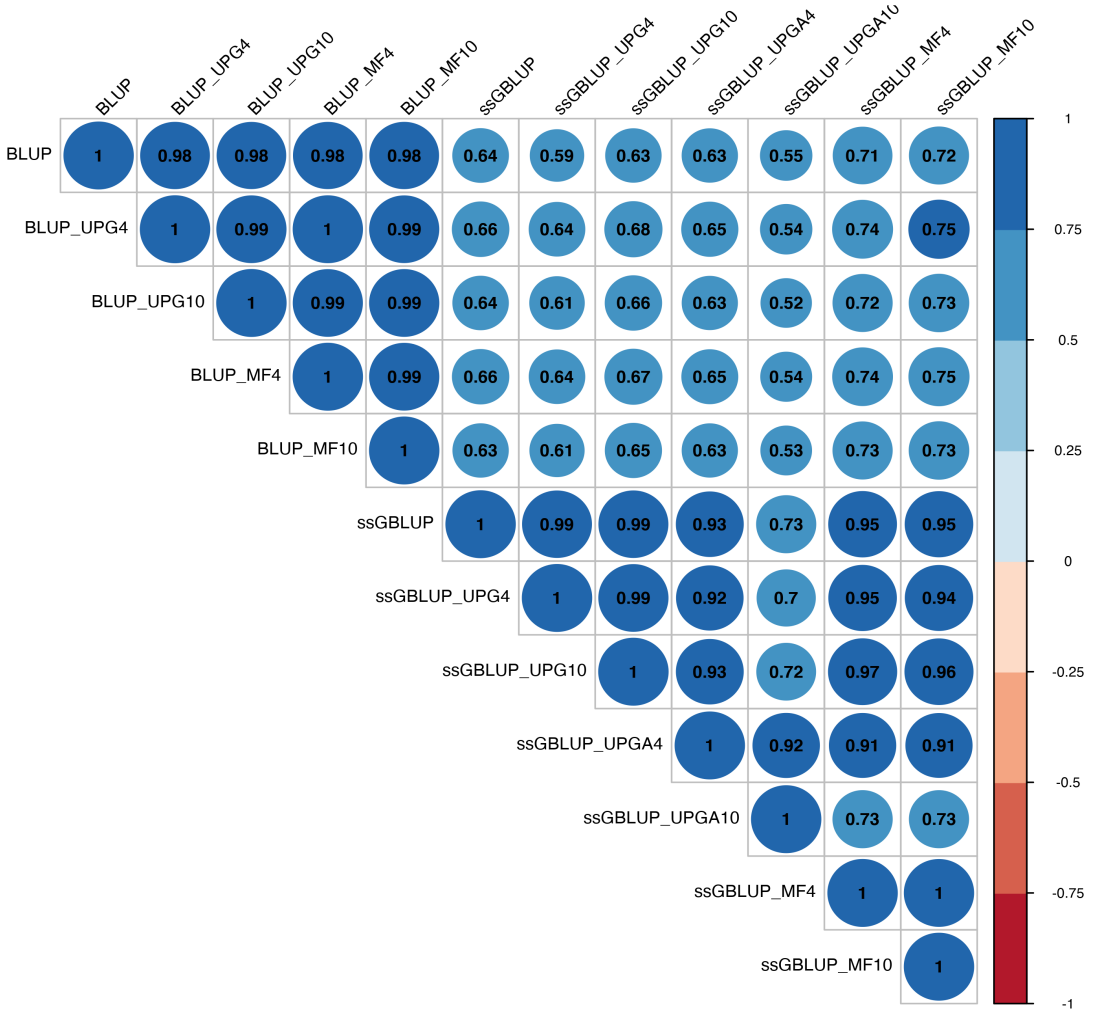

Figure 1. Correlation between breeding values for scrotal circumference at 12 months of age estimated with different models with and without genetic groups.

We observed correlations ranging from 0.12 to 1.0 for PWG (Figure 2). Correlations among pedigree-based models ranged from 0.94 to 1.0, and among genomic models from 0.29 to 1.0. The lowest correlations were between ssGBLUP_UPGA10 and the other models.

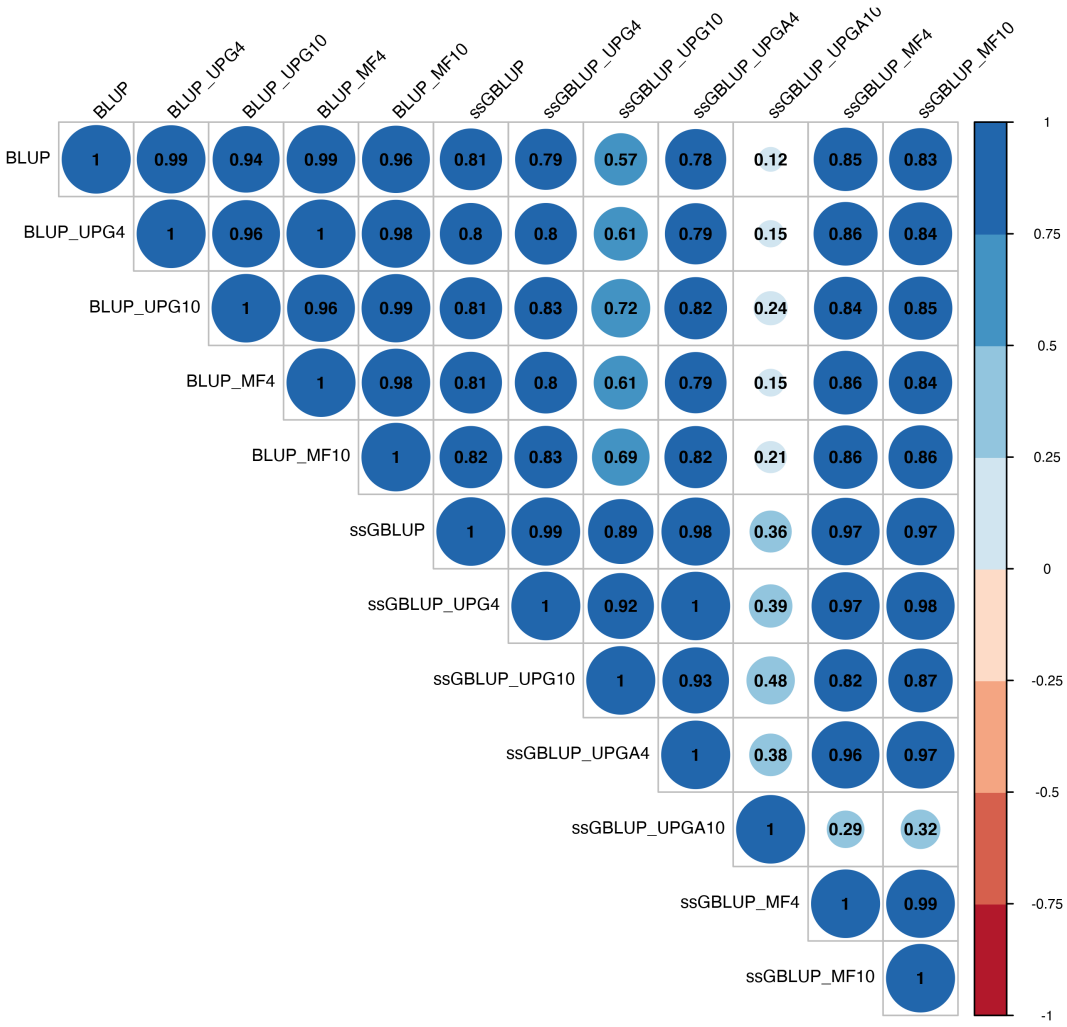

Figure 2. Correlation between breeding values for postweaning weight gain estimated with different models with and without genetic groups.

