UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

PREDICTION ABILITY OF CUSTOMIZED SNP ARRAYS WITH DIFFERENT DENSITIES USING THE SINGLE-STEP GENOMIC BLUP METHOD IN A BEEF CATTLE POPULATION

Juan Diego Rodríguez Neira

Zootecnista

2021

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

PREDICTION ABILITY OF CUSTOMIZED SNP ARRAYS WITH DIFFERENT DENSITIES USING THE SINGLE-STEP GENOMIC BLUP METHOD IN A BEEF CATTLE POPULATION

Juan Diego Rodríguez Neira

Orientador: Dr. Fernando Baldi

Coorientador: Dr. Ignacio Aguilar

Coorientador: Dr. Rafael Medeiros de Oliveira Silva

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

2021

R696p

Rodríguez-Neira, Juan Diego

Prediction ability of customized SNP arrays with different densities using the single-step genomic BLUP method in a beef cattle population / Juan Diego Rodríguez-Neira. -- Jaboticabal, 2021

78 p. : il., tabs.

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

Orientador: Fernando Baldi

Coorientador: Ignacio Aguilar

1.Melhoramento animal. 2. Seleção genômica. 3. Bovino de corte. 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.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: PREDICTION ABILITY OF CUSTOMIZED SNP ARRAYS WITH DIFFERENT DENSITIES USING THE SINGLE-STEP GENOMIC BLUP METHOD IN A BEEF CATTLE POPULATION

AUTOR: JUAN DIEGO RODRIGUEZ NEIRA
ORIENTADOR: FERNANDO SEBASTIAN BALDI REY
COORIENTADOR: RAFAEL MEDEIROS DE OLIVEIRA SILVA
COORIENTADOR: IGNACIO AGUILAR

Aprovado como parte das exigências para obtenção do Título de Doutor 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

Dra. LUDMILLA COSTA BRUNES (Participaçao Virtual)
Universidade Federal de Goiás / Goiânia/GO

Prof.Dr. LUIS GABRIEL GONZALEZ HERRERA (Participaçao Virtual)
Universidad Nacional de Colombia / Medellin

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

Jaboticabal, 15 de dezembro de 2021

DADOS CURRICULARES DO AUTOR

JUAN DIEGO RODRÍGUEZ NEIRA – nascido no 25 de maio de 1988 na cidade de Medellín, Estado de Antioquia - Colômbia. Filho de Diego Rodríguez Bermudez e Ana Edith Neira Espinosa. Iniciou em fevereiro de 2005 o curso de graduação em Zootecnia na Universidad Nacional de Colombia -Sede Medellín, obtendo o título de Zootecnista em dezembro de 2010. Durante a graduação, fez parte do grupo de pesquisa "Biología y Genética Molecular – BIOGEM" da mesma universidade. No período de 2011 a 2012 atuou como pesquisador auxiliar e participou do desenvolvimento de projetos de pesquisa e extensão do departamento de produção animal da Universidad Nacional de Colombia – Sede Medellín. No 2012 obteve a distinção de "Joven investigador – COLCIENCIAS". Iniciou em fevereiro de 2013 o programa de pós-graduação em "Ciencias Agrarias – Mejoramiento y genética molecular" na Universidad Nacional de Colombia – Sede Medellín, obtendo o titulo de Mestre em Ciencias Agrarias em agosto de 2016. Nesse mesmo ano atuou como professor da disciplina "Biotecnología Agropecuária" e participou no desenvolvimento de projetos de pesquisa do grupo BIOGEM na mesma universidade. No 2017, atou como professor das disciplinas "Biología Molecular y Genética", "Mejoramiento Animal I" y "Mejoramiento Animal II" na Facultad de Ciencias Agropecuarias da Universidad de Cundinamarca – Sede Fusagasugá. Em março de 2018, ingressou no curso de doutorado do programa de Pós-graduação em Genética e Melhoramento Animal da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista "Júlio de Mesquita Filho", campus de Jaboticabal, sob orientação do Prof. Dr. Fernando Sebastián Baldi Rey, como bolsista do Programa Estudante Convenio de Pós-graduação "PEC-PG" (Edital no. 32/2017) da CAPES.

Financial support

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Dedico

Aos meus pais Diego e Ana,

minhas irmãs Isabel Cristina e Maria Alejandra e

aos meus sobrinhos Sebastián e Helen.

SUMMARY

HABILIDADE DE PREDIÇÃO DE PAINÉIS DE SNP CUSTOMIZADOS COM DIFERENTES DENSIDADES USANDO O MÉTODO DE PASSO ÚNICO GENÔMICO BLUP EM UMA POPULAÇÃO DE GADO DE CORTE

RESUMO – A implementação da seleção genômica (SG) nos programas de melhoramento de bovinos de corte no Brasil tem sido suportado por o incremento do progresso genético das populações de bovinos Nelore. No entanto, um desafio de custo-benefício surge com base na quantidade e qualidade dos marcadores SNPs usados para predições acuradas dos valores genômicos (GEBVs). Portanto, o objetivo deste estudo é avaliar a habilidade de predição de marcadores SNPs em dados simulados e reais sobre características de alta e baixa herdabilidade por meio do método de passo único GBLUP, empregando painéis customizados de baixa e moderada densidade. No capítulo 2, oito réplicas de 18.000 animais genotipados foram simulados com painéis de 335k SNPs, dos quais 6.000 foram a população de validação para predizer os GEBVs. Critérios de alta (H-I) e baixa (L-I) informatividade além da localização equidistante (E-D) dos marcadores foram usados para customizar quatro densidades (35k, 16k, 4k e 2k) de painéis. Assim, em conjunto com o painel de 335k SNPs, treze cenários foram avaliados. Adicionalmente, os painéis customizados foram imputados a 335k SNPs e a acurácia da imputação foi estimada. Em geral, pelo menos 5% dos marcadores H-I ou E-D do painel de 335k SNPs alcançaram predições genômicas confiáveis e imputações acuradas. No capítulo 3, registros de 945 animais genotipados com alta densidade (HD) nascidos entre 1974 e 2018 para peso aos 210 (W210) e 450 (W450) dias de idade e o efeito materno para W210 (MW210) foram usados, dos quais 247 animais nascido despois do 2008 foram a população de validação. Cinco painéis de diferentes densidades (40k, 20k, 10k, 5k e 2k) foram customizados pelos critérios H-I e E-D e o painel HD foi usado como o cenário desejável. As predições dos GEBVs e a acurácia BIF (Beef Improvement Federation) foram obtidas com os programas da família BLUPF90. O método de regressão linear foi usado para avaliar a habilidade de predição, inflação e viés dos GEBVs para cada painel customizado. Uma superestimação da acurácia BIF foi observada quando diminuiu a densidade dos painéis. Para todas as características, a habilidade de predição apresentou aumento com altas densidades e foi similar para painéis customizados com densidades superiores a 10k SNPs, mesmo assim, a inflação foi baixa com altas densidades e o MW210 presentou as maiores inflações. O viés foi susceptível a superestimação dos GEBVs quando a densidade dos painéis customizados diminuiu. Em geral, a acurácia BIF e o nível do viés foram sensíveis aos painéis customizados de baixa densidade enquanto a habilidade de predição com pelo menos 5.000 H-I ou E-D

marcadores a partir do painel HD, apresentou predições acuradas e com menos viés. Esses resultados indicaram que o desenvolvimento de painéis customizados de baixa densidade poderia ser uma abordagem viável para SG utilizando o método ssGBLUP e com melhor custo-benefício nos programas de melhoramento de bovinos de corte.

Palavras-chave: Acurácia da predição, simulação, Acurácia da imputação, seleção genômica, bovinos de corte, inflação, MAF, peso aos 210 dias, peso aos 450 dias.

PREDICTION ABILITY OF CUSTOMIZED SNP ARRAYS WITH DIFFERENT DENSITIES USING THE SINGLE-STEP GENOMIC BLUP METHOD IN A BEEF CATTLE POPULATION

ABSTRACT - The implementation of genomic selection (GS) in Brazilian beef cattle breeding programs has been supported by the increase in genetic progress of Nellore cattle populations. However, from the fundamental principles of GS, a cost-benefit challenge arises based on the quantity and quality of SNP markers used for the prediction of reliable genomic values (GEBVs). Therefore, the objective of this study is to evaluate the prediction ability of SNP markers in simulated and real data on high and low heritability traits through of single-step GBLUP method, employing lower and moderate density customized arrays. In chapter 2, eight replicates of 18,000 genotyped animals were simulated with 335k SNPs array, of which 6,000 were the validation population to predict GEBVs. Four densities of customized SNPs arrays (35k, 16k, 4k and 2k) created from high (H-I) and low (L-I) informativeness and evenly distanced SNP markers (E-D) of SNPs, were the criteria to marker select. Thus, in conjunction with the 335k SNPs array, thirteen scenarios were evaluated. In addition, the customized arrays were imputed to 335k SNPs and the accuracy of imputation was estimated. Overall, at least 5% H-I or E-D markers from 335k array achieved reliable genomic predictions and accurate imputations. In chapter 3, records of 945 genotyped animal with high-density (HD) born between 1974 and 2018 for weight to 210 (W210) and 450 (W450) days of age and maternal effect of W210 (MW210) were used, of which 247 animals born after 2008 were the validation population. Five density arrays were customized (40k, 20k, 10k, 5k and 2k) by H-I and E-D criteria and the HD array was used as desirable scenario. The GEBV predictions and accuracy BIF (beef improvement federation) were obtained with BLUPF90 family programs. The linear regression method was used to evaluate the prediction ability, inflation, and bias of GEBV of each customized array. An overestimation of BIF accuracy was observed when the density arrays decreased. For all traits, the prediction ability increased with higher densities and was similar for customized arrays above 10k, as well, the inflation was low with higher densities and the MW210 effect displayed the higher inflations. The bias was susceptible to overestimation of GEBVs when the density customized arrays decrease. Overall, the BIF accuracy and the level of bias were sensible to of low-density customized arrays while the prediction ability at least 5,000 H-I or E-D markers from HD array, displayed accurate and less biased predictions. These results indicated that the development of low-density customized arrays might be an approach feasible to GS under ssGBLUP method and cost-effective in beef cattle breeding programs.

Keywords: Prediction accuracy, simulation, accuracy of imputation, genomic selection, beef cattle, inflation, MAF, weight at 210 days, weight at 450 days.

CHAPTER 1 – General considerations

INTRODUCTION

In the middle of the last century, Brazil begun to incursionary in the selection of representative animals of the Nellore breed, with the importation of the founders' individuals of the current population (Ferreira de Oliveira et al., 2002) and, after three decades the first Nellore breeding programs was structured (Ferraz and Fries, 2004; Fries and Ferraz, 2006). The genetic evaluation in Brazil involved independent groups formed by universities, research institutes, consultant geneticists and quantitative geneticist service companies (Berry et al., 2016; Albuquerque et al., 2017). These programs reflect differences in the strategies to improvement based in the propped conditions of each group, thus is expected variations and diversity in the current populations of Nellore cattle (Carvalheiro, 2014). However, the high correlation between growth traits and the potential of mature of individuals achieved a common interest among the genetic programs to be included in genetic evaluations for selection purposes (Laureano et al., 2011). Subsequently, the progress of Nellore genetic evaluations from structured programs allowed the incorporation of complex traits as female fertility and the inclusion maternal effects (Lôbo et al., 2010).

The advances in molecular technologies and the employed of them in genetic evaluations impacted strongly the genetic progress of cattle populations. In 2010, Brazil began with the genomic evaluation of Nellore and achieved a notable increase of science in the application of genomic technologies in Brazilian beef breeding programs has occurred (Berry et al., 2016). The inclusion of genome-wide association (GWAS) procedure to identify genomic regions associated with economic importance traits, increased the accuracy of breeding values and the genetic progress (Albuquerque et al., 2017). Additionally, the genomic selection (GS) approach through of prediction of the breeding values of young animals or animals without phenotype information, based on GWAS performed on training populations to estimate marker effects (Montaldo et al., 2012) helped to

complement and accelerate the genetic evaluations and obtain reliable genetic values (Carvalheiro, 2014).

Since early decades of implementation of the genetic improvement programs, only the animal's own records and records of its closest relatives and eventually all relatives were considered (VanRaden, 2020). However, in order to improve the genetic progress of the traits of economic importance, Smith, (1967), claim that consider the information coming from DNA could improvement the estimates of breeding values. Around 80's, Stam, (1980), derivate and demonstrated the distribution of the genome fraction identical by descending and Soller and Beckmann, (1983), proposed structure a genomic relationship more precise employing DNA markers that subsequently improvement the parentage determination and identification of QTL.

In 2001, (Meuwissen et al., 2001), introduced the genomic selection approach that use the genomic information through of inclusion of arrays composted by molecular markers of single-nucleotide polymorphism (SNP) distributed into whole genome to predictions of genomic breeding values (GEBVs). The GS allows more parentage control by capture more precise DNA inheritance (VanRaden, 2020), increase accuracy of models, identified higher proportion of additive genetic variances, increase the genetic gain by decrease generational interval and displays advantage on complex traits (VanRaden, 2008). The genomic prediction arises from the estimation of effects for each SNPs include in the analysis (Meuwissen, 2009), through of non-parametric (Meuwissen et al., 2001) and parametric methods (Misztal et al., 2009; VanRaden, 2008). The single-step Genomic BLUP (ssGBLUP) is a parametric method that combines information from genotyped and non-genotyped animals in a unique analysis (Aguilar et al., 2010) and has been relevant to be use in genomic improvement programs by display less bias and improve the accuracy of genomic predictions (Mäntysaari et al., 2020).

The GS through of estimation of marker effects allow estimate the independent chromosome segments (ICS) and knowing the shared alleles of genotyped animals, and indirectly the effective population size (Ne) (Misztal et al.,

2020). Thus, the optimal density of markers to detected QTLs is relate directly with Ne (Lee et al., 2017). The farm animals present a small Ne due that were originated from a narrow genetic basis thus the number of marker and ICS to estimate is lower to obtain accurate genomic predictions (Misztal et al., 2020). In Nellore cattle, few studies have been development to knowing the Ne; however, Cardoso et al. (2018) achieved identified a diverse and structured population from Ne above of 100 when the selection is implemented. Therefore, to achieve accurate predictions in GS, it's possible the use of low and moderate density arrays varying of 10k to 50k SNPs (Georges et al., 2019).

Commonly the large number of arrays to cattle protected by intellectual property and built principally to Taurus cattle are frequently employed in Indicus cattle, being that both species no segregate the same markers (Boddhireddy et al., 2014). The advances in genotyping techniques in the last decade, the genotyping cost decreased approximately 80%, promoting customize and standardize low, as well as, moderate density arrays to carry out in genomic evaluations (VanRaden, 2020).

The use of a lower number of markers for genomic predictions has been showed accurate predictions and slightly different in contrast whit the use of more dense arrays, highlining the LD, MAF and position of SNPS into the genome as criteria to select markers (Barjasteh et al., 2010; Boddhireddy et al., 2014; Su et al., 2012; Wu et al., 2016). Likewise, the incorporation of approaches to imputation of genotypes in genomic predictions from low and moderate customized arrays has displayed reliable imputation with low imputed allele errors (Aliloo et al., 2018; H. Aliloo et al., 2018; Barjasteh et al., 2010; Lopez et al., 2020; Mulder et al., 2012; Shashkova et al., 2021). Therefore, the development of customized arrays is a great chose for genomic predictions due could reduce the cost of implementation of genetic improve programs as well as the use in other approaches such the imputations of genotypes.

GENERAL OBJECTIVE

Evaluate the prediction ability of SNP markers for simulate and real data on high and low heritability traits through of single-step GBLUP method, employing lower and moderate density customized arrays.

Specific objectives

Evaluate the impact of SNP marker density and informativeness on prediction ability and imputation performance using the single-step genomic BLUP in a simulated beef cattle population.

Evaluate the prediction ability for growth and maternal-related traits using SNPs arrays with different marker densities based on SNP informativeness employing the ssGBLUP methodology in a Nellore beef cattle population.