Correlations among models for WW ranged from 0.40 to 1.0 (Figure 3). The lowest correlations were observed between pedigree-based BLUP and ssGBLUP_UPGA4. High correlations (~0.80) between pedigree-based and genomic models were obtained when metafounders were taken into account in ssGBLUP.

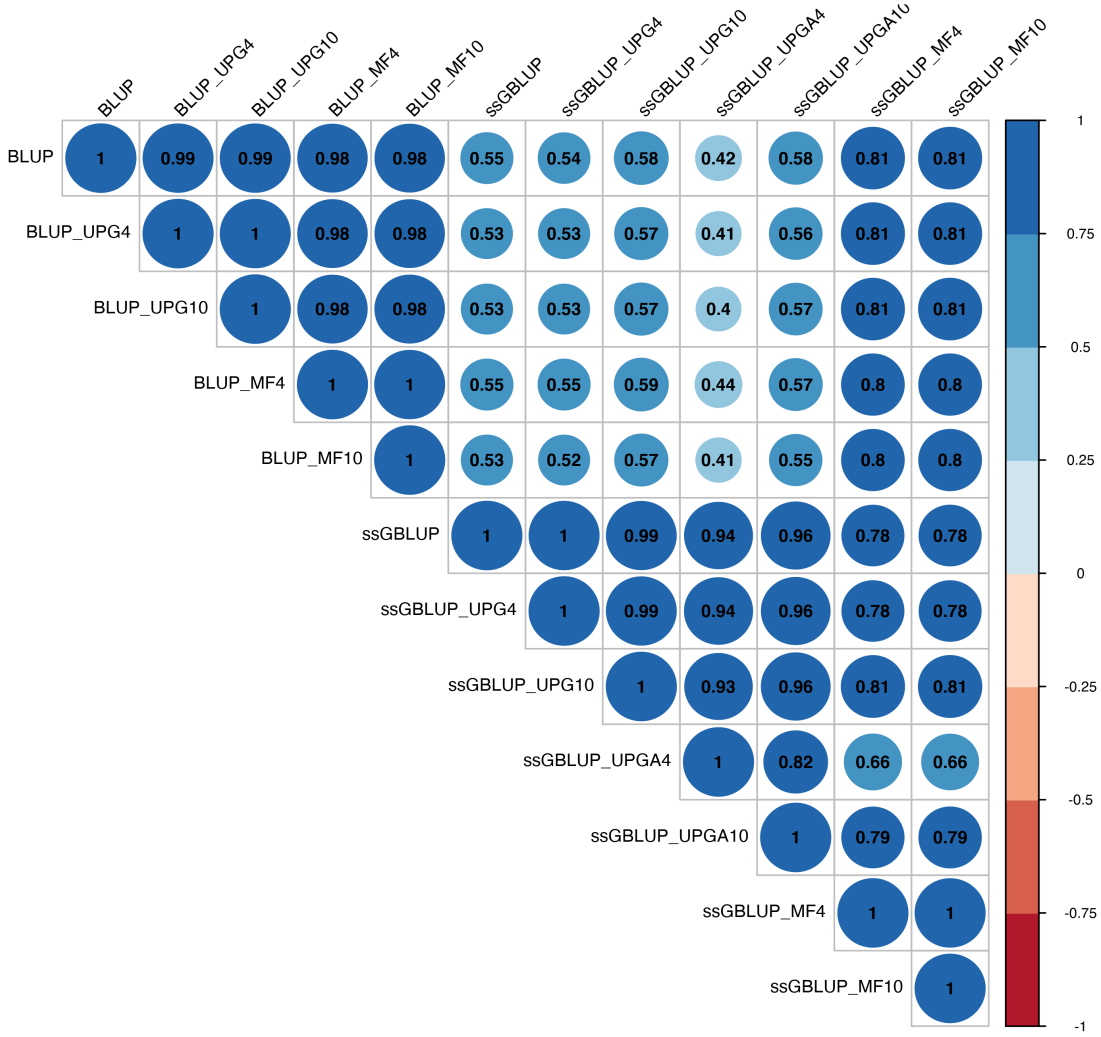

Figure 3. Correlation between breeding values for weaning weight estimated with different models with and without genetic groups.

Finally, the greatest correlations were observed for BW (0.81 to 1.0). The correlations between pedigree-based models were around 1.0 (0.98 to 1.0), between genomic models ranged from 0.95 to 1.0, whereas correlations between pedigree-based and genomic models were around 0.80. The GEBV from ssGBLUP with MF had the greatest overall correlation with all other methods.

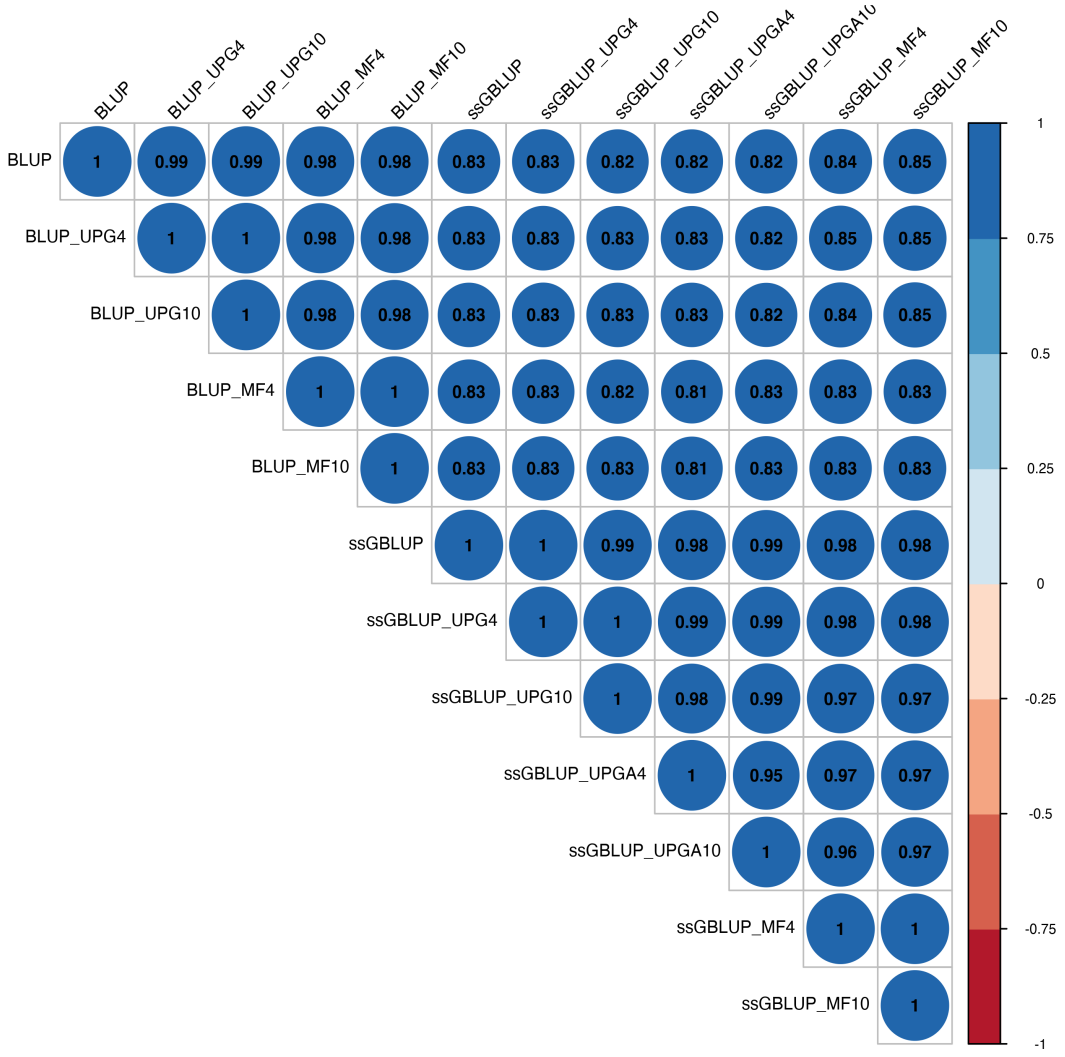

Figure 4. Correlation between breeding values for birth weight estimated with different models with and without genetic groups.

Discussion

To be considered as Montana, the animals should have at least three breeds in their composition with minimum percentages of biological type A of 12.5%; and N and A biological types summed up to 25%. In addition, the maximum percentage of each biological type allowed is 37.5% for group N, 87.5% for group A; and 75% for groups B and C (Santana et al., 2013). These

information shows how heterogeneous can be a Montana animal. In this sense, for this population to model differences between breeds is very important.

Additionally, nowadays the Montana breeding program does not account for missing pedigrees in the genetic evaluation. Does not account for those implies that animals with missing pedigrees have their breeding values regressed toward zero. However, under the selection, animals in different generation must have different breeding values. Missing pedigrees can be accounted by using genetic groups or metafounders. It is a big challenge working with composite beef cattle as Montana due to the heterogeneity of population. Thus, the development of tools able to model these situations as base population, breed differences and missing pedigrees in populations where these factors are present as in Montana beef cattle are very important.

Relationship within and across MF $(\boldsymbol{\Gamma})$

In animal breeding, missing pedigrees and uncertainty regarding the base population may be accounted for by adding genetic groups called unknown-parent groups (UPG), based on some criteria as year of birth, generation, breed, line, sex or a combination of these. Recently, Legarra et al. (2015) proposed the use of MF to account for relationships within and among base populations that are ignored by using UPG. Therefore, the main difference between UPG and MF is that the latter act as related, inbred UPG.

Relationships within MF less than one indicate negative inbreeding, which implies a higher frequency of heterozygotes relative to the average of the population. It means that the base population has a large genetic variability.

Conversely, when these values are higher than one, the base populations were inbred (Legarra et al., 2015). Additionally, positive relationships across MF indicate that ancestor populations overlap, whereas negative values indicate population divergence, and zero means the base populations are unrelated. The estimation of $\boldsymbol{\Gamma}$ different than zero allows for the full consideration of metafounders in genetic evaluation, because $\boldsymbol{\Gamma}$ equal to zero is equivalent to having UPG for $\mathbf{A}$ and $\mathbf{A}_{22}$ (Bradford et al., 2019).

Relationships within and across MF in this study were lower than one and different than zero, respectively. Similar findings were also reported in simulated and real datasets (Xiang et al., 2017;van Grevenhof et al., 2018;Bradford et al., 2019). In contrast with those studies, we found negative relationships across some MF (purebred) when using ten groups. This was between the founder groups N and B, and N and C and can be explained by the fact Montana is composed by two subspecies of Bos Taurus, *i.e., Bos taurus indicus* in the group N and *Bos taurus taurus* in the groups A, B, and C. However, negative coefficients were not observed when using only four groups, possibly because the intermediate crosses were implicitly nested within the four biological types. This created stronger ties among the biological types based on SNP information and, therefore, positive coefficients.

The relationship coefficients within MF in our study were similar to those reported by other authors. In a simulated study, van Grevenhof et al. (2018) found relationship within MFs of 0.17 and 0.74 if parental lines were related or not , respectively. Bradford et al. (2019) showed relationships within MF ranging from 0.54 to 0.71 in a simulated dairy cattle population. Colleau et al. (2017) found relationship within MF equal to 0.47 in sheep, and Legarra et al. (2015) showed

relationship of 0.55 and 0.77 for Holsteins and Jersey breeds, respectively. In general, the relationships within groups in GG10 were higher than in GG4, implying more variability in the latter. This can be justified by the method used to setup the groups in each scenario. In GG4 the animals with a higher proportion of a certain biological type (N, A, B and C) were allocated to a purebred group in a way the groups were not homogeneous (*i.e.*, were composed by purebred and crossbred). Conversely, in GG10 the strategy was to split purebred and crossbreds in different groups that were more homogeneous.

Accuracy and stability of (G)EBV

Adding genomic information increase the accuracy of GEBV for all traits but PWG. The genomic information allows to improve the relationship coefficient among animals and a better estimation of mendelian sampling, which is not possible through pedigree-based relationship matrix ($\mathbf{A}$). Additionally, the marker information allows a better estimation of relationship between animals. For instance, two unrelated individuals in $\mathbf{A}$ will be related through $\mathbf{H}$ if these are related through $\mathbf{G}$ matrix, even if pedigree not shows it (Legarra et al., 2014). When there is no information of phenotypes and progeny (young animals) the GEBV is composed by the sum of parent average, genomic information coming from $\mathbf{G}$ and pedigree information coming from $\mathbf{A}_{22}$ (Lourenco et al., 2015a). On the other hand, for ungenotyped animals the EBV is composed by half of the parents' (G)EBV plus the Mendelian sampling (Quaas, 1988). In this sense, take genomic information into account allows to combine more information to estimate breeding values for young animals without records and progenies. How much this extra amount of information contributes to improve the estimation of breeding

values can be measured through the increasing in accuracy by using models with marker information.

Indeed, increases in the accuracy of predicting futures phenotypes or GEBVs have been reported by using genomic models, supporting the benefits of genomic selection in livestock breeding programs (Aguilar et al., 2010;Chen et al., 2011;Baloche et al., 2014;Garcia et al., 2018). However, increasing in accuracy differ based on a lot of factors and can be different for each trait. In this study, we found lower or no increase in accuracy for PWG trait by using genomic information. These findings can be justified by the low heritability and accuracy of animals' EBV for this trait when compared to the other ones. Additionally, predictive of GEBVs is related to reference population size and composition of this population. These factors should be taken into account to set up genomic selection schemes in livestock production. A good way to choose genotyping strategies aiming to increase the accuracy would be take into account older important animals with higher EBV accuracy and large amounts of young animals (Lourenco et al., 2015b). However, our reference population was limited to few animals with low accuracy of EBV especially for PWG trait which can be understood as a limitation factor in this population.

The accuracy of the models is an important feature to be considered in genetic evaluation because a good model needs to be capable to predict (G)EBVs or future phenotypes in an accurate way to produce genetic gain. When the number of genotyped animals is small, the proportion of information coming from $\mathbf{G}$ to GEBV estimates is also small, and then the increase in accuracy due to genomic information is likely lower (Lourenco et al., 2015b). The accuracy of genomic predictions is strongly dependent on the number of genotyped animals,

number of markers and heritability of the recorded phenotype. Another important thing to be considered is the relationship between reference population and target animals (Calus, 2010). In such way, accuracy of GEBV may be very similar to EBV if all these points are not taken into account (VanRaden et al., 2009). Is important to point out that the abovementioned factors can be more important in composite or crossbred populations since their genome is a mosaic of genome regions inherited from purebred progenitors (Sevillano et al., 2019).