LITERATURE REVIEW

Genomic selection and Density of markers

Genomic selection (GS) is a technique provided by inclusion of molecular markers of single-nucleotide polymorphism (SNP) distributed into whole genome to predictions of genomic breeding values (GEBVs) of future animals' candidates to selection (Meuwissen et al., 2001). The GS estimate in a representative portion of population called reference animals, the values of SNP marker effects to apply in young animals or candidate animals and estimate the GEBVs (Mäntysaari et al., 2020). The accuracy of GEBVs could be influenced by several elements, as effective population size, the genetic architecture of trait (Daetwyler et al., 2010), the response variable to estimate the SNPs effect (Garrick et al., 2009), the minor allele frequency (MAF), the linkage disequilibrium (LD) between quantitative trait loci (QTL) and the marker, heritability of trait, the method used to genomic

predictions and the density of markers arrays (de los Campos et al., 2013). In general, the larger genomes display through the generations a higher disruption of LD (Ballesta et al., 2020), and this fact is associated with a higher number of independent chromosome segments (Me), being necessary a higher number of available markers to detect QTL (Lee et al., 2017).

Prediction of GEBVs of individuals from genomic information employing SNPs High Density (HD) arrays has been implemented in several species (Goddard and Hayes, 2009), as beef and dairy cattle (Fragomeni et al., 2019; Harder et al., 2020; Liu et al., 2019; Lopes et al., 2020; Rezende et al., 2019; Silva et al., 2021), pigs (Christensen et al., 2012; Silveira et al., 2020; Song et al., 2019; Waide et al., 2018), goats (de Lima et al., 2020; Gipson, 2019; Molina et al., 2017; Teissier et al., 2019), aquaculture (Garcia et al., 2018; Joshi et al., 2020; Liu et al., 2020; Vallejo et al., 2018; Wang et al., 2017), and plants (Cappa et al., 2019; Heffner et al., 2009; Resende et al., 2012; Sousa et al., 2019). Should expected that with the use of great number of markers through of HD arrays, higher levels of LD between SNPs could observe and consequently, find QTLs associated with productive important traits. However, the implementation of HD arrays makes more expensive the approaches based in genomic selection (Salvian et al., 2020). In this aspect, Misztal et al. (2020) reported that the genome in most of the cattle breeds were originated from a narrow genetic basis, therefore, the current breeds display low effective population size (Ne), furthermore, exhibit a structure with lower number of Me and claim is not necessary a great number of markers to predict GEBVs (Boichard et al., 2012; Hayes et al., 2012; Rolf et al., 2010; Weigel et al., 2010). In this sense, Su et al. (2012) contrasted the genomic predictions between medium (54k) and HD (777k) arrays, with different traits in Holstein and Red Dairy cattle. Boddhireddy et al. (2014) in Nellore cattle compared 54k SNPs and HD arrays (777k) for reproductive, productive, and visual body conformation scores; Wu et al. (2018) evaluated genomic predictions for growth and reproductive traits from Bos Indicus low-density (~35k) and HD (777k) in Nellore cattle and Barjasteh et al. (2020), in several simulated cattle populations working traits with heritability

varying from 0.25 to 0.50, employed medium-density SNPs arrays (50k) and HD SNPs arrays (777k).

In general, all studies concluded that the density arrays not influenced the GEBVs and the accuracy of genomic predictions in response that the HD arrays increase slightly the accuracy of genomic predictions compared with medium and low-density arrays. Therefore, the use of different low-density SNPs arrays in genetic improvement programs is a feasible strategy due that reduce of genotyping cost and computational time that implies the use of HD arrays in genomic predictions.

Customized arrays to prediction and imputation

The use of moderate or low-density customized arrays in genomic predictions or implemented for imputation to infer HD genotypes are considered as an alternative strategy that open opportunity for the animal breeding programs (Aliloo et al., 2018). Several criteria as the marker effects (Sousa et al., 2019), positions of SNPs (Zhang et al., 2015), haplotype blocks (Ma et al., 2015), markers significant in GWAS (Subedi et al., 2013), markers with high linkage disequilibrium (LD) and MAF, could be contemplated to select representative SNPs with high effect to customized arrays. Zhu et al. (2017), considered different levels of MAF and evenly distribution of SNPs to customize low density arrays provided relevant information for use in Simmental Chinese cattle. Likewise, Chen et al. (2011) in broilers displayed higher genomic predictions when considered a threshold of 0.4 for MAF for selected markers. Both studies demonstrated that selection of markers impact in genetic structure and change the G matrix properties. Other facts in conjunction with position and MAF of markers as the strongly relationship between reference and validation population, allow identify representative markers for customize low-density arrays and provide similar genomic predictions than HD arrays (Barjasteh et al., 2020). In this sense, to obtain reliable genomic predictions, the G matrix due achieve special conditions given specific criteria of markers and populations. Thus, the use of a low proportion of available markers from HD array

is feasible for provide reliable genomic predictions (Rolf et al., 2010; Salvian et al., 2020).

Additionally, in the imputation approach, the quality of imputed genotypes from customized arrays, detected through of accuracy of imputation is an important approach influenced by several factors as chromosomal position and MAF of selected SNPs (Shashkova et al., 2021), the structure of the specific target population (Aliloo et al., 2018) and the method used to impute. Different researches have focused to development and implement of low-density arrays to imputation in cattle (Aliloo et al., 2018; Judge et al., 2016; Korkuć et al., 2019; Zhang and Druet, 2010) based on that the genomic predictions with HD are similar than imputed genotypes, when the accuracy of imputation is above 90% (Wellmann et al., 2013). In general, the selection of SNPs based on MAF, LD and position of markers to customize arrays, report values of imputation accuracy adequate for apply in genomic evaluations (Judge et al., 2016; Mulder et al., 2012). Carvalheiro et al. (2014), in Nellore cattle using low and medium SNPs arrays created based on MAF and LD reported high imputation accuracy and displayed slight difference of results using FIMPUTE and BEAGLE software. Likewise, Kranjčevičová et al. (2019) employed customized arrays based on randomly selected SNPs markers and observed decay of accuracy of imputation when the number of markers to impute is above of 70% from reference array. The same tendence has been observed in other species as the sheep. O'Brien et al. (2019) reported variations in allele concordance rate associated with multi-breeds and showed that employing customized arrays based on MAF, LD and markers evenly distributed, achieve a high accuracy of imputation when at least 12% of markers available from 50k arrays are used.

Bias and dispersion of genomic predictions

The genomic evaluation depends of adequate statistic methods to remove biases of predictions (Vitezica et al., 2011). Therefore, considerer the bias and inflation of genomic predictions is a necessary approach to compare GEBVs of

young and older individuals (Mäntysaari et al., 2010). The bias, is define with the correctly prediction of mean GEBV of young genotyped animals (Granado-Tajada et al., 2020) through the difference between predicted and modeled values (Mäntysaari et al., 2020). The preselection of young animals without phenotype could impact to increase of bias of the predictions by the rise in the genetic level and reduction in genetic variance (Legarra and Reverter, 2018; Macedo et al., 2020; Vitezica et al., 2011), as well as the proportion of genotyped individual selected, the structure of trait and the scenarios used to selected (Gowane et al., 2018). Further, the genomics models display biases associated with sample error of genomic matrix and the number of markers necessary for knowing genomic relationship (Goddard et al., 2011). This fact was observed by Chen et al. (2011), that with a low number of markers increased the sample error and the genomic relationship bias between individuals. Goddard et al. (2011) proposed to diminish the bias, including the sampling error into the estimation of matrix G through of regressed elements of G against relationship matrix (A). Actually, the use of ssGBLUP procedure has increase due to displayed a diminished of genomic prediction bias and improve accuracy from candidate animals (Mäntysaari et al., 2020). This method includes all information (phenotypic and genomic) and only considers genotyped animals into the matrix G to predicted the GEBVs of non-genotyped individuals (Vitezica et al., 2011). The structure of matrix G is a key component to bias of predictions due that the dimensionality given for the density arrays used (Pocrnic et al., 2016) and the errors of genotyped could displayed bias of predictions between genotyped and non-genotyped individuals (Mulder et al., 2012; Nordbø et al., 2019). Criteria via ssGBLUP has been detected to controlling the bias of predictions as include the inbreeding in a selected population, strong selection on a trait and the compatibility between pedigree and genomic relationship (Tsuruta et al., 2019) that could improve with employing of metafounders (Legarra et al., 2015).

The dispersion of GEBVs is measured though regression coefficient modeled GEBVs on reduced GEBVs (Vitezica et al., 2011). Over or under dispersion cause inflation or deflation of GEBVs and occur when the regression

coefficient is less or over to 1 (Neves et al., 2012). The reduction of inflation of genomic predictions for young genotyped individuals its possible through of selection and the use of genetic parameters recently estimated, adjusting the weight of the relationship matrix for genotyped animals to guarantee a smaller additive genetic variance (Tsuruta et al., 2019). This parameter is an indicator of quality of genomic prediction and allows the fair comparison of GEBVs among animals and consequently proper selection decisions (Piccoli et al., 2018). This key component has commonly been called bias, which can be justified because, in selected traits, bias and dispersion are related (Vitezica et al., 2011). In the same way, Ma et al. (2015) in Danish Jersey, reported that ssGBLUP approach including all female's information and pedigree as well as information from genotyped bulls and cows displayed reduced inflation GEBVs and reported that the incorporation of approximately 5% of available SNPs from HD array displayed similar inflation of predictions when compared to HD. Salvian et al. (2020) in broilers reported alike results and obtained less inflated predictions when used arrays close to 5% of available SNPs from HD array (370.600 SNPs). Additionally, implemented imputed genotypes from customized arrays. Mulder et al. (2012) reported lower inflated predictions with the diminishes of SNPs to be imputed. Likewise, Lopez et al. (2020) identified lower and deflated prediction with imputed genotypes to HD from 50k SNPs arrays when employed a single-trait models in commercial Hanwoo beef cattle.

REFERENCES

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., 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. https://doi.org/10.3168/jds.2009-2730

Albuquerque, L.G., Fernandes Júnior, G.A., Carvalheiro, R., 2017. Beef cattle genomic selection in tropical environments. Proceedings of the 22nd Conference of the Association for the Advancement of Animal Breeding and Genetics (AAABG), Townsville, Queensland, Australia, 2-5 July 2017 255–263.

Aliloo, H., Mrode, R., Okeyo, A., Ni, G., Goddard, M., Gibson, J., 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. Journal of Dairy Science 101, 9108–9127. https://doi.org/10.3168/jds.2018-14621

Aliloo, H., Mrode, R., Okeyo, M., Ojango, J., Dessie, T., Rege, E., Goddard, M., Gibson, J., 2018. Optimal design of low-density marker panels for genotype imputation. Proceedings of the World Congress on Genetics Applied to Livestock Production 11, 146.

Ballesta, P., Bush, D., Fonseca Silva, F., Mora, F., 2020. Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species Eucalyptus cladocalyx. Plants (Basel, Switzerland) 9(1):99 https://doi.org/10.3390/plants9010099

Barjasteh, S., Dashab, G.R., Rokouei, M., Shariati, M.M., Vafaye Valleh, M., 2010. Comparing Different Marker Densities and Various Reference Populations Using Pedigree-Marker Best Linear Unbiased Prediction (BLUP) Model, Iranian Journal of Applied Animal Science. Islamic Azad University - Rasht Branch.

Berry, D.P., Garcia, J.F., Garrick, D.J., 2016. Development and implementation of genomic predictions in beef cattle. Animal Frontiers 6, 32–38. https://doi.org/10.2527/af.2016-0005

Boddhireddy, P., Prayaga, K., Barros, P., Lôbo, R., Denise, S., 2014. Proceedings, 10 th World Congress of Genetics Applied to Livestock Production Genomic Predictions of Economically Important Traits in Nelore Cattle of Brazil.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., 2012. Design of a bovine low-density snp array optimized for imputation. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0034130

Cappa, E.P., Marco De Lima, B., da Silva-Junior, O.B., Garcia, C.C., Mansfield, S.D., Grattapaglia, D., 2019. Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. https://doi.org/10.1016/j.plantsci.2019.03.017

Cardoso, D.F., de Albuquerque, L.G., Reimer, C., Qanbari, S., Erbe, M., do Nascimento, A. v., Venturini, G.C., Scalez, D.C.B., Baldi, F., de Camargo, G.M.F., Mercadante, M.E.Z., do Santos Gonçalves Cyrillo, J.N., Simianer, H., Tonhati, H., 2018. Genome-wide scan reveals population stratification and footprints of recent selection in Nelore cattle. Genetics Selection Evolution 50, 22. https://doi.org/10.1186/s12711-018-0381-2

Carvalheiro, R., 2014. Proceedings, 10 th World Congress of Genetics Applied to Livestock Production Genomic Selection in Nelore Cattle in Brazil.

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science 89, 2673–2679. https://doi.org/10.2527/jas.2010-3555

Christensen, O.F., Madsen, P., Nielsen, B., Ostersen, T., Su, G., 2012. Single-step methods for genomic evaluation in pigs. animal 6, 1565–1571. https://doi.org/10.1017/S1751731112000742

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185, 1021–1031. https://doi.org/10.1534/genetics.110.116855

de Lima, L.G., de Souza, N.O.B., Rios, R.R., de Melo, B.A., dos Santos, L.T.A., Silva, K. de M., Murphy, T.W., Fraga, A.B., 2020. Advances in molecular genetic techniques applied to selection for litter size in goats (Capra hircus): a review. Journal of Applied Animal Research. https://doi.org/10.1080/09712119.2020.1717497

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. https://doi.org/10.1534/genetics.112.143313

Ferraz, J.B.S., Fries, L.A., 2004. V V Simpósio Nacional da Sociedade Brasileira de Simpósio Nacional da Sociedade Brasileira de Melhoramento Animal Melhoramento Animal Beef Cattle Genetic Evaluation programs in Brazil.

Ferreira de Oliveira, J.H., Magnabosco, C. de U., Souza, D., 2002. Nelore: Base Genética e Evolução Seletiva no Brasil.

Fragomeni, B., Lourenco, D., Legarra, A., VanRaden, P., Misztal, I., 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. Journal of Dairy Science 102, 10012–10019. https://doi.org/10.3168/jds.2019-16262

Fries, L.A., Ferraz, J.B.S., 2006. 8th World Congress on Genetics Applied to Livestock Production. BEEF CATTLE GENETIC PROGRAMMES IN BRAZIL.