Overall, the inclusion of UPGs in all relationship matrices of ssGBLUP, either four or 10 (ssGBLUP_UPG), was not able to increase the accuracy in our study, just slight differences in these methods were observed. However, ssGBLUP_UPG4 model yielded the greatest accuracy compared to the other genomic models for WW trait. Conversely, ssGBLUP_UPGA10 model yielded higher accuracy than other genomic models for SC12 and PWG. Although the accuracy for those methods were bigger it is likely unreliable because genetic trends were weird (Appendix). Additionally, increasing in the number of UPG is not always the best choice, since UPG solutions are essentially related to the number of animals and phenotypes in each group (Tsuruta et al., 2014). The best way to set UPGs is try to model the differences between groups and keep a good number of phenotypes and animals in each group but it is not always feasible in real data. In composite cattle modeling unknown-parent groups can be even worse because there are differences among animal breed compositions as well.

Accuracy of models was more stable in relation to difference in number of UPGs when those were also added to $\mathbf{G}$ matrix (ssGBLUP_UPG). Removing UPGs from $\mathbf{G}$ matrix seems have more influence of the number of UPGs. However, genetic trends were more reliable when UPGs were kept in $\mathbf{G}$ matrix as

well (results no show). There are some evidences that for purebred populations it is not necessary to include UPGs into $\mathbf{G}$, because this matrix is not affected by missing pedigrees. However, $\mathbf{G}$ matrix can be affected by lines or breed differences (Plieschke et al., 2015), so in composite populations likely a good way is set UPG for $\mathbf{G}$ matrix as well in terms of genetic trends.

The main reason to account UPG in genetic evaluation is that genetic trends could have large bias when genetic difference among groups are ignored and it is worse in populations strongly selected. On the other hand, poor definition and wrong assignments of UPGs can also introduce bias. Therefore, UPGs needs to be estimated accurately with enough information to avoid these issues (Tsuruta et al., 2019). However, when the goal is predict the EBVs of the youngest generation removing UPGs from the model should not have large effect on the accuracy of those mainly if genotyped animals does not have missing parents (Misztal et al., 2013).

Our finding indicates that metafounders may not improve the accuracy when compared to ssGBLUP without those (slight differences) mainly for traits with lower selection. Our hypothesis suggests that MF can works better for traits with more selection as PWG. Similar results were also reported by other authors in previous studies with simulated and real dataset. In a simulation study, Bradford et al. (2019) showed that the accuracy of models is more related to the trait heritability than inclusion or not of metafounders or UPGs in ssGBLUP. The authors showed that traits with higher heritability have higher accuracy even just with pedigree-based models than traits with lower heritabilities. These authors found a slight increase in accuracy in ssGBLUP with metafounders (0.01 to 0.04

to the heritability 0.3 and 0.01 to 0.03 to the trait with heritability 0.1). In contrast, we found lower accuracies with MF, however, these models are still accurate.

In two different simulation studies Garcia-Baccino et al. (2017) and van Grevenhof et al. (2018) found a small extra gain (0.02) when include metafounders in ssGBLUP model and no difference in prediction accuracy, respectively. For a real dataset, Xiang et al. (2017) showed that the inclusion of metafounders in the ssGBLUP performed as well as ssGBLUP with breed of origin of alleles that requires phasing genotypes and can be performed in a simple way. Metafounders were developed aiming make $\mathbf{G}$ and $\mathbf{A}$ matrices compatible. It is important to point out that metafounders are applied to $\mathbf{A}$ matrix. Incompatibility between these matrices is related to different base populations used to building them, however this issue is more related to bias than to accuracy or stability of GEBVs and it can be verified in our study as well.

Overall, the inclusion of genomic information to the models helped increase the correlation in subsequent evaluations, except for PWG trait. As in accuracy, we did not expect increasing in stability by using MF in the models. Slight differences of ssGBLUP_MF from ssGBLUP for SC12, WW and BW traits were observed. Conversely, for PWG the inclusion of MF helped to get higher correlations than pedigree-based models. These results indicate that MF likely can help traits and models in which traditional ssGBLUP does not work very well, in other words, those that produce accuracy or correlation similar to the pedigree-based models.

Slope or Dispersion

When the slope of the regression of (G)EBV with complete data on (G)EBV with reduced data is equal to 1, both sets of (G)EBV are in a similar scale. Otherwise, inflation or deflation will occur when b_1 is less or over 1. It is important to note that the scale of (G)EBV is a key component in selection schemes because it allows the fair comparison of (G)EBV among animals and, consequently, proper selection decisions (Piccoli et al., 2018). Inflation causes overdispersion, which is detrimental for genomic predictions, especially when the selection candidates are from different generations or have different amount of information (Neves et al., 2012).

The inclusion of genomic information in our study helped to reduce the inflation for WW predictions; however, the inclusion of UPG or metafounders had a small impact across traits. Bradford et al. (2019) showed that the dispersion of BLUP models without accounting for missing pedigrees can be greater than in the genomic models. Inflation in genomic models is likely caused by a mismatch between the scale of pedigree and genomic relationship matrices (Misztal et al., 2017). Vitezica et al. (2011) showed that inflation is related to the heritability of the trait and selection pressure. According to Tsuruta et al. (2019), the inflation of GEBV can be reduced by weighting $\mathbf{A}_{22}^{-1}$ by a factor smaller than 1.0 or by reducing the additive genetic variance of the trait. Inflation of GEBV using ssGBLUP may be also observed when the pedigree is long but incomplete and when inbreeding is not considered in $\mathbf{A}$. Consequently, inflation/deflation can be reduced by a combination of pedigree truncation, incorporation of inbreeding in $\mathbf{A}$, and inbreeding for unknown parents (Misztal et al., 2017).

Our results showed small or no change in inflation/deflation of (G)EBV when UPG were added to any BLUP or ssGBLUP models. This indicates UPG

or metafounders were not able to reduce the inflation for traits like WW that had severely inflated EBV. Conversely, the inclusion of genomic information eliminated the inflation for such a trait. Under genomic models, the inclusion of UPG or metafounders helped to alleviate the inflation. Overall, fitting ten groups was slightly better than four groups. It is important to point out that to estimate UPG effects, animals with and without phenotypes sharing the same UPG must be related. In the same way, if animals in $\mathbf{A}^{22}$ (block of $\mathbf{A}^{-1}$ for genotyped animals) and $\mathbf{A}^{11}$ (block of $\mathbf{A}^{-1}$ for non-genotyped animals) are not related (*i.e.,* $\mathbf{A}^{12}$ = 0), $\mathbf{H}^{-1}$ will not contribute to the estimation of UPG effects (Tsuruta et al., 2019). The estimation of metafounders effects relies on $\boldsymbol{\Gamma}$ and, consequently, on the number of genotyped animals with phenotypes connected to each metafounder.

Regarding UPG, a decrease in inflation/deflation was observed in ssGBLUP_UPGA when compared to ssGBLUP_UPG. This raises a question on whether UPG should be fitted in $\mathbf{G}$, given genomic relationships do not rely on pedigree missingness. According to Plieschke et al. (2015), $\mathbf{G}$ accounts for line or breed differences (Plieschke et al., 2015) and, therefore, using UPG in this matrix may be beneficial in crossbred and multibreed populations. Based on our results, it is enough to fit UPG in $\mathbf{A}$ and $\mathbf{A}_{22}$ for obtaining the least inflated/deflated genomic predictions in this Montana composite beef cattle population. However, our study was based on a small number of genotyped animals and research in larger populations is necessary to confirm our findings. Overall, the metafounders model provided similar b_1 as ssGBLUP_UPGA; it is important to recall that in the metafounders approach, $\mathbf{A}$ and $\mathbf{A}_{22}$ are modified to be compatible to $\mathbf{G}$ centered by 0.5 allele frequencies, therefore, there are no extra modifications in $\mathbf{G}$.

Bias

Bias defines the ability to correctly predict the average of breeding values for selection candidates (Granado-Tajada et al., 2020). If this value is different than zero, the ability to correctly estimate genetic trends and genetic gain is compromised (Legarra and Reverter, 2018). A negative value for bias indicates an underestimation of (G)EBV using reduced data. We found negative values for all models and traits in our study, with overall stronger biases for ssGBLUP with and without UPG compared to BLUP models. However, similar values of bias were observed between BLUP and ssGBLUP_MF models for all traits. This result indicates that the inclusion of MF into ssGBLUP can reduce bias of genomic models to the same level as BLUP, although the latter was still biased probably because of model artifacts and pre-selection. A way to eliminate bias in ssGBLUP models in composite populations would likely involve a pedigree-based model with minimum bias and the addition of genomic information together with MF to model the heterogeneous base population. Because artificial selection can generate bias due to an increase in the genetic level and a reduction in genetic variance, finding an unbiased model can be challenging (Legarra and Reverter, 2018).

In a simulation study, Bradford et al. (2019) found no bias in pedigree-based and genomic models with complete pedigree. The same authors reported an increase in bias of pedigree-based models when UPG were used to account for missing pedigree; this agrees with our findings. Bias with the inclusion of UPG is mainly caused by inaccurate estimates of UPG effects, which reinforces the importance of a robust group definition. Tsuruta et al. (2014) showed that

combining groups with less information helped to reduce bias of genomic predictions in the US Holstein population.

Contrasting to our results, Bradford et al. (2019) found a greater bias in pedigree-based models than genomic models when missing pedigree was not taken into account. However, our results are in agreement with Garcia-Baccino et al. (2017) who found greater bias in ssGBLUP predictions compared to BLUP; the latter was in fact unbiased. It is important to note that the models proposed by those authors were simpler than ours that have genetic and non-genetic maternal effects and several fixed effects. The complex structure of the Montana population also plays an important whole in bias.

Correlation of (G)EBVs among models

The correlation of (G)EBV across models shows the similarity of estimated breeding values between models. Ours results indicate similarity between EBV with different models for all traits, meaning that the inclusion of UPG or MF into pedigree-based models was not able to considerably change predictions. Overall, correlations between genomic models were lower than between BLUP models, which indicates more changes. Correlations between BLUP and ssGBLUP_MF were greater than between BLUP and ssGBLUP with or without UPG. It is important to point out that (G)EBV coming from UPG and MF models include group effects. Therefore, changes in (G)EBV for those models depend on the estimation of group effects. Additionally, if animals with and without phenotypes in the same group are not related, the group effects are not estimable and the model is similar to ignoring UPG. Furthermore, if genotyped animals are not related to non-genotyped animals (animals in $\mathbf{A}^{22}$ and $\mathbf{A}^{11}$,

respectively), $\mathbf{H}^{-1}$ will not contribute to the estimation of group solutions and it is again similar to ignoring UPG (Tsuruta et al., 2019). Since UPG solutions are trait-dependent, changes in (G)EBV also are related to the traits.

Conclusions

Genomic information helps to increase accuracy and persistency of predictions in a composite beef cattle population. Using unknown parent groups only in the pedigree or in both pedigree and genomic relationship matrices in ssGBLUP to account for heterogeneous base population and pedigree missingness does not improve accuracy, inflation/deflation, and bias of genomic predictions; therefore, using UPG in this Montana population is not recommended. Although the addition of metafounders to a ssGBLUP model was not able to promote an increase in accuracy and a reduction in inflation/deflation, the models with MF provides the least biased genomic predictions. Thus, ssGBLUP with MF should be the model of choice for evaluations of small composite populations.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T.J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93**,** 743-752.

Baloche, G., Legarra, A., Sallé, G., Larroque, H., Astruc, J.M., Robert-Granié, C., and Barillet, F. (2014). Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. *Journal of Dairy Science* 97**,** 1107-1116.

Bradford, H.L., Masuda, Y., Vanraden, P.M., Legarra, A., and Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. *Journal of Dairy Science* 102**,** 2336-2346.

Calus, M.P.L. (2010). Genomic breeding value prediction: methods and procedures. *Animal* 4**,** 157-164.

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., and Muir, W.M. (2011). Effect of different genomic relationship matrices on accuracy and scale. *Journal of Animal Science* 89**,** 2673-2679.

Colleau, J.-J., Palhière, I., Rodríguez-Ramilo, S.T., and Legarra, A. (2017). A fast indirect method to compute functions of genomic relationships concerning genotyped and ungenotyped individuals, for diversity management. *Genetics Selection Evolution* 49**,** 87.

Ferraz, J., Eler, J.P., Dias, F., and Golden, B. (Year). "(Co)variance component estimation for growth weights of Montana TropicalÆ, a Brazilian beef composite").

Garcia, A.L.S., Bosworth, B., Waldbieser, G., Misztal, I., Tsuruta, S., and Lourenco, D.a.L. (2018). Development of genomic predictions for harvest and carcass weight in channel catfish. *Genetics Selection Evolution* 50**,** 66.

Garcia-Baccino, C.A., Legarra, A., Christensen, O.F., Misztal, I., Pocrnic, I., Vitezica, Z.G., and Cantet, R.J.C. (2017). Metafounders are related to F (st) fixation indices and reduce bias in single-step genomic evaluations. *Genetics, selection, evolution : GSE* 49**,** 34-34.

Granado-Tajada, I., Legarra, A., and Ugarte, E. (2020). Exploring the inclusion of genomic information and metafounders in Latxa dairy sheep genetic evaluations. *Journal of Dairy Science* 103**,** 6346-6353.

Legarra, A., Bertrand, J.K., Strabel, T., Sapp, R.L., Sánchez, J.P., and Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics* 124**,** 286-295.

Legarra, A., Christensen, O.F., Aguilar, I., and Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livestock Science* 166**,** 54-65.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., and Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. *Genetics* 200**,** 455-468.

Legarra, A., and Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution* 50**,** 53.

Lourenco, D.A., Fragomeni, B.O., Tsuruta, S., Aguilar, I., Zumbach, B., Hawken, R.J., Legarra, A., and Misztal, I. (2015a). Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet Sel Evol* 47**,** 56.

Lourenco, D.a.L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T.J., Forni, S., and Weller, J.I. (2014). Are evaluations on young genotyped animals benefiting from the past generations? *Journal of Dairy Science* 97**,** 3930-3942.

Lourenco, D.a.L., Tsuruta, S., Fragomeni, B.O., Chen, C.Y., Herring, W.O., and Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices1. *Journal of Animal Science* 94**,** 909-919.

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., and Misztal, I. (2015b). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus1. *Journal of Animal Science* 93**,** 2653-2662.

Macedo, F.L., Christensen, O.F., Astruc, J.-M., Aguilar, I., Masuda, Y., and Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genetics Selection Evolution* 52**,** 47.

Misztal, I., Bradford, H.L., Lourenco, D.a.L., Tsuruta, S., Masuda, Y., Legarra, A., and Lawlor, T.J. (2017). Studies on inflation of GEBV in single-step GBLUP for type. *Interbull Bulletin* 51**,** 38-42.

Misztal, I., Tsuruta, S., Lourenco, D.a.L., Masuda, Y., Aguilar, I., Legarra, A., and Vitezica, Z.G. (2014). *Manual for BLUPF90 family of programs* [Online]. Available: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf [Accessed].

Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I., and Swan, A.A. (2013). Unknown-parent groups in single-step genomic evaluation. *Journal of Animal Breeding and Genetics* 130**,** 252-258.

Neves, H.H.R., Carvalheiro, R., and Queiroz, S.A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC genetics* 13**,** 100-100.

Piccoli, M., Brito, L., Braccini Neto, J., Brito, F., Cardoso, F., Cobuci, J., Sargolzaei, M., and Schenkel, F. (2018). A comprehensive comparison between single- and two-step GBLUP methods in a simulated beef cattle population. *Canadian Journal of Animal Science* 98.