Garcia, A.L.S., Bosworth, B., Waldbieser, G., Misztal, I., Tsuruta, S., Lourenco, D.A.L., 2018. Development of genomic predictions for harvest and carcass weight in channel catfish 06 Biological Sciences 0604 Genetics. Genetics Selection Evolution 50, 66. https://doi.org/10.1186/s12711-018-0435-5

Garrick, D.J., Taylor, J.F., Fernando, R.L., 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution 2009 41:1 41, 1–8. https://doi.org/10.1186/1297-9686-41-55

Georges, M., Charlier, C., Hayes, B., 2019. Harnessing genomic information for livestock improvement. Nature Reviews Genetics 20, 135–156. https://doi.org/10.1038/s41576-018-0082-2

Gipson, T.A., 2019. — Special Issue — Recent advances in breeding and genetics for dairy goats. Asian-Australasian Journal of Animal Sciences 32, 1275–1283. https://doi.org/10.5713/ajas.19.0381

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381–391. https://doi.org/10.1038/nrg2575

Goddard, M.E., Hayes, B.J., Meuwissen, T.H.E., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128, 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Gowane, G.R., Lee, S.H., Clark, S., Moghaddar, N., Al-Mamun, H.A., Werf, J.H.J. van der, 2018. Effect of selection on bias and accuracy in genomic prediction of breeding values. bioRxiv 298042. https://doi.org/10.1101/298042

Granado-Tajada, I., Legarra, A., Ugarte, E., 2020. Exploring the inclusion of genomic information and metafounders in Latxa dairy sheep genetic evaluations. https://doi.org/10.3168/jds.2019-18033

Harder, I., Stamer, E., Junge, W., Thaller, G., 2020. Estimation of genetic parameters and breeding values for feed intake and energy balance using pedigree relationships or single-step genomic evaluation in Holstein Friesian cows. Journal of Dairy Science 103, 2498–2513. https://doi.org/10.3168/jds.2019-16855

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Animal Genetics 43, 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x

Heffner, E.L., Sorrells, M.E., Jannink, J.L., 2009. Genomic selection for crop improvement. Crop Science. https://doi.org/10.2135/cropsci2008.08.0512

Joshi, R., Skaarud, A., de Vera, M., Alvarez, A.T., Ødegård, J., 2020. Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (Oreochromis niloticus). Aquaculture 516, 734641. https://doi.org/10.1016/j.aquaculture.2019.734641

Judge, M.M., Kearney, J.F., McClure, M.C., Sleator, R.D., Berry, D.P., 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations. Journal of Animal Science 94, 949–962. https://doi.org/10.2527/jas.2015-0044

Korkuć, P., Arends, D., Brockmann, G.A., 2019. Finding the optimal imputation strategy for small cattle populations. Frontiers in Genetics 10. https://doi.org/10.3389/FGENE.2019.00052

Kranjčevičová, A., Kašná, E., Brzáková, M., Přibyl, J., Vostrý, L., 2019. Impact of reference population size and marker density on accuracy of population imputation. Czech Journal of Animal Science 64, 405–410. https://doi.org/10.17221/148/2019-CJAS

Laureano, M.M.M., Boligon, A.A., Costa, R.B., Forni, S., Severo, J.L.P., Albuquerque, L.G., 2011. Estimativas de herdabilidade e tendências genéticas para características de crescimento e reprodutivas em bovinos da raça Nelore: Estimates of heritability and genetic trends for growth and reproduction traits in Nelore cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia 63, 143–152. https://doi.org/10.1590/S0102-09352011000100022

Lee, S.H., Clark, S., van der Werf, J.H.J., 2017. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0189775

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., Misztal, I., 2015. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics 200, 455–468. https://doi.org/10.1534/GENETICS.115.177014

Legarra, A., Reverter, A., 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method 01 Mathematical Sciences 0104 Statistics. Genetics Selection Evolution 50, 53. https://doi.org/10.1186/s12711-018-0426-6

Liu, J., Yang, G., Kong, J., Xia, Z., Sui, J., Tang, Q., Luo, K., Dai, P., Lu, X., Meng, X., Luan, S., 2020. Using single-step genomic best linear unbiased prediction to improve the efficiency of genetic evaluation on body weight in Macrobrachium rosenbergii. https://doi.org/10.1016/j.aquaculture.2020.735577

Liu, Y., Xu, Lei, Wang, Z., Xu, Ling, Chen, Y., Zhang, L., Xu, Lingyang, Gao, X., Gao, H., Zhu, B., Li, J., 2019. Genomic prediction and association analysis with models including dominance effects for important traits in Chinese simmental beef cattle. Animals 9. https://doi.org/10.3390/ani9121055

Lôbo, R.B., Bittnecourt, T.C.B. dos S.C. de, 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. https://doi.org/10.1590/S1516-35982010001300025

Lopes, F.B., Baldi, F., Passafaro, T.L., Brunes, L.C., Costa, M.F.O., Eifert, E.C., Narciso, M.G., Rosa, G.J.M., Lobo, R.B., Magnabosco, C.U., 2020. Genome-enabled prediction of meat and carcass traits using Bayesian regression, single-step genomic best linear unbiased prediction and blending methods in Nelore cattle-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). https://doi.org/10.1016/j.animal.2020.100006

Lopez, B.I., Lee, S.H., Shin, D.H., Oh, J.D., Chai, H.H., Park, W., Park, J.E., Lim, D., 2020. Accuracy of genomic evaluation using imputed high-density genotypes for carcass traits in commercial Hanwoo population. Livestock Science 241, 104256. https://doi.org/10.1016/j.livsci.2020.104256

Ma, P., Lund, M.S., Nielsen, U.S., Aamand, G.P., Su, G., 2015. Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. Journal of Dairy Science 98, 9026–9034. https://doi.org/10.3168/jds.2015-9703

Macedo, F.L., Reverter, A., Legarra, A., 2020. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. Journal of Dairy Science 103, 529–544. https://doi.org/10.3168/jds.2019-16603

Mäntysaari, E., Zengting, L., Vanraden, P., 2010. Interbull validation test for genomic evaluations. Interbull bulletin 17.

Mäntysaari, E.A., Koivula, M., Strandén, I., 2020. Symposium review: Single-step genomic evaluations in dairy cattle. Journal of Dairy Science 103, 5314–5326. https://doi.org/10.3168/JDS.2019-17754

Meuwissen, T., 2009. Genetic management of small populations: A review. Acta Agriculturae Scandinavica A: Animal Sciences 59, 71–79. https://doi.org/10.1080/09064700903118148

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157.

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–4655. https://doi.org/10.3168/jds.2009-2064

Misztal, I., Lourenco, D., Legarra, A., 2020. Current status of genomic evaluation. Journal of Animal Science 98, 1–14. https://doi.org/10.1093/jas/skaa101

Molina, A., Muñoz, E., Díaz, C., Menéndez-Buxadera, A., Ramón, M., Sánchez, M., Carabaño, M.J., Serradilla, J.M., 2017. Goat genomic selection: Impact of the integration of genomic information in the genetic evaluations of the Spanish Florida goats. https://doi.org/10.1016/j.smallrumres.2017.12.010

Montaldo, H.H., Casas, E., Ferraz, J.B.S., Vega-Murillo, V.E., Román-Ponce, S.I., 2012. Opportunities and challenges from the use of genomic selection for beef cattle breeding in Latin America. Animal Frontiers 2, 23–29. https://doi.org/10.2527/AF.2011-0029

Mulder, H.A., Calus, M.P.L., Druet, T., Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science 95, 876–889. https://doi.org/10.3168/JDS.2011-4490

Neves, H.H., Carvalheiro, R., Queiroz, S.A., 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics 2012 13:1 13, 1–17. https://doi.org/10.1186/1471-2156-13-100

Nordbø, Ø., Gjuvsland, A.B., Eikje, L.S., Meuwissen, T., 2019. Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. Genetics Selection Evolution 51, 1–8. https://doi.org/10.1186/s12711-019-0517-z

O'Brien, A.C., Judge, M.M., Fair, S., Berry, D.P., 2019. High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep. Journal of Animal Science 97, 1550–1567. https://doi.org/10.1093/jas/skz043

Piccoli, M.L., Brito, L.F., Braccini, José, Brito, F. v, Cardoso, F.F., Cobuci, J.A., Sargolzaei, M., Schenkel, F.S., Piccoli, M., Brito, L., Schenkel, F., Braccini, J, Cardoso, F., Cobuci, J., 2018. ARTICLE A comprehensive comparison between single-and two-step GBLUP methods in a simulated beef cattle population. Can. J. Anim. Sci 98, dx. https://doi.org/10.1139/cjas-2017-0176

Pocrnic, I., Lourenco, D.A.L., Masuda, Y., Legarra, A., Misztal, I., 2016. The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics 203, 573. https://doi.org/10.1534/GENETICS.116.187013

Resende, J.F.R., Muñoz, P., Resende, M.D.V., Garrick, D.J., Fernando, R.L., Davis, J.M., Jokela, E.J., Martin, T.A., Peter, G.F., Kirst, M., 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190, 1503–1510. https://doi.org/10.1534/genetics.111.137026

Rezende, F.M., Pablo Nani, J., Peñagaricano, F., 2019. Genomic prediction of bull fertility in US Jersey dairy cattle. Journal of Dairy Science 102, 3230–3240. https://doi.org/10.3168/jds.2018-15810

Rolf, M.M., Taylor, J.F., Schnabel, R.D., McKay, S.D., McClure, M.C., Northcutt, S.L., Kerley, M.S., Weaber, R.L., 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genetics 11. https://doi.org/10.1186/1471-2156-11-24

Salvian, M., Costa, G., Moreira, M., Spangler, M.L., Mourão, G.B., 2020. Estimation of Breeding Values Using Different Densities of Snp to Inform Kinship in Broiler Chickens. https://doi.org/10.21203/rs.3.rs-32429/v1

Shashkova, T.I., Martynova, E.U., Ayupova, A.F., Shumskiy, A.A., Ogurtsova, P.A., Kostyunina, O. v., Khaitovich, P.E., Mazin, P. v., Zinovieva, N.A., 2021. Development of a low-density panel for genomic selection of pigs in Russia1. Translational Animal Science 4, 264–274. https://doi.org/10.1093/TAS/TXZ182

Silva, R.P., Espigolan, R., Berton, M.P., Lôbo, R.B., Magnabosco, C.U., Pereira, A.S.C., Baldi, F., 2021. Genomic prediction ability for carcass composition indicator traits in Nellore cattle. https://doi.org/10.1016/j.livsci.2021.104421

Silveira, L.S., Lima, L.P., Nascimento, M., Nascimento, A.C.C., Silva, F.F., 2020. Regression trees in genomic selection for carcass traits in pigs. Genetics and Molecular Research 19. https://doi.org/10.4238/gmr18498

Smith, C., 1967. Improvement of metric traits through specific genetic loci. Animal Production 9, 349–358. https://doi.org/10.1017/S0003356100038642

Soller, M., Beckmann, J.S., 1983. Genetic polymorphism in varietal identification and genetic improvement. Theoretical and Applied Genetics 1983 67:1 67, 25–33. https://doi.org/10.1007/BF00303917

Song, H., Zhang, J., Zhang, Q., Ding, X., 2019. Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig. Frontiers in Genetics 10. https://doi.org/10.3389/FGENE.2018.00730

Sousa, T.V., Caixeta, E.T., Alkimim, E.R., Oliveira, A.C.B., Pereira, A.A., Sakiyama, N.S., Zambolim, L., Resende, M.D.V., 2019. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. Frontiers in Plant Science 9. https://doi.org/10.3389/fpls.2018.01934

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res., Camb 35, 1. https://doi.org/10.1017/S0016672300014002

Su, G., Brøndum, R.F., Ma, P., Guldbrandtsen, B., Aamand, G.P., Lund, M.S., 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95, 4657–4665. https://doi.org/10.3168/jds.2012-5379

Subedi, S., Feng, Z., Deardon, R., Schenkel, F.S., 2013. SNP selection for predicting a quantitative trait. Journal of Applied Statistics 40, 600–613. https://doi.org/10.1080/02664763.2012.750282

Teissier, M., Larroque, H., Robert-Granie, C., 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. Journal of Dairy Science 102, 3142–3154. https://doi.org/10.3168/jds.2018-15650

Tsuruta, S., Lourenco, D.A.L., Masuda, Y., Misztal, I., Lawlor, T.J., 2019. Controlling bias in genomic breeding values for young genotyped bulls. Journal of Dairy Science 102, 9956–9970. https://doi.org/10.3168/jds.2019-16789

Vallejo, R.L., Cheng, H., Fragomeni, B.O., Gao, G., Silva, R.M.O., Martin, K.E., Evenhuis, J.P., Wiens, G.D., Leeds, T.D., Palti, Y., 2018. The accuracy of genomic predictions for bacterial cold water disease resistance remains higher than the pedigree-based model one generation after model training in a commercial rainbow trout breeding population. https://doi.org/10.1016/j.aquaculture.2021.737164

VanRaden, P.M., 2020. Symposium review: How to implement genomic selection. Journal of Dairy Science 103, 5291–5301. https://doi.org/10.3168/jds.2019-17684

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genetics Research 93, 357–366. https://doi.org/10.1017/S001667231100022X

Waide, E.H., Tuggle, C.K., Serão, N.V.L., Schroyen, M., Hess, A., Rowland, R.R.R., Lunney, J.K., Plastow, G., Dekkers, J.C.M., 2018. Genomic prediction of piglet response to infection with one of two porcine reproductive and respiratory syndrome virus isolates. Genetics Selection Evolution 50, 3. https://doi.org/10.1186/s12711-018-0371-4

Wang, Q., Yu, Y., Li, F., Zhang, X., Xiang, J., 2017. Predictive ability of genomic selection models for breeding value estimation on growth traits of Pacific white shrimp Litopenaeus vannamei. Chinese Journal of Oceanology and Limnology 35, 1221–1229. https://doi.org/10.1007/s00343-017-6038-0

Weigel, K.A., van Tassell, C.P., O'Connell, J.R., VanRaden, P.M., Wiggans, G.R., 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229–2238. https://doi.org/10.3168/jds.2009-2849

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K., Bennewitz, J., 2013. Genomic selection using low density marker panels with application to a sire line in pigs. Genetics Selection Evolution 2013 45:1 45, 1–11. https://doi.org/10.1186/1297-9686-45-28

Wu, X.-L., Li, H., Xu, J., Ferraz, J.B.S., Silva, L.R., Garcia, J.F., Tait, R., Bauck, & S., 2018. Evaluation of genomic prediction accuracies of growth and reproduction traits in Nellore cattle using the new GGP® indicus low density SNP chip.

Wu, X.L., Xu, J., Feng, G., Wiggans, G.R., Taylor, J.F., He, J., Qian, C., Qiu, J., Simpson, B., Walker, J., Bauck, S., 2016. Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. PLoS ONE 11, e0161719. https://doi.org/10.1371/journal.pone.0161719

Zhang, Z., Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science 93, 5487–5494. https://doi.org/10.3168/jds.2010-3501

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., Simianer, H., Li, J., 2015. Accuracy of Whole-Genome Prediction Using a Genetic Architecture-Enhanced Variance-Covariance Matrix. G3: Genes|Genomes|Genetics 5, 615. https://doi.org/10.1534/G3.114.016261

Zhu, B., Zhang, J. jing, Niu, H., Guan, L., GUO, P., XU, L. yang, CHEN, Y., ZHANG, L. pei, GAO, H. jiang, GAO, X., LI, J. ya, 2017. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture 16, 911–920. https://doi.org/10.1016/S2095-3119(16)61474-0

CHAPTER 2 - Effect of SNP marker density and informativeness on imputation performance and prediction ability using the single-step genomic BLUP in a simulated beef cattle population

ABSTRACT

In beef cattle populations, there are few evidence about the minimum number of genetic markers need to obtain reliable genomic prediction and imputed genotypes. This study aimed evaluate the impact of single nucleotide polymorphism (SNP) markers densities and informativeness on genomic predictions and imputation performance for high and low heritability traits using the single-step genomic Best Linear Unbiased Prediction methodology (ssGBLUP) in a simulated beef cattle population. The simulated genomic and phenotypic data were obtained through QMsim software. 735,293 SNPs markers and 7000 quantitative trait loci (QTL) were randomly simulated. The mutation rate ($1x10^{-5}$), QTL effects distribution (gamma distribution 0.4) and minor allele frequency (MAF≥0.02) of markers were used for quality control. A total of 335k SNPs (HD) and 1000 QTLs were finally considered. Densities of ~35k, ~16k, ~4k and ~2k SNPs were customized through windows of 10, 20, 80 and 160 SNPs by chromosome, respectively. Three marker selection criteria were used within windows: 1) informative markers with MAF values close to 0.5 (H-I); 2) less informative markers with the lowest MAF values (L-H); 3) markers evenly distributed (E-D). The HD array and twelve scenarios of customized SNPs arrays likewise the performance of imputation was evaluated. The genomic predictions and imputed genotypes were obtained with BLUPF90 and FIMPUTE software's, respectively, and statistics parameters were applied to evaluate the accuracy of genotypes imputed. The Pearson's correlation, the difference between predicted and simulated predictions, and the coefficient of regression were used to evaluate the prediction ability (*PA*), inflation (*b*), and bias (*d*), respectively. Densities above 16k SNPs using H-I and E-D criteria displayed lower *b*, and higher *PA* and imputation accuracy, and consequently, similar values of *PA, b* and *d* were observed with the use of imputed

genotypes. The L-I criteria showed higher *PA*, adequate accuracy imputation and lower *b* with densities higher to 35k SNPs, however the quality of imputed genotypes was low. The results obtained showed that at least 5% of H-I or E-D SNPs available in the HD are necessary to achieve reliable genomic predictions and imputed genotypes. The development of low-density customized arrays based on criteria of high informativeness and evenly distribution of SNPs, might be a cost-effective and feasible approach to implement genomic selection in beef cattle.