Plieschke, L., Edel, C., Pimentel, E.C.G., Emmerling, R., Bennewitz, J., and Götz, K.-U. (2015). A simple method to separate base population and segregation effects in genomic relationship matrices. *Genetics Selection Evolution* 47**,** 53.

Quaas, R.L. (1988). Additive Genetic Model with Groups and Relationships. *Journal of Dairy Science* 71**,** 1338-1345.

Santana, M.L., Eler, J.P., Cardoso, F.F., Albuquerque, L.G., and Ferraz, J.B.S. (2013). Phenotypic plasticity of composite beef cattle performance using reaction norms model with unknown covariate. *animal* 7**,** 202-210.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC genomics* 15**,** 478-478.

Sevillano, C.A., Bovenhuis, H., and Calus, M.P.L. (2019). Genomic Evaluation for a Crossbreeding System Implementing Breed-of-Origin for Targeted Markers. *Frontiers in Genetics* 10.

Simeone, R., Misztal, I., Aguilar, I., and Vitezica, Z.G. (2012). Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *Journal of Animal Breeding and Genetics* 129**,** 3-10.

Song, H., Zhang, J., Jiang, Y., Gao, H., Tang, S., Mi, S., Yu, F., Meng, Q., Xiao, W., Zhang, Q., and Ding, X. (2017). Genomic prediction for growth and reproduction traits in pig using an admixed reference population1. *Journal of Animal Science* 95**,** 3415-3424.

Tsuruta, S., Lourenco, D.a.L., Masuda, Y., Misztal, I., and Lawlor, T.J. (2019). Controlling bias in genomic breeding values for young genotyped bulls. *Journal of Dairy Science* 102**,** 9956-9970.

Tsuruta, S., Misztal, I., and Lawlor, T.J. (2013). Short communication: Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *Journal of Dairy Science* 96**,** 3332-3335.

Tsuruta, S., Misztal, I., Lourenco, D.a.L., and Lawlor, T.J. (2014). Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. *Journal of Dairy Science* 97**,** 5814-5821.

Van Grevenhof, E.M., Vandenplas, J., and Calus, M.P.L. (2018). Genomic prediction for crossbred performance using metafounders1. *Journal of Animal Science* 97**,** 548-558.

Vanraden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91**,** 4414-4423.

Vanraden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., and Schenkel, F.S. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* 92**,** 16-24.

Vitezica, Z.G., Aguilar, I., Misztal, I., and Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research* 93**,** 357-366.

Westell, R.A., Quaas, R.L., and Van Vleck, L.D. (1988). Genetic Groups in an Animal Model. *Journal of Dairy Science* 71**,** 1310-1318.

Xiang, T., Christensen, O.F., and Legarra, A. (2017). Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. *Journal of Animal Science* 95**,** 1472-1480.

APPENDIX

Figure 1. Genetic trends in standard deviation units of estimated breeding values (EBV) for scrotal circumference at 12 months of age, postweaning weight gain, weaning weight and birth weight.

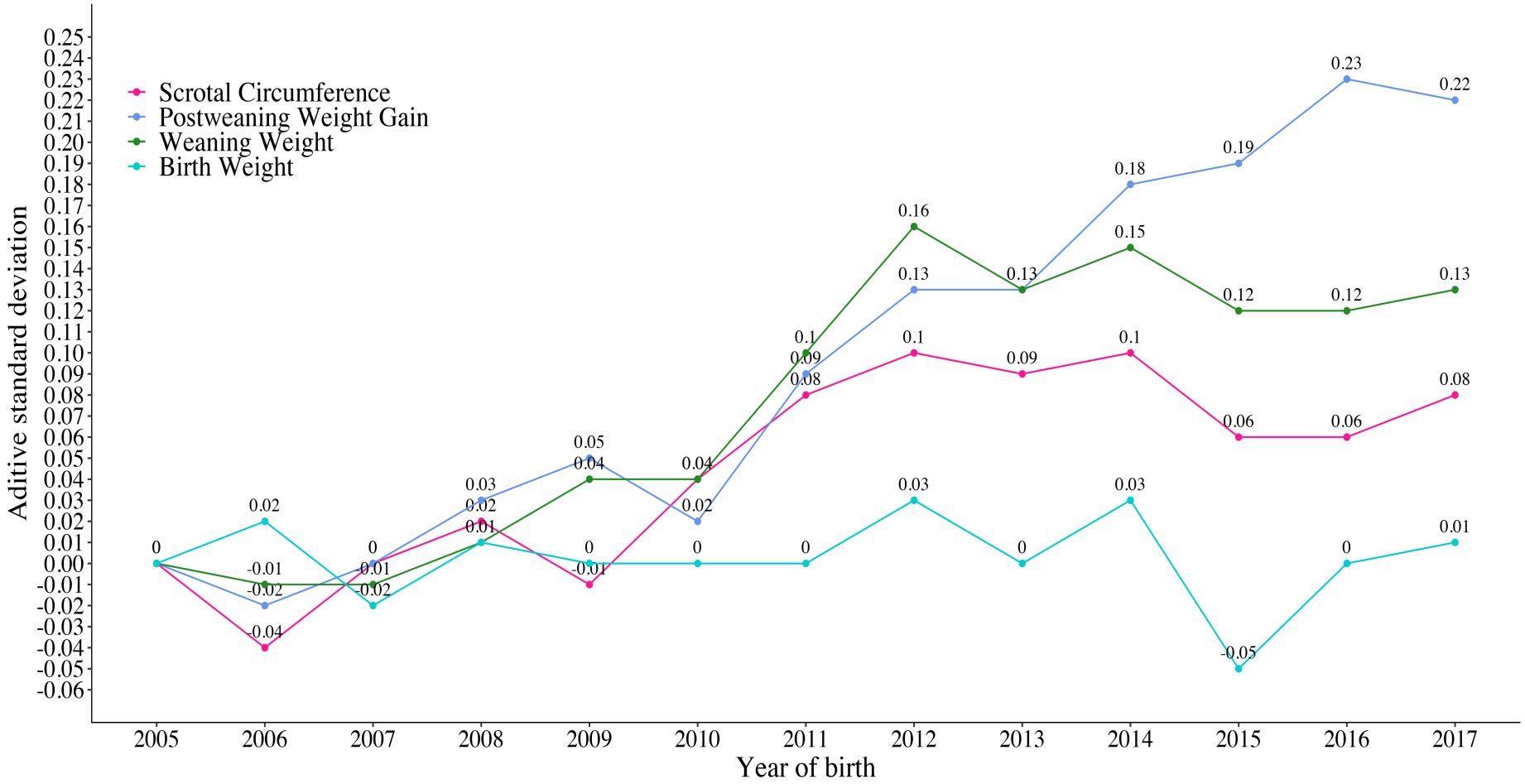

Figure 2. Genetic trends in standard deviation units of genomic estimated breeding values (GEBV) for scrotal circumference at 12 months of age, postweaning weight gain, weaning weight and birth weight.

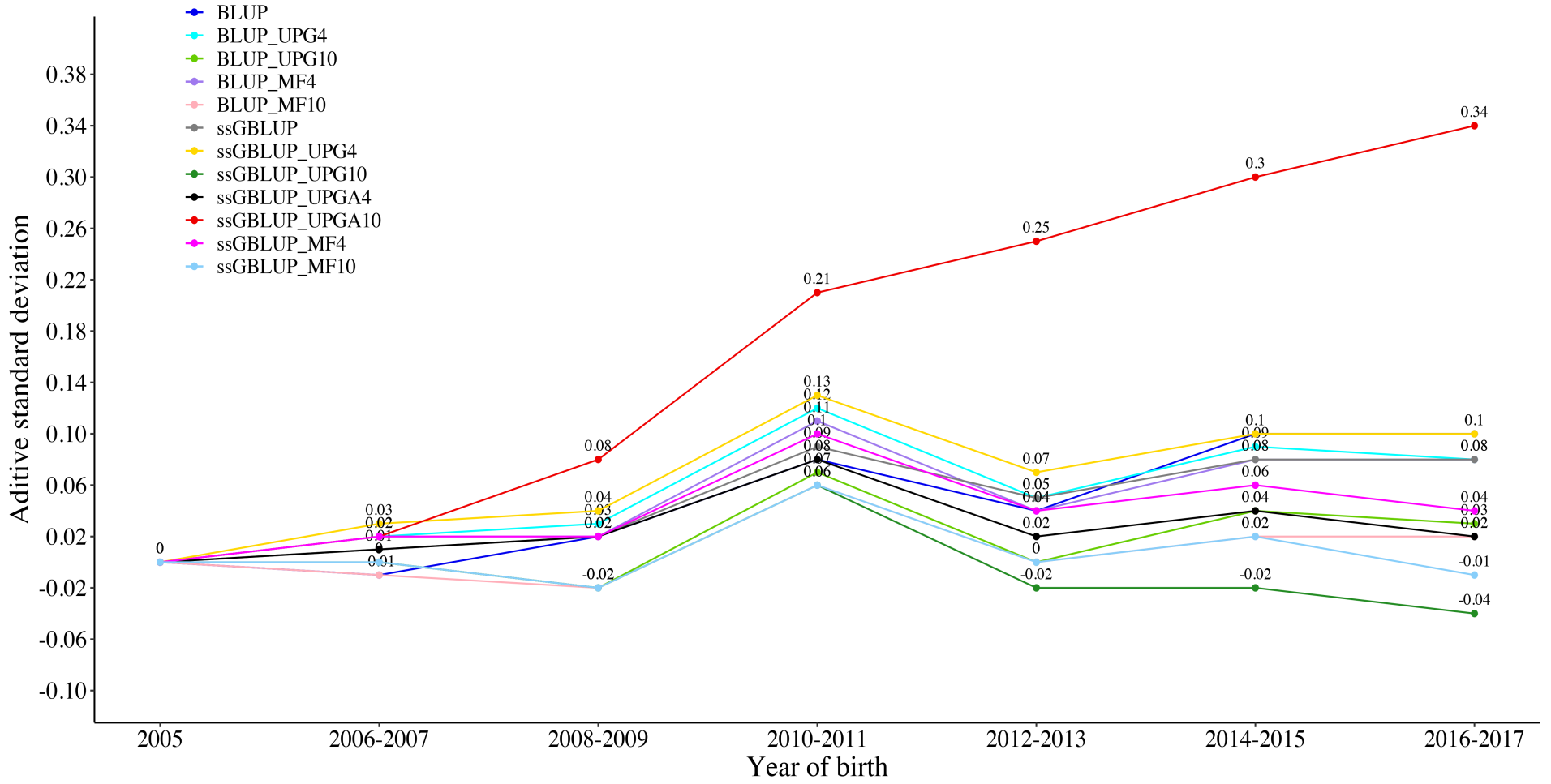

Figure 3. Genetic trends in standard deviation units of genomic estimated breeding values (GEBV) for scrotal circumference at 12 months of age.

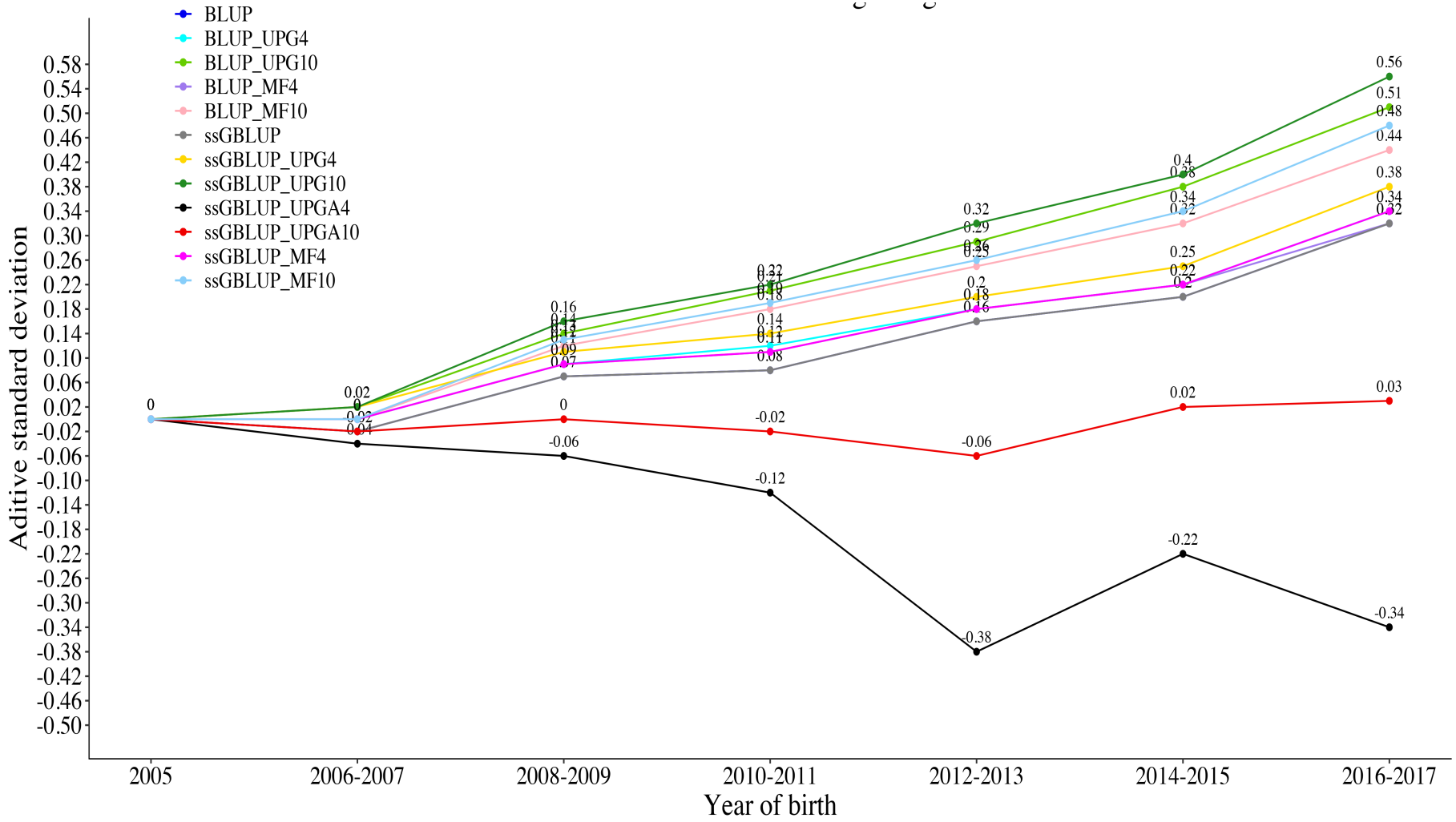

Figure 4. Genetic trends in standard deviation units of genomic estimated breeding values (GEBV) for postweaning weight gain.