KEYWORDS: customized SNP arrays, genomic selection, imputation accuracy, inflation, simulation.

INTRODUCTION

Genomic selection has enabled the genetic improvement of quantitative traits using whole genome molecular markers and it has been successfully implemented in several livestock breeding programs (de Lima et al., 2020; Liu et al., 2019; Silveira et al., 2020). Genomic prediction combines high-density markers obtained from high throughput genotyping with phenotypic and pedigree data to increase the prediction accuracy of breeding values (Meuwissen et al., 2001). It is widely reported that genomic prediction accuracy is influenced by several factors like the correlation between marker allele frequencies and quantitative trait loci (QTL), number of phenotypic records, trait genetic architecture and heritability, minor allele frequency (MAF), marker density and population genetic structure (Garrick et al., 2009; Goddard and Hayes, 2009; Daetwyler et al., 2010; de los Campos et al., 2013; Zhang et al., 2019).

The number of chromosome segments (M_e) segregating independently is a population specific parameter and is associated with the kinship between individuals and population effective size (Goddard et al., 2011). Populations with small effective size (N_e) displayed lower number of M_e segregating independently, increasing the level of linkage disequilibrium (LD) between genetic markers and QTL and the proportion of the additive variance explained by the genetic markers and consequently, the prediction ability of genomic information (Goddard, 2009;

Meuwissen et al., 2013; Pocrnic et al., 2016). In this sense, the number of markers necessary to capture the additive genetic variance is proportional to population N_e (Lee et al., 2017). Most of the farm animals come from common ascendants; thus, a small N_e is expected (Misztal et al., 2020). Therefore, the use of moderate to low-density marker arrays would be sufficient to capture the additive genetic variance and predict accurately the genomic breeding values (GEBV) (Rolf et al., 2010; Weigel et al., 2010; Boichard et al., 2012; Hayes et al., 2012).

In the last years, several genomic companies have developed different low to moderate-density SNPs arrays for taurine and indicine cattle to improve the imputation process (Aliloo et al., 2018) and genomic prediction reliabilities (Aliloo et al., 2018; Wu et al., 2016). The customization of low-density SNPs arrays is specific by company and prioritize the selection of markers with high effect for productive traits to improve the genomic prediction reliability (Zhu et al., 2017). Notwithstanding, several different low-density SNPs arrays were developed with low number of common SNPs, which sometimes hinders and limits the reliability of imputation strategies from low density SNPs arrays to medium or high-density arrays (Meuwissen et al., 2016). In this sense, the accuracy of imputation was adequate to impute above 15K markers in Nellore cattle, expecting that the GEBVs accuracy from imputed genotypes would be similar to that obtained with non-imputed genotypes with high-density SNP arrays (Carvalheiro et al., 2014).

The genotype imputation uses different techniques and algorithms enabling to estimate a dense genotype from low dense genotype (Dassonneville et al., 2012). Basically, the most important parameters for genotype imputation are the LD information between IBD regions, number of animals in the reference population, the number of markers to be imputed and the rare allele frequency (Wang et al., 2016). Nevertheless, the genomic imputation displayed mistakes with structured patterns based on the ambiguity of algorithm used to imputation, the frequency and the effect of the haplotype on the trait (Pimentel et al., 2015), leading to underestimation or overestimation of genomic values altering their accuracy (Goddard., 2017).

Normally, a small proportion of population is genotyped in beef cattle and there are scarce genotypes of proven sires to in the training population, and the single-step genomic best linear unbiased prediction (ssGBLUP) is a suitable procedure (Silva et al., 2016) for dealing with this scenario. The ssGBLUP combines information from genotyped and non-genotyped animals in a unique analysis (Aguilar et al., 2010), joining pedigree, phenotypic records, and genomic information to predict the GEBV (Legarra et al., 2009). The performance of ssGBLUP method depends on several parameters to control and fit the biased predictions mainly the compatibility between genomic matrix and traditional relationship matrix (Nordbø et al., 2019). Genotyping errors due to imputation and pedigree errors, missing pedigree information and sample misidentification or identification mistakes (Pimentel et al., 2015; Wang et al., 2017) would influence the compatibility between genomic matrix and traditional relationship matrix (Meuwissen et al., 2015).

To make feasible the genomic selection at commercial scale, the use of imputed genotypes from different SNPs arrays to perform genomic predictions is a common practice in livestock programs (Moghaddar et al., 2015). However, the imputation process could incur in genotyping errors or missing data that bias genomic predictions, mainly for genomic models based on the genomic relationship matrix (Nordbø et al., 2019). Thus, errors or missing data close to 1% of SNPs in the imputed genotypes affect the levels of bias and consequently the genomic predictions (Nordbø et al., 2019). Few studies have evaluated the impact of SNP arrays density and SNP informativeness on imputation accuracy and genotype imputation error (He et al., 2014; Wang et al., 2020) and their influence on GEBVs obtained through the ssGBLUP using imputed genotypes. In this sense, Lopez et al. (2020) applied the ssGBLUP method in Hanwoo cattle and reported that the prediction accuracies with high-density SNP arrays (Bovine HD) imputed genotypes from 50k SNPs were slightly higher than those obtained with 50k SNPs non-imputed genotypes. The objective of this study was to evaluate the impact of SNP marker density and informativeness on prediction ability and imputation

performance using the single-step genomic BLUP in a simulated beef cattle population.

MATERIAL AND METHODS

Pedigree and phenotypic data simulation

The pedigree, phenotypes and genotypes were simulated using the software QMSim version 1.10 (Sargolzaei et al., 2009). Two traits with low ($h^2 = 0.12$) and high heritability ($h^2 = 0.42$) were simulated. The phenotypic variance was assumed to be 1.0, and the result obtained for each trait, was the average of eight replicates that were performed.

A historical population was created from generation zero to generation 2,020. The different levels of linkage disequilibrium (LD) were created using an equal number of animals per generation (2,000 animals) until generation 1,000. To generate the “bottleneck effect” the number of animals were reduced gradually from 2,000 to 600 producing consequently, genetic drift ad LD from generation 1,001 to 2,020.

The expansion of population was created selecting 200 animals (males and females equally distributed) from the 600 animals of the last generation of historical population, and the simulation of effective size was based on the real population (Brito et al., 2011). To enlarge the population, were considered the absence of selection, average of five progeny per dam, exponential growth of the number of dams, and the random union of gametes. After this process, 240 males and 6,000 females were randomly selected from the last generation including founder animals from the selection populations. This population was spanned over 8 generations and the selected males and females from each generation were randomly mated, generating a single progeny with equal probability of being a male or a female. The replacement rate of sires and dams were 20% and 60% respectively and kept constant for all generations.

Genome simulation

The genome length was based on Base_4.6.1 (Snelling et al., 2007) with a total of 2333 cM. 29 *Bos Taurus* autosomes (BTA) were simulated with 735,293 markers and 7,000 QTLs randomly distributed and was assumed that the genetic variance was explained entirely by QTLs. The number of markers and QTLs per chromosome ranged from 12,931 to 46,495 and from 121 to 438, respectively. All markers were bi-allelic, mimicking SNPs presents in the bovine commercial panels. For QTLs, the number of alleles per loci randomly ranged from two to four. Minor allele frequencies (MAF) were assumed equal for marker and QTLs alleles. The QTLs allele effects were sampled from a gamma distribution with a shape parameter equal to 0.4 (Hayes and Goddard, 2001). A recurrent mutation rate of $1x10^{-5}$ for both markers and QTLs was considered in the historical population. A total of 335,000 markers whit MAF>= 0.02 and 1000 QTLs were randomly selected. The phenotypes were generated by the sum of the QTLs additive effects and random residuals with normal distribution of zero mean and variance equal to 0.48.

ssGBLUP method

The traditional genetic evaluation was performed using pedigree and phenotypic information. The model can be represented as follows:

$$\mathbf{y = Xb + Zu + e}$$

where ***y*** is the vector of phenotype, ***b*** is the vector of fixed effects, ***u*** is the vector of additive genetic effects, ***X*** e ***Z*** are incidence matrices and ***e*** is the vector of random residuals. Considering an infinitesimal model, ***var*** (***u***) $= A\sigma_u^2$, where ***A*** is the numerator relationship matrix obtained from pedigree information and σ_u^2 is the variance of genetic effect. In the single-step genomic BLUP (ssGBLUP) proposed by Misztal et al. (2009), the inverse of the numerator relationship matrix **(A^{-1})** was replaced by H^{-1} that combines pedigree and genomic information. The H^{-1} was constructed according to Aguilar et al. (2010), as follows:

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

where $\mathbf{H^{-1}}$ is the inverse of the realized relationship matrix that incorporates the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$) and the inverse of the numerator relationship matrix of genotyped animals $\mathbf{A}_{\mathbf{22}}^{-\mathbf{1}}$ and the G matrix was created according to VanRaden, (2008):

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

where **M** is a matrix of marker alleles with **m** columns (m = total number of markers) and **n** rows (n = total number of genotyped individuals), and **P** is a matrix containing the frequency of the second allele (**pj**), expressed as **2pj**. **Mij** was 0 if the genotype of individual **i** for SNP **j** was homozygous for the first allele, was 1 if heterozygous, or 2 if the genotype was homozygous for the second allele. The variance component estimation and solutions were obtained by BLUPF90 family programs (Aguilar et al., 2014; Misztal et al., 2002).

Customized marker arrays, imputation and imputation accuracy

The impact of marker density array on imputation and genomic prediction was evaluated using customized SNPs arrays obtained from a HD panel defined by the total of simulated markers. In these sense, four SNP densities were customized ~35k, ~16k, ~4k and ~2k SNPs through windows of 10, 20, 80 and 160 SNPs by chromosome, respectively. Also, three marker selection criteria were used to pick up the SNPs within windows: 1) select informative markers with MAF values close to 0.5; 2) select less informative markers with the lowest MAF values; 3) select markers evenly distributed. Thus, the HD array with 335K SNPs and twelve scenarios of customized SNPs arrays with eight replicates, were evaluated. The reference population for imputation consisted of animals from the 6th and 7th generation (12,000) that were genotyped with the HD array. To evaluate imputation genotyped error on genomic predictions, the animals from 8th generation (6,000) were used as validation population and their genotype was imputed from

customized arrays to HD array using the FIMPUTE version 2.2 software (Sargolzaei et al., 2014). The impact of accuracy imputation was evaluated using four statistics parameters: 1) the correlation between true and imputed genotypes (R^2); 2) the correlation between true and imputed genotypes adjusted by the respective SNP allele frequency of each genotype (R^2 adjust) (Berry et al., 2017; Mulder et al., 2012); 3) the genotype concordance rate (C-R), designated as the mean proportions of correctly imputed genotypes within individuals (Berry et al., 2013) and 4) the imputed allelic error (A-E), defined as the percentage of wrongly imputed alleles per animal. All parameters were calculated through the IMPUTACC software (Aguilar, 2013).

Prediction ability, inflation and bias

The 18,000 genotyped animals from 6th, 7th and 8th generations were divided in reference group (12,000 animals), including animals from 6th and 7th generations, and the validation group (6,000 animals) included animals from 8th generation, without progeny and phenotype records. For each scenario, the prediction ability, bias and inflation were calculated in the validation set. The prediction ability was estimated as the Pearson's correlation between the true breeding value (TBV) and genomic estimated breeding value (GEBV). Bias was measured as the difference between predicted and simulated breeding values (Vitezica et al., 2011). The coefficient of regression (b) of TBV on GEBV was used as measure of inflation of the prediction method, where b=1 denotes no inflation.

RESULTS AND DISCUSSION

Prediction ability

The prediction ability, inflation and bias for validation animals obtained with the customized SNPs arrays are shown in Table 1. For both simulated traits, the prediction ability improved as the marker density array increased. The H-I and E-D

criteria produced almost the same prediction ability with 16k and 35k density arrays, nonetheless, for density arrays below 16k SNPs lower prediction ability was observed. The L-I criteria displayed a higher decrease of prediction ability with density arrays below 35k SNPs. Similarly, Barjasteh et al. (2020), in a simulated population with high relationship between training and validation population using three density arrays (5k, 50k and 777k SNPs), reported that the prediction accuracy increased as the density array rise for two traits with different heritability (0.25 and 0.50). In Angus cattle, Rolf et al. (2010) recommended arrays with densities varying between 2,500 to 10,000 SNPs markers extracted from the BovineSNP50 array to obtain reliable genomic predictions for feed efficiency related traits. Further, Salvian et al. (2020) using the ssGBLUP method obtained similar genomic predictions for growth traits in broilers when applied a density of markers above 5% of available markers in the HD array (600k SNPs). In Chinese Simental cattle, Zhu et al. (2017), reported higher prediction ability for growth traits using a MAF ranging from 0.01 to 0.1 for customized SNPs arrays.

Table 1. Mean prediction ability, inflation and bias for low and high heritability traits by SNP density array and SNP informativeness using non-imputed markers.

		Array Densities														
Trait	**Criteria**	**2k**	**4k**	**16k**	**35k**	**HD**	**2k**	**4k**	**16k**	**35k**	**HD**	**2k**	**4k**	**16k**	**35k**	**HD**
		$r_{(GEBV.TBV)}$					$b_{TBV/GEBV}$					$d_{GEBV.TBV}$				
	H-I	0.61±0.02	0.67±0.01	0.72±0.01	0.73±0.02		0.86±0.03	0.92±0.02	0.98±0.03	0.99±0.03		0.02±0.08	0.06±0.08	0.08±0.08	0.09±0.08	
h^2:0.42	**L-I**	0.38±0.02	0.51±0.02	0.68±0.01	0.72±0.01	0.74±0.01	0.66±0.04	0.77±0.03	0.94±0.03	0.97±0.03	0.99±0.02	0.04±0.08	0.02±0.08	0.09±0.08	0.09±0.08	0.10±0.08
	E-D	0.60±0.02	0.67±0.01	0.73±0.01	0.73±0.01		0.85±0.03	0.92±0.02	0.98±0.03	0.99±0.03		0.02±0.08	0.06±0.08	0.09±0.08	0.09±0.08	
	H-I	0.47±0.03	0.51±0.03	0.56±0.02	0.57±0.02		0.84±0.03	0.93±0.03	1.05±0.04	1.02±0.05		0.03±0.05	0.04±0.05	0.05±0.05	0.05±0.05	
h^2:0.12	**L-I**	0.16±0.04	0.22±0.03	0.52±0.04	0.55±0.04	0.57±0.03	0.51±0.07	0.80±0.02	0.97±0.04	1.04±0.04	1.07±0.04	0.05±0.06	0.04±0.06	0.06±0.05	0.06±0.05	0.06±0.05
	E-D	0.47±0.03	0.51±0.03	0.56±0.3	0.57±0.03		0.82±0.03	0.92±0.04	1.05±0.06	1.03±0.05		0.02±0.07	0.03±0.07	0.05±0.05	0.05±0.05	

GEBV= Genomic Breeding Value estimate; **TBV=** True Breeding Value; **h^2**= heritability of trait; **$r_{(GEBV.TBV)=}$** correlation between GEBV and TBV; **$b_{TBV/GEBV}$ =** regression coefficient of TBV on GEBV; **$d_{(GEBV,TBV)}$ =** difference between GEBV and TBV; **H-I=** selected markers with high informativeness; **L-I=** selected markers with less informativeness; **E-D=** markers evenly distributed.