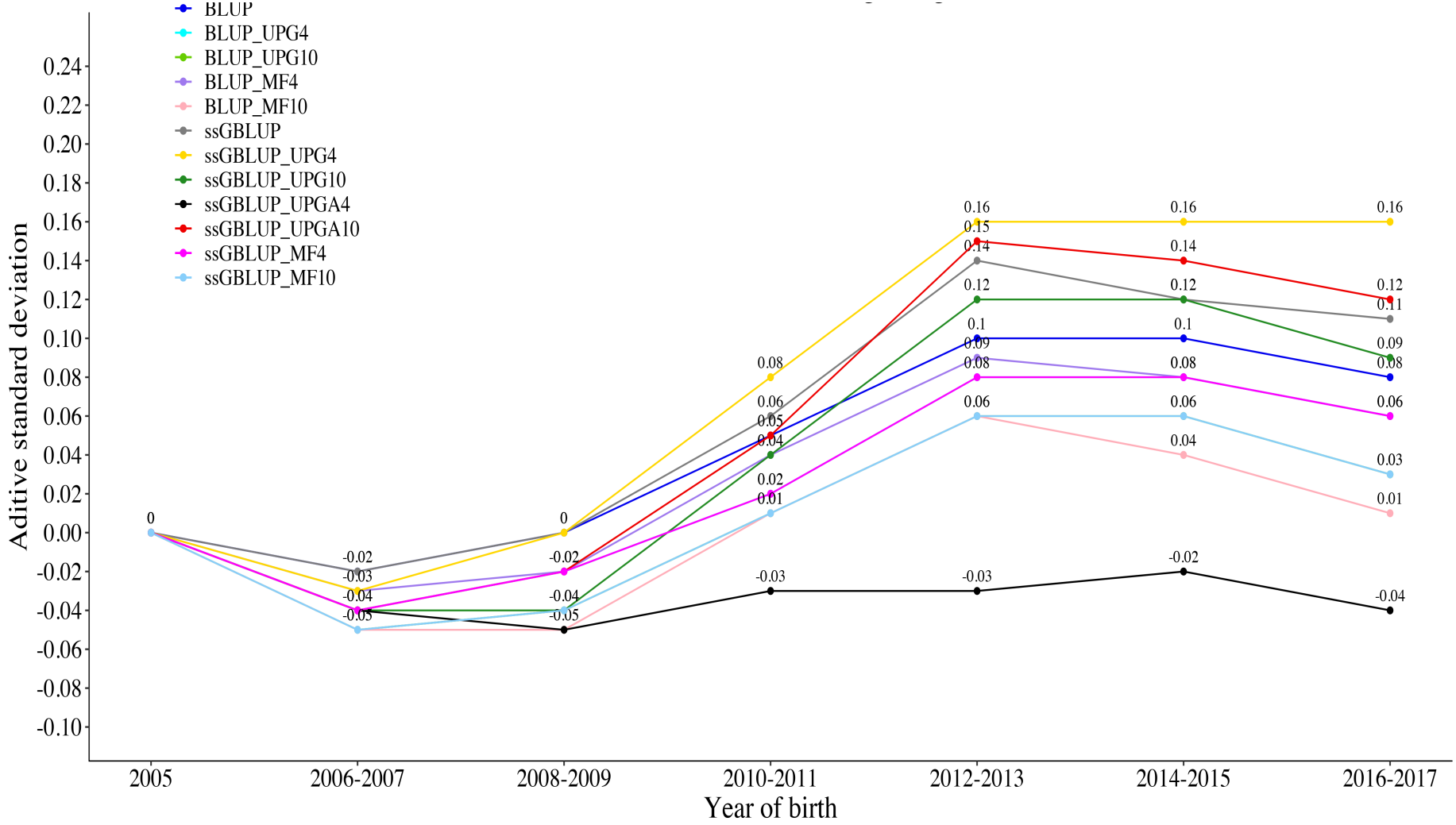

Figure 5. Genetic trends in standard deviation units of genomic estimated breeding values (GEBV) for weaning weight.

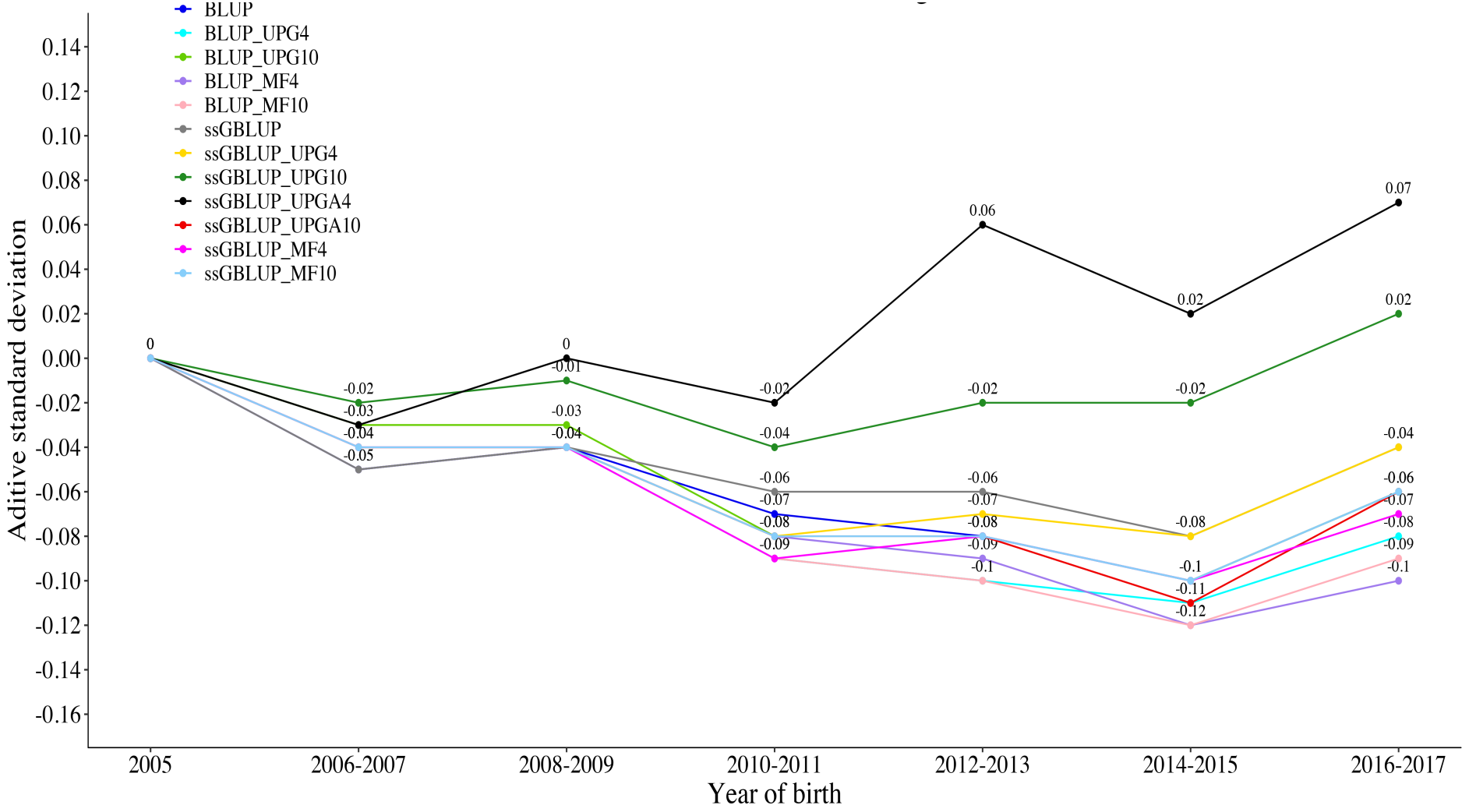

Figure 6. Genetic trends in standard deviation units of genomic estimated breeding values (GEBV) for birth weight.