The L-I criterion to select the SNPs showed lower prediction ability than H-I and E-D criteria when the SNP density was below 35k. The results obtained in this study suggested that the reduction of customized marker density had higher impact on prediction ability when less informative markers were used compared to more informative markers or evenly spaced criteria. Chen et al. (2011) applied the ssGBLUP method varying the MAF in broilers for body weight at 6 weeks and area of breast meat traits and reported that a MAF threshold close to 0.4 was necessary to maximize the prediction ability for the studied traits. The authors also reported that variations associated with genetic structure and properties of G matrix were affected by the MAF threshold. In this sense, Pocrnic et al. (2016), in a simulated population reported that the breeding values could be a linear function of effective SNP markers when capture a higher proportion of genetic variance, allowing shape a genomic matrix with a reliable and ideal dimensionality for predictions.

Inflation and bias

For both simulated traits, the prediction inflation increased as the customized density arrays decreased (Table 1). The 4k and 2k SNP marker density arrays displayed the highest inflated genomic predictions. The H-I and E-D criteria presented similar inflation and lower dispersion with marker density arrays above 16k SNPs, while the L-I criteria displayed more inflated predictions for SNP density arrays below 16k. The inflated predictions could be related to the low additive genetic variance captured by the SNP markers due to the low levels of LD in the low-density arrays (Ma et al., 2015). This study showed that the incorporation of approximately 5% of available SNPs from HD array displayed similar inflation of predictions when compared to HD. Salvian et al. (2020) in broilers reported alike results and obtained less inflated predictions when used arrays close to 5% of available SNPs from HD array (600k SNPs).

The prediction bias decreased as the SNP density array diminished, however with 2k array and L-I criteria a slightly increase of bias was observed for both traits (Table 1). The criteria to pick up the SNP and density arrays influenced the prediction bias for both traits. The inclusion of low-density arrays overestimated

the GEBVs due to mistakes of kindship, increasing the similarity between individuals given by false identical segments by descending detected (Wang, 2014).

Accuracy of imputation

The imputation accuracies from customized SNP arrays to HD array are shown in Table 2. As expected, the imputation accuracy improved as the density of marker arrays to be imputed increased, due to the number of markers to be imputed diminished (Mulder et al., 2012). With the H-I and E-D criteria, the correlations parameters (R^2 and R^2adj) and C-R (mean allele concordance rate) parameter of imputed genotypes displayed high and almost the same values favoring the accuracy and quality of imputation when the SNP density arrays were imputed above 16k, while the A-E (mean of proportion to allelic error imputed) parameter displayed a lower difference between H-I and E-D criteria, favoring the quality of imputation of H-I criteria with SNP density arrays to impute higher to 16k. On the other hand, the L-I criteria with SNP density arrays to be imputed below 16k SNPs displayed lower accuracy and quality of imputation evaluated trough the R^2, R^2adj and C-R parameters, respectively. However, the A-E parameter displayed high values of imputed allelic error being higher for SNP density arrays to impute below 35k. Carvalheiro et al. (2014) worked with Nellore cattle and evaluated the imputation performance using the FIMPUTE and BEAGLE software for low and medium SNPs arrays densities based on MAF, LD and distance between markers, and reported high imputation accuracy (0.97) and high percentage of correctly imputed genotypes (99.1%) from low-density arrays (above 7k SNPs) to high density arrays (Bovine HD).

In this sense, Kranjčevičová et al. (2019), evaluated the imputation accuracy in cattle from low density arrays to Ilumina BovineSNP50v,2 BeadChip (50k), using customized arrays of 15, 30, 55, 70 and 95% of randomly selected SNPs markers from the 50k array, and reported that the imputation accuracy was almost same deleting until 30% of SNPs. In sheep, O'Brien et al. (2019) selected informative

SNPs based on position, MAF and LD to customize lower-density arrays (0.384k, 1k, 2k, 3k, 6k and 9k) to impute for medium-density array (50k). The authors reported variations in allele concordance rate associated with the different breeds used and showed that 6k SNPs arrays to impute to 50k with markers selected by MAF, LD and evenly distributed attained a reliable imputation (above 0.98). In this way, in beef cattle populations, Judge et al. (2016), evaluated the imputation accuracy from different low-density arrays customized by MAF, LD and evenly distribution of SNPs markers and reported that SNP density arrays of 3k and 6k SNPs achieved an imputation accuracy of 0.90 and 0.95, respectively. Likewise, Mulder et al. (2012) worked with Holstein cattle and customized three low-density SNPs arrays (0.3k, 3k and 6k SNPs) from available markers of 50k SNPs array, selecting markers either by high MAF or by position adjacent to midpoint of each window designed and reported imputation accuracies varying from 0.81 to 0.99. According with this study, the higher imputation accuracy increased as the number of SNPs to be imputed decreased, moreover, the criteria used to select the SNPs markers displayed almost the same results, with the highest imputation accuracy for H-I and E-D criteria while the L-I criteria displayed the highest imputation error.

Table 2. Accuracy to high-density marker imputation and standard errors from different low-density chips customized by informativeness and distribution of markers using four statistics parameters.

Criteria	Imputed	R^2	R^2_{adj}	C-R	A-E
H-I	**35k-HD**	0.99±0.00004	0.99±0.0001	0.99±0.00003	0.02±0.0015
	16k-HD	0.99±0.0002	0.99±0.0003	0.99±0.0001	0.05±0.005
	4k-HD	0.99±0.0004	0.98±0.0007	0.99±0.0002	0.30±0.010
	2k-HD	0.97±0.001	0.95±0.003	0.98±0.001	0.82±0.048
L-I	**35k-HD**	0.99±0.0002	0.99±0.001	0.99±0.0001	0.19±0.007
	16k-HD	0.97±0.001	0.95±0.002	0.98±0.0007	0.86±0.03
	4k-HD	0.56±0.01	0.27±0.01	0.71±0.005	14.97±0.282
	2k-HD	0.38±0.01	0.08±0.004	0.61±0.004	21.61±0.230
E-D	**35k-HD**	0.99±0.00004	0.99±0.0001	0.99±0.00002	0.02±0.001
	16k-HD	0.99±0.0001	0.99±0.0002	0.99±0.0001	0.06±0.004
	4k-HD	0.98±0.001	0.97±0.001	0.99±0.0003	0.47±0.02
	2k-HD	0.96±0.003	0.92±0.01	0.97±0.02	1.32±0.10

R^2= mean of correlation between true and imputed genotype; **R^2-adj=** mean adjusted genotype correlation between true and imputed genotype; **C-R=** mean allele concordance rate; **A-E (%) =** mean of proportion to allelic error imputed; **H-I=** selected markers with high informativeness; **L-I=** selected markers with less informativeness; **E-D=** markers evenly distributed.

Prediction ability of imputed arrays

The prediction ability with the customized low-densities SNPs arrays imputed to HD in the validation subset are shown in Table 3. For both simulated traits, similar prediction ability of imputed HD genotyped was observed between the H-I and E-D criteria of SNP selection. However, the prediction ability obtained with the L-I criteria was lower when low density arrays (4k and 2k arrays) were imputed. Similar prediction abilities to non-imputed HD array were obtained with the imputed genotypes when at least 5% of SNP markers available from HD and using the H-I or E-D criteria was applied. In Hanwoo cattle, Lopez et al. (2020) used the ssGBLUP for single and multi-trait genetic models with marker density arrays of 50k and imputed to HD from 50K SNPs and reported prediction accuracy differences of 0.6 to 2% between non-imputed and imputed arrays, as observed in this study. In this way, Mulder et al. (2012) displayed that the similarity between accuracy of predictions of non-imputed (50k) with imputed genotypes (0.3k, 3k and 6k SNPs) increased with the decrease in the number of markers to be imputed.

Table 3. Mean prediction ability, inflation and bias for low and high heritability traits by SNP density array and SNP informativeness using imputed markers.

		Array Imputed														
Trait	**Criteria**	**2k-HD**	**4k-HD**	**16k-HD**	**35k-HD**	**HD**	**2k-HD**	**4k-HD**	**16k-HD**	**35k-HD**	**HD**	**2k-HD**	**4k-HD**	**16k-HD**	**35k-HD**	**HD**
		$r_{(GEBV.TBV)}$					$b_{TBV/GEBV}$					$d_{GEBV.TBV}$				
	H-I	0.73±0.01	0.74±0.01	0.74±0.01	0.74±0.01		1.00±0.03	1.01±0.03	1.01±0.03	1.01±0.03		0.11±0.09	0.10±0.09	0.10±0.08	0.10±0.08	
h2:0.42	**L-I**	0.30±0.02	0.45±0.03	0.73±0.01	0.74±0.01	0.74±0.01	0.51±0.04	0.67±0.05	1.00±0.03	1.01±0.03	0.99±0.02	0.18±0.07	0.21±0.07	0.13±0.08	0.10±0.08	0.10±0.08
	E-D	0.72±0.01	0.74±0.01	0.74±0.01	0.74±0.01		0.99±0.03	1.01±0.03	1.01±0.03	1.01±0.03		0.12±0.08	0.11±0.08	0.10±0.08	0.10±0.08	
	H-I	0,56±0.03	0,57±0.03	0,57±0.03	0.57±0.03		1.06±0.04	1.07±0.04	1.07±0.04	1.07±0.04		0.05±0.05	0.05±0.05	0.06±0.05	0.06±0.05	
h2:0.12	**L-I**	0.17±0.01	0.31±0.03	0.56±0.02	0.57±0.03	0.57±0.03	0.45±0.01	0.71±0.06	1.05±0.04	1.07±0.04	1.07±0.04	0.07±0.06	0.09±0.06	0.07±0.06	0.06±0.06	0.06±0.05
	E-D	0.56±0.02	0.57±0.02	0.57±0.03	0.57±0.03		1.04±0.05	1.06±0.05	1.07±0.04	1,07±0.04		0.06±0.06	0.06±0.06	0.06±0.05	0.06±0.05	

GEBV= Genomic Breeding Value estimate; **TBV=** True Breeding Value; h^2= heritability of trait; **35k-HD** = imputed marker from 35K; **16k-HD** = imputed marker from 16K, **4k-HD** = imputed marker from 4K; **2k-HD** = imputed marker from 2K; $r_{(GEBV.TBV)=}$ correlation between GEBV and TBV; $b_{TBV/GEBV}$ **=** regression coefficient of TBV on GEBV; $d_{(GEBV,TBV)}$ **=** difference between GEBV and TBV; **H-I=** selected markers with high informativeness; **L-I=** selected markers with less informativeness; **E-D=** markers evenly distributed.

Inflation and bias using imputed arrays

The inflation and bias with the different customized low-densities SNPs arrays imputed to HD array are shown in Table 3. Slightly differences in prediction inflation between imputed arrays were detected. For both simulated traits, the L-I criteria for imputed SNP density arrays below 16k displayed the highest inflation. However, the imputed arrays from different scenarios showed lower and similar inflation of predictions than non-imputed HD arrays, being the high heritability trait the less inflated. Similar with current study, Mulder et al. (2012), for three customized arrays (0.3k, 3k and 6k SNPs) reported inflated predictions (0.97, 0.98 and 0.98) as the number of SNPs to be imputed decreased, and the criteria to selected SNPs did not affect the prediction inflation. This fact, support the overdispersion of the imputed genotype additive variance and consequently the overestimation of GEBVs. Likewise, Lopez et al. (2020) reported deflated predictions in single-trait models imputed genotypes from 50K to HD (777,962 SNPs) in commercial Hanwoo cattle.

For both simulated traits, almost the same bias of imputed SNP arrays was observed. The H-I and E-D criteria displayed slightly differences between imputed customized SNP arrays. However, the bias achieved for L-I criteria was higher when low-density arrays below 35k SNPs were used for imputation to HD arrays. Mulder et al. (2012), associated the bias with the genetic variance captured by the imputed arrays which are conditionate to errors of allele frequencies imputed. Thus, the current study support that the prediction bias using imputed genotypes was associated with the informativeness and marker density array due to genotype errors originated during the imputation process, mainly for with L-I criteria.

Imputation is an attractive cost-saving to implementing genomic selection in livestock, however, combining different commercial SNPs arrays with a varying number of common SNPs is a challenge to attain a reliable imputation together with a low error rate. Additionally, imputation to higher density panels in some cattle populations, i.e., indicine cattle, is not sufficiently accurate using commercially available low-density panels (Bernardes et al., 2019; Carvalheiro et

al., 2014). It is important to highlight that imputation errors inflated Mendelian errors, and it can have some side effects on subsequent genetic analyses (IBD analysis) by affecting the estimation of the Mendelian sampling variation and the compatibility between pedigree and genomic information in the ssGBLUP method. The results of this study supported that a minimum number highly informative or evenly distributed SNPs are necessary to obtained accurately imputed genotypes and reliable genomic predictions using ssGBLUP with imputed or non-imputed genotypes.

CONCLUSION

The customized arrays including at least 5% of highly informative or evenly distributed SNPs available in the HD array are necessary to achieve similar genomic predictions than those obtained with the HD array. Therefore, the development of specific moderate and low-density customized arrays based on criteria of high informativeness and evenly distribution of SNPs, might be cost-effective to perform genomic predictions and approaches that involve genotype imputation in beef cattle breeding programs.

ACKNOWLEDGEMENTS

The authors thank National Association of Breeders and Researchers (ANCP), the Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias (FCAV/Unesp), the Universidad de la Republica, Facultad de Veterinaria (UdelaR), Departamento de Genética y Mejoramiento Animal and the Instituto Nacional de Investigación Agropecuaria of Uruguay (INIA).

FUNDING

This study was supported by the Programa Estudantes Convênio de Pós-Graduação da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (PECPG-CAPES : 88881.154576/2017-01).

REFERENCES

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., 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. https://doi.org/10.3168/jds.2009-2730

Aguilar, Ignacio, Misztal, Ignacy, Tsuruta, Shogo, Legarra, Andres, Wang, Huiyu, Aguilar, I, Misztal, I, Tsuruta, S, Legarra, A, Wang, H, 2014. PREGSF90-POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs.

Aliloo, H., Mrode, R., Okeyo, A.M., Ni, G., Goddard, M.E., Gibson, J.P., 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. Journal of Dairy Science 101, 9108–9127. https://doi.org/10.3168/jds.2018-14621

Aliloo, Hassan, Mrode, R., Okeyo, M., Ojango, J., Dessie, T., Rege, E., Goddard, M., Gibson, J., 2018. Optimal design of low-density marker panels for genotype imputation. Proceedings of the World Congress on Genetics Applied to Livestock Production 11, 146.

Barjasteh, S., Dashab, G.R., Rokouei, M., Shariati, M.M., Vafaye Valleh, M., 2020. Comparing Different Marker Densities and Various Reference Populations Using Pedigree-Marker Best Linear Unbiased Prediction (BLUP) Model, Iranian Journal of Applied Animal Science. Islamic Azad University - Rasht Branch.

Bernardes, P.A., Nascimento, G.B. do, Savegnago, R.P., Buzanskas, M.E., Watanabe, R.N., de Almeida Regitano, L.C., Coutinho, L.L., Gondro, C., Munari, D.P., 2019. Evaluation of imputation accuracy using the combination of two high-density panels in Nelore beef cattle. Scientific Reports 9, 1–10. https://doi.org/10.1038/s41598-019-54382-w

Berry, D.P., Mcclure, M.C., Mullen, M.P., 2013. Within-and across-breed imputation of high-density genotypes in dairy and beef cattle from medium-and low-density genotypes. https://doi.org/10.1111/jbg.12067

Berry, D.P., Mchugh, N., Randles, S., Wall, E., Mcdermott, K., Sargolzaei, M., O'brien, A.C., 2017. Imputation of non-genotyped sheep from the genotypes of their mates and resulting progeny. https://doi.org/10.1017/S1751731117001653

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., 2012. Design of a bovine low-density snp array optimized for imputation. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0034130

Brito, F. v, Neto, J.B., Sargolzaei, M., Cobuci, J.A., Schenkel, F.S., 2011. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle.

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science 89, 2673–2679. https://doi.org/10.2527/jas.2010-3555

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185, 1021–1031. https://doi.org/10.1534/genetics.110.116855

Dassonneville, R., Fritz, S., Ducrocq, V., Boichard, D., 2012. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. https://doi.org/10.3168/jds.2011-5133

de Lima, L.G., de Souza, N.O.B., Rios, R.R., de Melo, B.A., dos Santos, L.T.A., Silva, K. de M., Murphy, T.W., Fraga, A.B., 2020. Advances in molecular genetic techniques applied to selection for litter size in goats (Capra hircus): a review. Journal of Applied Animal Research. https://doi.org/10.1080/09712119.2020.1717497

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. https://doi.org/10.1534/genetics.112.143313

Garrick, D.J., Taylor, J.F., Fernando, R.L., 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution 2009 41:1 41, 1–8. https://doi.org/10.1186/1297-9686-41-55

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257. https://doi.org/10.1007/s10709-008-9308-0

Goddard., M.E., 2017. Can we make genomic selection 100% accurate?, Journal of Animal Breeding and Genetics. https://doi.org/10.1111/jbg.12306

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381–391. https://doi.org/10.1038/nrg2575

Goddard, M.E., Hayes, B.J., Meuwissen, T.H.E., 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128, 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Hayes, B., Goddard, M.E., 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol 33, 209–229.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D., Kijas, J.W., van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Animal Genetics 43, 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x

He, S., Wang, S., Fu, W., Ding, X., Zhang, Q., 2014. Imputation of missing genotypes from low-to high-density SNP panel in different population designs. Animal genetics 46, 1–7. https://doi.org/10.1111/age.12236

Judge, M.M., Kearney, J.F., McClure, M.C., Sleator, R.D., Berry, D.P., 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations. Journal of Animal Science 94, 949–962. https://doi.org/10.2527/jas.2015-0044

Kranjčevičová, A., Kašná, E., Brzáková, M., Přibyl, J., Vostrý, L., 2019. Impact of reference population size and marker density on accuracy of population imputation. Czech Journal of Animal Science 64, 405–410. https://doi.org/10.17221/148/2019-CJAS

Lee, S.H., Clark, S., van der Werf, J.H.J., 2017. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0189775

Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–4663. https://doi.org/10.3168/jds.2009-2061

Liu, Y., Xu, Lei, Wang, Z., Xu, Ling, Chen, Y., Zhang, L., Xu, Lingyang, Gao, X., Gao, H., Zhu, B., Li, J., 2019. Genomic prediction and association analysis with models

including dominance effects for important traits in Chinese simmental beef cattle. Animals 9. https://doi.org/10.3390/ani9121055

Lopez, B.I., Lee, S.H., Shin, D.H., Oh, J.D., Chai, H.H., Park, W., Park, J.E., Lim, D., 2020. Accuracy of genomic evaluation using imputed high-density genotypes for carcass traits in commercial Hanwoo population. Livestock Science 241, 104256. https://doi.org/10.1016/j.livsci.2020.104256

Ma, P., Lund, M.S., Nielsen, U.S., Aamand, G.P., Su, G., 2015. Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. Journal of Dairy Science 98, 9026–9034. https://doi.org/10.3168/jds.2015-9703

Meuwissen, T., Hayes, B., Goddard, M., 2016. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers 6, 6–14. https://doi.org/10.2527/af.2016-0002

Meuwissen, T., Hayes, B., Goddard, M., 2013. Accelerating improvement of livestock with genomic selection. Annual Review of Animal Biosciences 1, 221–237. https://doi.org/10.1146/annurev-animal-031412-103705

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157.

Meuwissen, T.H.E., Svendsen, M., Solberg, T., Ødegård, J., 2015. Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian Red cattle. Genet Sel Evol 47, 79. https://doi.org/10.1186/s12711-015-0159-8

Misztal, I., Lourenco, D., Legarra, A., 2020. Current status of genomic evaluation. Journal of Animal Science 98, 1–14. https://doi.org/10.1093/jas/skaa101

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D.H., 2002. BLUPF90 AND RELATED PROGRAMS (BGF90).

Moghaddar, N., Gore, K.P., Daetwyler, H.D., Hayes, B.J., van der Werf, J.H.J., 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genetics Selection Evolution 47, 97. https://doi.org/10.1186/s12711-015-0175-8

Mulder, H.A., Calus, M.P.L., Druet, T., Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science 95, 876–889. https://doi.org/10.3168/JDS.2011-4490

Nordbø, Ø., Gjuvsland, A.B., Eikje, L.S., Meuwissen, T., 2019. Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. Genetics Selection Evolution 51, 1–8. https://doi.org/10.1186/s12711-019-0517-z

O'Brien, A.C., Judge, M.M., Fair, S., Berry, D.P., 2019. High imputation accuracy from informative low-to-medium density single nucleotide polymorphism genotypes is achievable in sheep. Journal of Animal Science 97, 1550–1567. https://doi.org/10.1093/jas/skz043

Pimentel, E.C.G., Edel, C., Emmerling, R., Götz, K.U., 2015. How imputation errors bias genomic predictions. Journal of Dairy Science 98, 4131–4138. https://doi.org/10.3168/jds.2014-9170

Pocrnic, I., Lourenco, D.A.L., Masuda, Y., Legarra, A., Misztal, I., 2016. The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics 203, 573. https://doi.org/10.1534/GENETICS.116.187013

Rolf, M.M., Taylor, J.F., Schnabel, R.D., McKay, S.D., McClure, M.C., Northcutt, S.L., Kerley, M.S., Weaber, R.L., 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genetics 11. https://doi.org/10.1186/1471-2156-11-24

Salvian, M., Costa, G., Moreira, M., Spangler, M.L., Mourão, G.B., 2020. Estimation of Breeding Values Using Different Densities of Snp to Inform Kinship in Broiler Chickens. https://doi.org/10.21203/rs.3.rs-32429/v1

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15. https://doi.org/10.1186/1471-2164-15-478

Sargolzaei, M., Schenkel, F.S., Bateman, A., 2009. QMSim: a large-scale genome simulator for livestock. BIOINFORMATICS APPLICATIONS NOTE 25, 680–681. https://doi.org/10.1093/bioinformatics/btp045

Silva, R.M.O., Fragomeni, B.O., Lourenco, D.A.L., Magalhães, A.F.B., Irano, N., Carvalheiro, R., Canesin, R.C., Mercadante, M.E.Z., Boligon, A.A., Baldi, F.S., Misztal, I., Albuquerque, L.G., 2016. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. Journal of Animal Science 94, 3613–3623. https://doi.org/10.2527/jas2016-0401

Silveira, L.S., Lima, L.P., Nascimento, M., Nascimento, A.C.C., Silva, F.F., 2020. Regression trees in genomic selection for carcass traits in pigs. Genetics and Molecular Research 19. https://doi.org/10.4238/gmr18498

Snelling, W.M., Chiu, R., Schein, J.E., Hobbs, M., Abbey, C.A., Adelson, D.L., Aerts, J., Bennett, G.L., Bosdet, I.E., Boussaha, M., Brauning, R., Caetano, A.R., Costa, M.M., Crawford, A.M., Dalrymple, B.P., Eggen, A., Wind, A.E. der, Floriot, S., Gautier, M., Gill, C.A., Green, R.D., Holt, R., Jann, O., Jones, S.J., Kappes, S.M., Keele, J.W., Jong, P.J. de, Larkin, D.M., Lewin, H.A., McEwan, J.C., McKay, S., Marra, M.A., Mathewson, C.A., Matukumalli, L.K., Moore, S.S., Murdoch, B., Nicholas, F.W., Osoegawa, K., Roy, A., Salih, H., Schibler, L., Schnabel, R.D., Silveri, L., Skow, L.C., Smith, T.P., Sonstegard, T.S., Taylor, J.F., Tellam, R., Tassell, C.P. van, Williams, J.L., Womack, J.E., Wye, N.H., Yang, G., Zhao, S., Consortium, the I.B.B.M., 2007. A physical map of the bovine genome. Genome Biology 8, R165. https://doi.org/10.1186/GB-2007-8-8-R165

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genetics Research 93, 357–366. https://doi.org/10.1017/S001667231100022X

Wang, J., 2014. Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. Journal of Evolutionary Biology 27, 518–530. https://doi.org/10.1111/jeb.12315

Wang, Q., Yu, Y., Li, F., Zhang, X., Xiang, J., 2017. Predictive ability of genomic selection models for breeding value estimation on growth traits of Pacific white shrimp Litopenaeus vannamei. Chinese Journal of Oceanology and Limnology 35, 1221–1229. https://doi.org/10.1007/s00343-017-6038-0

Wang, X., Su, G., Hao, D., Lund, M.S., Kadarmideen, H.N., 2020. Comparisons of improved genomic predictions generated by different imputation methods for genotyping by sequencing data in livestock populations. Journal of Animal Science and Biotechnology 2020 11:1 11, 1–12. https://doi.org/10.1186/S40104-019-0407-9

Wang, Y., Lin, G., Li, C., Stothard, P., 2016. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Science Reviews 4, 79–98. https://doi.org/10.1007/s40362-017-0041-x

Weigel, K.A., de los Campos, G., Vazquez, A.I., Rosa, G.J.M., Gianola, D., van Tassell, C.P., 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in JerseWeigel, K.A. et al., 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of. Journal of Dairy Science 93, 5423–5435. https://doi.org/10.3168/jds.2010-3149

Wu, X.L., Xu, J., Feng, G., Wiggans, G.R., Taylor, J.F., He, J., Qian, C., Qiu, J., Simpson, B., Walker, J., Bauck, S., 2016. Optimal design of low-density SNP

arrays for genomic prediction: Algorithm and applications. PLoS ONE 11, e0161719. https://doi.org/10.1371/journal.pone.0161719

Zhang, H., Yin, L., Wang, M., Yuan, X., Liu, X., 2019. Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. Frontiers in Genetics 0, 189. https://doi.org/10.3389/FGENE.2019.00189

Zhu, B., Zhang, J. jing, Niu, H., Guan, L., GUO, P., XU, L. yang, CHEN, Y., ZHANG, L. pei, GAO, H. jiang, GAO, X., LI, J. ya, 2017. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture 16, 911–920. https://doi.org/10.1016/S2095-3119(16)61474-0

CHAPTER 3 - Prediction ability for growth and maternal traits using SNP arrays based on different marker density in Nellore cattle using the ssGBLUP

ABSTRACT

This study aimed to investigate the prediction ability for growth and maternal traits using different low-density customized SNP arrays selected by informativeness and distribution of markers across the genome employing single-step genomic BLUP (ssGBLUP). Phenotypic records for adjusted weight at 210 days and 450 days of age of animals born between 1974 and 2018 were utilized. A total of 945 genotyped animals with high-density chip were used, and 267 individuals born after 2008 were selected as validation population. A total of 11 scenarios were evaluated, where high informativeness and evenly distributed criteria were used to selected SNPs. By criteria, five density arrays were customized (40k, 20k, 10k, 5k and 2k) and the HD array was used as desirable scenario. The GEBV predictions and accuracy BIF (beef improvement federation) were obtained with BLUPF90 family programs. The linear regression method was used to evaluate the prediction ability, inflation, and bias of GEBV of each customized array. An overestimation of partial GEBVs (GEBV_p) in contrast with complete GEBVs (GEBV_c) and increase of BIF accuracy with the density arrays diminished were observed. For all traits, the prediction ability was higher as the array density increased and it was similar with customized arrays higher than 10k SNPs. The level of inflation was lower as the density array increased and was higher for MW210 effect. The bias was susceptible to overestimation of GEBVs when the density customized arrays decreased. These results revealed that the BIF accuracy is sensible to overestimation using a low-density customized arrays while the prediction ability with least 5,000 informative SNPs obtained from the Illumina BovineHD BeadChip properly distributed and highly informative shows accurate and less biased predictions. Thus, this study reported that genomic selection with low-density customized arrays under ssGBLUP method could be feasible and cost-effective in Nellore beef cattle.

Keywords: Accuracy, beef cattle, genomic selection, inflation, minor allele frequency.

INTRODUCTION

The genomic selection is a form of marker-assisted selection proposed by Meuwissen et al. (2001), which assumes that all genomic segments contribute to the genetic variance, and each segment is in high linkage disequilibrium (LD) with at least one known genetic marker. Several methods have been postulated to predict the genomic breeding value (GEBV) (Habier et al., 2009; Erbe et al., 2012; Daetwyler et al., 2013; de los Campos et al., 2013; Chiaia et al., 2018;). Misztal et al. (2009) and Aguilar et al. (2010) postulated the single-step genomic Best Linear Unbiased Predictor (ssGBLUP) procedure by combining the traditional pedigree relationship matrix (**A**) and the genomic relationship matrix (**G**) into a new **H** matrix. The ssGBLUP procedure has been used in genomic evaluations in different livestock species, including beef and dairy cattle (Liu et al., 2019; Rezende et al., 2019), pigs (Waide et al., 2018; Silveira et al., 2020), goats (de Lima et al., 2020; Gipson, 2019), aquaculture (Wang et al., 2017; Garcia et al., 2018; Joshi et al., 2020) and, plants (Heffner et al., 2009; Resende et al., 2012; Sousa et al., 2019). The ssGBLUP has some advantages in which we can highlight its simplicity and lack of approximations, and it accounts for pre-selection bias of genotyped selected parents without phenotypes. Further, predictions of the GEBV do not rely on de-regressed proofs and accounts for the entire population structure (Legarra et al., 2014; Lourenco et al., 2014).

To predict the GEBV for complex traits, the highest the number of markers, the increased is the chance to pick-up more markers in LD with a quantitative trait locus (QTL), improving the prediction ability of the markers (Meuwissen et al.,

2001). However, the use of denser marker arrays does not guarantee better predictions since the additive genetic variance is captured by a limited number of markers (Goddard, 2009; Lourenco et al., 2020). In this regard, Lee et al. (2017) showed that the effective population size (N_e) is a limiting factor to assess the adequate number of markers to capture the maximum of the additive genetic variance. Populations with large N_e display a higher number of independent chromosome segments (M_e), and therefore, an increased number of markers are necessary. Misztal, (2016) proposed a theory to identify animals with high additive information through a linear relation of the animals with M_e, and subsequently, with the effective number of markers that explain the additive variation. In a simulated study, Pocrnic et al. (2016) demonstrated that the number of markers required to explain the variation of the genomic matrix increased as a function of the N_e. The authors concluded that the optimal dimension of the genomic matrix should include markers explaining 95% to 98% of the additive genetic variation to obtain reliable predictions with a less computational cost. Assuming that most of the cattle breeds were originated from a narrow genetic basis with current low N_e (Misztal et al., 2020), the use of high-density SNP arrays to predict GEBV is not crucial (Rolf et al., 2010; Weigel et al., 2010; Boichard et al., 2012; Hayes et al., 2012).

The customization of low-density SNPs arrays has been a topic of interest to reduce the genotyping costs as well as the computational requirements for the genomic evaluations (Habier et al., 2009; Wu et al., 2016). Barjasteh et al. (2020), working with several simulated cattle reference populations for traits with heritability varying from 0.25 to 0.50, demonstrated that the GEBV accuracy with medium-density SNPs arrays (50k) was like to that obtained using denser SNPs arrays (777k). (Salvian et al., 2020) working with broilers compared SNPs arrays with different densities in proportion to the HD panel (600k Affymetrix Axiom Genotyping Array) using the ssGBLUP method and showed that the accuracies were similar when 10% or 100% of markers were used. (Wu et al., 2018) evaluated the GEBV accuracies for growth and reproductive traits using the GeneSeek Profiler *Bos Indicus* low-density (~35k) and the Illumina BovineHD BeadChip

(777k) in Nellore cattle and reported slightly lower GEBV accuracies with the low-density array. Interestingly, Boddhireddy et al. (2014) evaluated the accuracy of the genomic prediction for reproductive, productive, and visual body conformation scores in Nellore cattle using the BovineHD Beadchip and 54k SNP arrays and reported that the prediction accuracy was not influenced by the array density.

It is well understood that *Bos taurus indicus* breeds need a different subset of informative SNPs than *Bos taurus taurus* breeds to adequately perform genomic selection in such breeds (Nayee et al., 2018). The Illumina BovineHD BeadChip opened opportunities to develop customized or *in-silico* SNPs arrays for *Bos taurus indicus* beef cattle breeds to be applied on genomic evaluations. Therefore, the knowledge of the population structure and the use of low-density SNP arrays that prioritize informative markers is an interesting strategy to obtain reliable and unbiased GEBV predictions with a lower cost and higher computational optimization (Wongpom et al.,2019). The minor allele frequency (MAF) is an important informative propriety of the population's background to predict the GEBV (Zhu et al., 2017), and the estimation of the genomic matrix allows the inclusion of more markers with high MAF (VanRaden, 2008). However, Yang et al. (2010) demonstrated that markers with low MAF would improve the GEBV prediction when the genetic structure presented QTL with low MAF. Therefore, the objective of this study was to evaluate the prediction ability for growth and maternal-related traits using SNPs arrays with different marker densities based on SNP informativeness employing the ssGBLUP methodology in a Nellore beef cattle population.

MATERIALS AND METHODS

Data

Phenotypic and genotypic records from the Nellore Brazil breeding program coordinated by the National Association of Breeders and Researchers (ANCP, Ribeirão Preto-SP, Brazil) were used. A total of 1,195,831 and 948,129 records for adjusted weight at 210 days (W210) and 450 days of age (W450), respectively, were used. Weight records were measured between 1974 and 2018. The animals were raised in pasture-based production systems, with or without the use of creep feeding and supplementation. The breeding season was from February to April and mid-November to January. Heifers were exposed to reproduction regardless of weight and body condition when attained 14 to 16 months of age. During the mating season, artificial insemination, controlled breeding, and multiple breeding systems were used, with a bull cow ratio of 1:30. Heifers were evaluated for pregnancy by rectal palpation roughly 60 days after the end of the breeding season, and those that did not conceive were exposed again at 24 months of age.

The phenotypic and genomic information of the population was divided into partial and complete datasets using information of animals born from 1974 to 2008 and 1974 to 2018, respectively. A total of 945 representative sires of the main Nellore lineages (i.e., Karvadi, Golias, Godhavari, Taj Mahal, Akasamu; and Nagpur) were genotyped with the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA, USA), which contains 777,962 markers. Quality control was performed on PreGSf90 program (Aguilar et al., 2010). Markers located on sexual chromosomes, with MAF lower than 0.05, and Hardy-Weinberg equilibrium with a p-value lower than 10^{-5} were removed from the dataset. Further, markers and animals were excluded for call rate lower than 0.90. After quality control, 427,495 markers (427k) and 914 animal samples were available for analyses. The descriptive statistic for the growth-related traits is presented in Table 1.

Table 1. Number of records and descriptive statistics for the complete and partial dataset for adjusted weight at 210 days (W210) and 450 days of age (W450).

Dataset	Genotypes	Trait	Offspring	Total records	Mean	SD
Partial	647	W210	84,510	523,935	181.49	31.14
		W450	71,833	426,326	271.50	55.50
Complete	914	W210	266,546	1195,831	189.10	31.10
		W450	227,210	948,129	281.60	57.19

Partial= dataset with phenotypic information from 1974 to 2008; **Complete=** dataset with phenotyp**ic** information from 1974 and 2018; **SD=** standard deviation of total records by trait for each dataset.

Marker arrays

The SNPs arrays were customized based on different length size windows per chromosome to select proportions of 10.0, 5.0, 2.5, 1.25 and 0.625% of markers from the 427k panel. As a result, the SNPs arrays were customized with densities of 40k, 20k, 10k, 5k and 2k, respectively. Two criteria were used to select the SNPs within a chromosome window: (i) highly informative SNPs (H-I) and (ii) evenly distributed SNPs (E-D). For the first criteria, markers with MAF close to 0.5 were selected by chromosome for each customized marker density. To select evenly distributed SNPs, the average distance between adjacent SNPs by chromosome for each customized marker density was considered. A total of 11 scenarios were evaluated considering the customized SNP arrays (427k, 40k_H-I, 40k_E-D, 20k_H-D, 20k_E-D, 10k_H-I, 10k_E-D, 5k_H-I, 5k_E-D, 2k_H-I, and 2k_E-D). The number of selected SNPs, and the average MAF and distance between adjacent markers by chromosome for each customized array are shown in Appendix. The customized SNP arrays were assembled using custom scripts written in Python 3.

Genetic structure analysis and effective population size

The genetic structure was estimated defining the covariance between animals from principal components (PC) of the genomic relationship matrix for each customized marker density arrays, applying the PreGSF90 program (Aguilar et al., 2010)

The effective population size ($\mathbf{N_e}$) was estimated for each customized marker density arrays using the SNP1101 v 1.0 software (Sargolzaei, 2014). The analysis was based on the extend of LD using the $\boldsymbol{r^2}$ statistic (Sved, 1971), represented as follows:

$$N_e = \left[\left(\frac{1}{E(r^2)}\right) - 1\right]\frac{1}{4c}$$

Where **c** is the distance in Morgans between two markers estimated for each chromosome in LD. The $E(r^2)$ is the expected $\mathbf{r^2}$ at a distance **c**, calculated as follows:

$$E(r^2) = \frac{1}{1 + 4N_e c}$$

Each genetic distance **(c)** corresponds to a value of **t** generation in the past (Hayes et al., 2003), obtained as follows:

$$t = \frac{1}{2c}$$

Genetic parameter estimation and prediction models

The contemporary groups (CG) for W450 were composed by farm, year and season of birth, sex, and management group at yearling *(n=21,537*). For W210, the CG were composed by farm, sex, year and season of birth, and management group at weaning (*n=36,196*). Records within ±3.5 standard deviations range from the CG mean, and CG with at least ten animals were considered in the analysis. The ssGBLUP is a modification of the BLUP model, which consist of replacing the

$\mathbf{A}^{-1}$ matrix with the $\mathbf{H}^{-1}$ by combining pedigree and genomic information (Aguilar et al., 2010), was applied as follows:

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

where $\mathbf{H}^{-1}$ is the inverse of the realized relationship matrix that incorporates the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$) and the inverse of the numerator relationship matrix for genotyped animals $\mathbf{A}_{22}^{-1}$. The **G** matrix was created according to VanRaden, (2008):

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

where **M** is a matrix of marker alleles with *m* columns (m = total number of markers) and *n* rows (*n* = total number of genotyped individuals), and **P** is a matrix containing the frequency of the second allele (p_j), expressed as $2p_j$. $\mathbf{M}_{ij}$ was 0 if the genotype of individual *i* for SNP *j* was homozygous for the first allele, 1 if heterozygous, or 2 if homozygous for the second allele. The variance component estimation and solutions for W210 and W450 direct effect and W210 maternal effect were obtained by BLUPF90 family programs (Aguilar et al., 2014; Misztal et al., 2002). Additionality, the overall linear animal model used was:

y= Xb + Za + Mm + Wmpe + e

where ***y*** is the vector of phenotypes, **b** is the vector of fixed effects (CG for both traits and the class of dam age at calving for W210 and W450), **a** is the vector of the additive genetic effect, **m** is the vector of random maternal additive genetic effects (only for W210), **mpe** is the vector of random maternal permanent environmental effects (only for W210), **X**, **Z, M** and **W** are the incidence matrices; and **e** is the vector of random residuals. It was assumed that $\mathrm{E}[\mathrm{y}] = \mathrm{X}\beta$; the direct additive genetic, genetic maternal, maternal permanent environmental, and

residual effects were normally distributed with means of zero and $\mathrm{Var(a)} = \mathrm{H} \otimes \mathrm{S_a}$; $\mathrm{Var(m)} = \mathrm{H} \otimes \mathrm{S_m}$; $\mathrm{Var(mpe)} = \mathrm{H} \otimes \mathrm{S_{mpe}}$; $\mathrm{Var(e)} = \mathrm{I} \otimes \mathrm{S_e}$; where $\mathbf{S_a}$ is the additive genetic covariance matrix; $\mathbf{S_m}$ is the covariance matrix of maternal additive genetic effects; $\mathbf{S_{mpe}}$ is the covariance matrix of maternal permanent environmental effects; $\mathbf{S}_e$ is the residual covariance matrix, and $\mathbf{I}$ is the identity matrix. The covariance between direct and maternal genetic effects was set to zero.

Prediction ability, inflation and bias

The prediction ability, inflation, and bias of the genomic prediction for each scenario was evaluated with the Linear Regression (LR) method (Legarra & Reverter, 2018), using the correlation (r), regression (b_1) coefficients and the difference from both datasets (partial and complete). The genotyped animals born after 2008 were selected for the validation dataset (*n=267*) and those before 2008 for the training dataset (*n=647*). A principal component analyses (PCA) was performed using the genomic relationship matrix to verify the existence of population substructure in the validation and training datasets for all 11 scenarios evaluated (Figure 1).

The correlation (r) between the GEBVs obtained with the complete dataset and 427k array (GEBVc_427k) with the GEBVs predicted with the partial dataset (GEBVp) for each scenario (GEBVp_427k, GEBVp_40k, GEBVp_20k, GEBVp_10k, GEBVp_5k, GEBVp_2k) was calculated. The inflation was evaluated through the regression of GEBVc_427k on GEBVp for each scenario considering only the animals of the validation dataset. The bias was measured as the difference between the GEBVp for each scenario proposed and the GEBVc_427K for the genotyped animals within the validation dataset.

RESULTS AND DISCUSSION

The variance component estimates for W210 and W450 were higher using the completed data than those obtained using the partial dataset (Table 2), and the maternal genetic additive variance for W210 showed the highest increase. The direct heritability for W450 displayed the highest increase using the complete dataset, probably, as consequence of larger bias correction or adjustment as consequence of more current records, since growth traits recorded after weaning are more affected by selection and Bulmer effect (van Grevenhof, van Arendonk, & Bijma, 2012). Cesarani et al. (2020) for average daily gain and conformation related traits in Simmental cattle reported changes in variance components over time and suggested that the genetic parameter estimated from the current population enabled the prediction of less biased GEBVs.

Table 2. Variance component estimation for adjusted weight at 210 days (W210) and 450 days of age (W450) using the complete and partial dataset.

Dataset	Trait	σ_a^2	σ_e^2	σ_m^2	σ_{mpe}^2	h^2	h_m^2	c^2
Partial	W210	75.70	268.78	33.64	75.41	0.22	0.07	0.17
	W450	269.07	563.70	-	-	0.32	-	-
Complete	W210	86.61	277.50	39.57	80.80	0.24	0.08	0.17
	W450	303.54	540.71	-	-	0.36	-	-

Partial= dataset with phenotype information from 1974 to 2008; **Complete=** dataset with phenotype information from 1974 and 2018; σ_a^2**=** genetic additive variance; σ_e^2**=** residual variance**;** σ_m^2**=** maternal variance; σ_{mpe}^2**=** maternal permanent environmental variance; h^2**=** direct heritability; h_m^2**=** maternal heritability and c^2**=** maternal permanent environmental effect by trait for each dataset.

The GEBV predictions for direct and maternal effect and their respective BIF accuracies for the validation dataset considering both datasets (partial and complete) with different customized SNP arrays for W210 and W450 are shown in Table 3. As expected, the GEBV accuracies for both traits were higher in the

complete than partial datasets. The GEBV accuracies were higher for W210, followed by W450 and maternal effect for W210 (MW210). Despite the higher heritability estimates for W450 in both datasets, the higher number of phenotypic records and offspring with records for W210 (Table 1) explained the higher GEBV accuracy (BIF) for such trait when considering the partial dataset, while with complete dataset the GEBV accuracy was proportional to heritability. Silva et al. (2016) in an experimental Nellore cattle population applied an index that combined parent average and direct genomic value to predict the GEBV for feed efficiency-related traits and reported higher GEBV accuracy for low-heritable traits due to the higher contribution of parental average (Yang & Su, 2016), working in a pig population indicated that the reliability of the GEBV predictions improved when phenotypic and pedigree data of relatives was incorporated. As expected, the reliability of GEBVs for MW210 were lower than those obtained for the GEBVs of W210 and W450 direct effect. The maternal effect is difficult to improve since it is expressed in the weaning weight of the bull's grand progeny and the heritability for the maternal effect is usually low (Kluska et al., 2018).

Table 3. Genomic breeding values (GEBV) predictions and BIF accuracies for the validation dataset for W210 (direct and maternal effect) and W450 (direct effect) considering different customized SNP arrays.

Trait	Array	Criteria	**GEBVp**±SD	**GEBVc**±SD	**AccBIF_p**±SD	**AccBIF_c**±SD
W210	**427k**	**-**	12.22±5.22	8.75±6.95	0.33±0.06	0.48±0.16
	40k	**H-I**	12.23±5.22	8.78±6.30	0.33±0.07	0.49±0.16
		E-D	12.25±5.28	8.75±6.33	0.33±0.07	0.48±0.16
	20k	**H-I**	12.12±5.24	8.80±6.26	0.33±0.07	0.49±0.16
		E-D	12.23±5.24	8.72±6.31	0.33±0.07	0.48±0.16
	10k	**H-I**	12.11±5.20	8.80±6.23	0.34±0.07	0.50±0.16
		E-D	12.23±5.29	8.75±6.24	0.34±0.07	0.49±0.16

	5k	H-I	12.04±5.33	8.83±6.37	0.35±0.07	0.51±0.15
		E-D	12.22±5.55	8.79±6.55	0.35±0.07	0.50±0.15
	2k	H-I	12.13±5.57	8.78±6.44	0.38±0.06	0.53±0.14
		E-D	12.12±5.44	8.73±6.28	0.38±0.06	0.53±0.14
MW210	427k	-	1.78±2.19	1.30±3.39	0.17±0.06	0.37±0.13
	40k	H-I	1.78±2.23	1.25±3.80	0.19±0.06	0.38±0.13
		E-D	1.77±2.22	1.29±3.79	0.19±0.06	0.38±0.13
	20k	H-I	1.77±2.22	1.25±3.81	0.19±0.06	0.38±0.13
		E-D	1.80±2.22	1.27±3.82	0.19±0.06	0.38±0.13
	10k	H-I	1.74±2.21	1.22±3.86	0.19±0.06	0.39±0.13
		E-D	1.76±2.22	1.26±3.86	0.19±0.06	0.38±0.13
	5k	H-I	1.71±2.24	1.26±3.85	0.20±0.06	0.40±0.13
		E-D	1.71±2.34	1.30±3.87	0.20±0.06	0.40±0.13
	2k	H-I	1.73±2.36	1.25±3.98	0.20±0.06	0.42±0.12
		E-D	1.73±2.39	1.20±4.02	0.20±0.06	0.43±0.12
W450	427k	-	23.09±9.37	18.48±11.64	0.19±0.04	0.60±0.13
	45k	H-I	23.17±9.35	18.54±11.67	0.20±0.04	0.61±0.12
		E-D	23.16±9.46	18.43±11.66	0.20±0.04	0.60±0.13
	20k	H-I	23.13±9.29	18.60±11.56	0.20±0.04	0.61±0.12
		E-D	23.17±9.38	18.42±11.65	0.20±0.04	0.61±0.12
	10k	H-I	23.00±9.44	18.58±11.48	0.20±0.04	0.61±0.12
		E-D	23.21±9.46	18.38±11.57	0.20±0.04	0.61±0.12
	5k	H-I	23.08±9.79	18.72±11.62	0.21±0.04	0.63±0.12

	E-D	23.21±9.84	18.53±11.76	0.21±0.04	0.62±0.12
2k	**H-I**	23.26±10.23	18.54±11.83	0.23±0.04	0.65±0.11
	E-D	22.91±9.90	18.51±11.50	0.23±0.04	0.65±0.11

GEBVp = Genomic breeding value estimate for the partial dataset; **GEBVc =** Genomic breeding value estimate for the complete dataset; **Acc_{BIF_p}=** BIF accuracy for the partial scenario; **Acc_{BIF_c}=** BIF accuracy for complete scenario; **W210=** Weight at 210 days of age; **MW210=** Maternal effect for weight at 210 days of age; **W450=** Weight at 450 days of age; **H-I=** Selected markers with high informativity; **E-D=** Markers evenly distributed; **BIF=** Beef Improvement Federation; **SD=** standard deviation.

The GEBVs for the validation dataset obtained with the complete dataset displayed lower mean values than did the partial dataset (Table 3). Such results might reflect the overestimation of the GEBVs prediction of the validation dataset when phenotypic information was omitted, and therefore, the prediction is based merely on genomic information. Several studies have reported that genomic predictions were mostly overestimated in young animals since proven animals are usually chosen in the reference population, which represents a highly selected sample of proven bulls (e.g., Wiggans et al., 2011; Tsuruta et al., 2013; Gunia et al., 2014; Gao et al., 2015).

For both criteria used to customize the SNPs arrays, the GEBV accuracies increased as the marker density decreased. A similar pattern was reported for milk fat percentage by Goddard et al. (2011), and the authors pointed out that the accuracies calculated from the elements of the diagonal of the inverse G matrix are sensitive to the dimensionality of such matrix. The use of low-density SNP arrays increased the average elements of the G inverse matrix and the genomic kinship bias among animals, overestimating the GEBV accuracies (Chen et al., 2011). Therefore, the variance of the G matrix and the sampling error increased as the SNP arrays density decreased.

It is important to highlight that for the customized arrays, the average of MAF was 0.49 (H-I) and 0.46 (E-D), and for the 427k array was 0.46 (Appendix).

The slight difference regarding the MAF informativeness were not large enough to cause noteworthy changes in the GEBV accuracies and prediction ability. Zhu et al. (2017) obtained higher prediction ability for growth-related traits in Chinese Simmental cattle using low density marker arrays with large SNP effect evenly distributed with MAF ranging from 0.01 to 0.1. Chen et al. (2011) working with broilers and using the ssGBLUP method, evaluated different MAF thresholds and reported that the highest genomic prediction ability for liveweight and breast meat was attained with MAF around 0.4. Zhu et al. (2017) and Chen et al. (2011) concluded that differences in genetic structure and changes in the G matrix properties could explain the variation in the prediction ability for the different MAF thresholds. Forni et al. (2011) also working with the ssGBLUP in a pig population evaluated different genomic relationship matrices using the same SNP density applying different allele frequencies and reported less inflated accuracies using a normalized G matrix compared to allele frequencies equal to 0.5 or observed allele frequencies.

The prediction ability or correlation (r) between the GEBVc_427k and the GEBV obtained with customized SNP densities in the validation subset (partial dataset) is shown in Table 4. For both traits, the prediction ability improved as the marker density increased. The prediction ability was similar and moderate to high for marker density arrays higher than 10k. Comparing the criteria to select SNPs (H-I or E-D), similar prediction abilities were obtained for marker densities higher than 2k arrays. For 2k arrays, higher prediction abilities were observed for E-D markers for both traits. Salvian et al. (2020) working with broilers, reported similar genomic predictions for SNP density arrays varying between 37k and 370k markers using the ssGBLUP method. These authors also observed that the genomic relationship matrix with at least 10% of markers from the HD panel (600k Affymetrix Axiom Genotyping Array) is necessary to obtain reliable genomic evaluations. In Angus cattle, SNP density arrays higher than 2.5k and lower than 10k markers were adequate to shape the G matrix that provided the highest GEBVs reliability for feed efficiency traits (Rolf et al., 2010). In a simulated study, Barjasteh et al. (2020) observed that the prediction ability was similar using 5k and

50k SNP densities arrays when the training population was strongly related with the validation set. As in our study, a strong relationship between the training and validation datasets was demonstrated through the PCA (Figure 01), and there was absence of population substructure formation among the different customized arrays. In this way, Lee et al. (2017), demonstrated that the validation populations could be a predictor of population genomic variations when the relationship between validation and training populations is strong. Furthermore, the authors pointed that information about population structure like numbers of chromosome segments and N_e are necessary before implementing genomic selection. In livestock, most of the genetic additive variation for complex traits is captured by a limited number of SNP markers without redundant effect (Misztal et al., 2015), due to small N_e of current livestock populations as result of few representative individuals selected in the past (Misztal et al. (2020). Minor differences in the N_e estimation for the current population (generation 1 and 2) were observed between customized arrays, although for generations 3, 4 and 5 ago the E-D marker selection criteria displayed slightly higher Ne values than H-I (Table 5). For all the evaluated scenarios, the estimated N_e displayed values above 124, reflecting that current population is genetically diverse. According to Meuwissen, (2009), the minimum N_e should be 100 to ensure the viability of populations over time. Cardoso et al. (2018) estimated the Ne in three Nellore cattle experimental populations, one selected and two non-selected (closed herd) for yearling weight, and the authors reported a genomic stratification between populations and higher N_e value in the selected herd (177) than in the non-selected herds (51 and 88). Studies with taurine beef cattle breed as Angus and Charolais breed, N_e estimates of 267 and 285 were reported, respectively, supporting the diversity and genetic structure of breeds (Lu et al., 2012).

Table 4. Prediction ability (r), inflation (b1), and bias (d) for the validation partial dataset for W210 (direct and maternal effect) and W450 (direct effect) considering different customized SNP arrays.

Trait	Array	Criteria	r	b_1±SD	d±SD
W210	427k	-	0.78	0.95±0.05	3.47±4.65
	40k	H-I	0.78	0.95±0.05	3.47±4.66
		E-D	0.78	0.94±0.05	3.50±4.68
	20k	H-I	0.78	0.94±0.05	3.37±4.70
		E-D	0.77	0.93±0.05	3.48±4.74
	10k	H-I	0.77	0.94±0.05	3.36±4.76
		E-D	0.77	0.92±0.05	3.47±4.75
	5k	H-I	0.76	0.90±0.05	3.29±4.85
		E-D	0.77	0.88±0.05	3.47±4.79
	2k	H-I	0.72	0.82±0.05	3.38±5.18
		E-D	0.74	0.86±0.05	3.39±5.03
MW210	427k	-	0.48	0.83±0.08	0.48±2.95
	40k	H-I	0.47	0.80±0.08	0.48±2.98
		E-D	0.47	0.81±0.08	0.46±2.96
	20k	H-I	0.46	0.79±0.08	0.47±2.99
		E-D	0.47	0.81±0.08	0.50±2.90
	10k	H-I	0.46	0.79±0.08	0.44±3.00
		E-D	0.47	0.81±0.08	0.45±2.98
	5k	H-I	0.45	0.77±0.08	0.45±3.01
		E-D	0.45	0.74±0.08	0.41±3.04

	2k	H-I	0.43	0.69±0.08	0.43±3.11
		E-D	0.42	0.67±0.08	0.43±3.15
W450	427k	-	0.72	0.89±0.05	4.50±7.87
	45k	H-I	0.72	0.89±0.05	4.57±7.96
		E-D	0.73	0.89±0.05	4.56±7.82
	20k	H-I	0.72	0.90±0.05	4.53±7.91
		E-D	0.72	0.89±0.05	4.57±795
	10k	H-I	0.71	0.87±0.05	4.41±8.13
		E-D	0.71	0.87±0.05	4.61±8.11
	5k	H-I	0.69	0.82±0.0	4.49±8.40
		E-D	0.72	0.84±0.04	4.62±8.07
	2k	H-I	0.66	0.75±0.05	4.66±9.03
		E-D	0.69	0.80±0.05	4.31±8.50

r = Correlation coefficient estimate between the GEBV of the complete dataset using arrays of 427k markers and GEBV of the partial dataset using arrays with different densities ; **b_1 =** Regression coefficient of the GEBV of the complete dataset using arrays of 427k markers on GEBV of the partial dataset using arrays with different densities; **d =** Value of the difference between GEBV of the partial dataset using arrays with different densities and GEBV of the complete dataset using only the array of 427k markers**; W210=** weight at 210 days of age; **MW210=** maternal effect for 210 days of age; **W450=** weight at 450 days of age; **H-I=** selected markers with high informativeness; **E-D=** markers evenly distributed; **SD=** standard deviation.

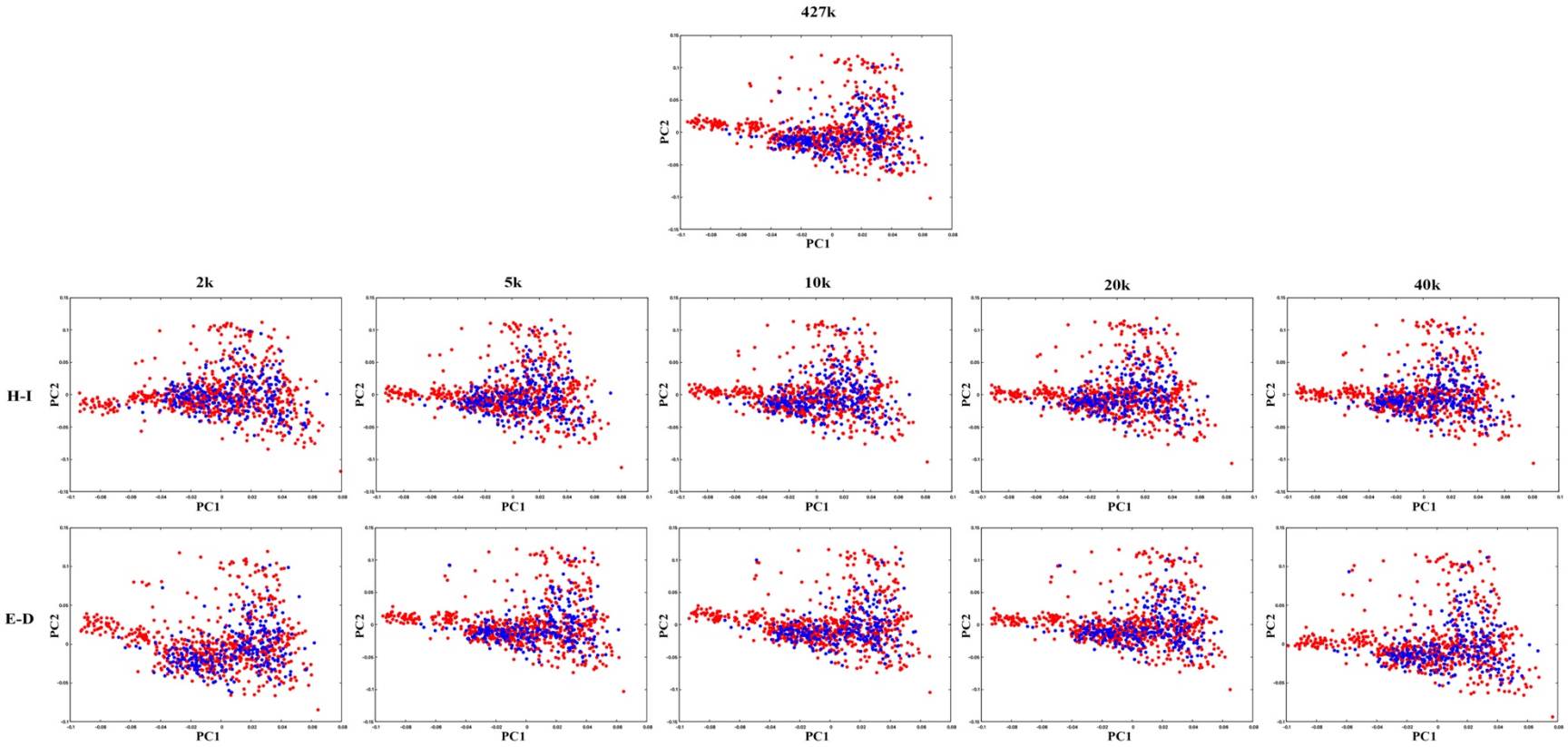

Figure 1. First and second principal components (PC) of the genomic relationship matrix (G) with the customized arrays of 40k (arrays with 45,930 and 42,734 markers for H-I and E-D respectively), 20k (arrays with 22,538 and 21,360 markers for H-I and E-D respectively), 10k (arrays with 11,010 and 10.671 markers for H-I and E-D respectively), 5k (arrays with 5,412 and 5,329 markers for H-I and E-D respectively), 2k (arrays with 2,669 and 2,656 markers for H-I and E-D respectively), 427k (array with 427,495 markers). Red dots = training population; Blue dots = validation population.

The regression coefficients (b_1; inflation) and the difference between the GEBVc_427k and GEBV (d; bias) obtained for the customized SNP density arrays for the validation dataset are shown in Table 4. The inflation of the predictions increased as the marker density decreased. The higher inflation was observed for marker densities arrays of 2k and 5k, while the 10k, 20k, and 40k displayed similar and less inflated predictions. Minor differences between H-I and E-D criteria were detected when considering marker densities arrays below 5k. The direct effect for W210 displayed less overdispersion than W450, while the MW210 presented higher overdispersion than the direct effects for W210 and W450. As the SNP density decays, the low heritability trait, such as the MW210, displayed the highest inflation, and the inflation was maximum at the lowest SNP density arrays. Salvian et al. (2020) in broilers reported prediction ability with lower inflation as the SNP density increased, however, the inflation was similar from the subset of ~18,500 to 370,600 SNPs. Ma et al. (2015), observed changes in the inflation rate for a Danish Jersey population where the markers did not fully capture the total genetic variance and insufficient LD between maker and causal genes. In our study, the population structure for the different SNP array densities based on the genomic relationship matrix was similar for training and validation subpopulations (Figure 1). As the marker density decreased, the similarity between animals increased due to detection of false identical segments by descent, increasing the mistakes in the coancestry index between individuals (Wang, 2014). In this regard, the inflation of the GEBV predictions with ssGBLUP method are affected due to changes in the genomic relationship matrix, which can vary in magnitude of $H^{-1}\alpha$ element of the model (Tsuruta et al., 2019).

The prediction bias tends to decrease as the SNP densities array decrease, and there were no major differences among the criteria applied to select such SNPs. Despite there was a trend in decreasing the prediction bias as the SNP density array decreased, the dispersion or variance of the predictions increased, elucidated by the standard deviation. For all customized SNP arrays, the bias was higher for direct effects than did for the maternal. For direct effects, the W210 showed less bias than did W450. Legarra & Reverter, (2018) reported that, over

time, populations suffer changes in their genetic structure, and this fact favors the prediction bias. In this study, the variance component estimates for W210 and W450 differed between training and validation population, probably due to strong selection applied after 2008 due to genomic selection implementation. The bias measured through the LR method has shown susceptibility, when incorrect genetic parameters were included in the evaluation or when some environmental effect was omitted, contributing to errors in the model that reflect a bias distant to the true value (Macedo et al., 2020). Thus, genomic evaluations considering the genetic structure of the population together with the changes in the variance components should contribute to more robust predictions.

Table 5. Estimated an effective population size (N_e) over time considering different customized SNP arrays.

		Generations ago				
Array	**Criteria**	**1**	**2**	**3**	**4**	**5**
427k	**-**	127.88	152.38	150.08	155.72	160.46
40k	**H-I**	131.44	148.85	136.77	132.00	132.98
	E-D	126.68	150.69	150.10	153.01	157.72
20k	**H-I**	134.36	151.86	138.95	134.50	134.70
	E-D	125.57	151.64	149.90	156.12	158.49
10k	**H-I**	128.47	152.21	141.89	134.95	133.21
	E-D	126.60	151.90	152.38	155.89	158.87
5k	**H-I**	126.24	143.57	131.81	126.98	126.54
	E-D	124.29	147.16	151.76	152.06	152.03
2k	**H-I**	120.75	146.11	139.14	130.37	130.65
	E-D	128.50	140.73	148.70	147.32	159.58

40k= SNP array with more than 40,000 markers; **20k=** SNP array with more than 20,000 markers; **10k=** SNP array with more than 10,000 markers; **5k=** SNP array with more than 5,000 markers; **2k=**

SNP array with more than 2,000 markers; **427k=** SNP array with 427,495 markers; **H-I=** selected markers with high informativeness; **E-D=** markers evenly distributed.

There is a constantly growing number of custom SNP arrays in cattle protected by intellectual property produced by two major companies (Illumina and Thermo Fisher). The continuous increase in the number of SNP arrays available was not accompanied by an organized effort to standardize the genomic data, making comparison and cross-compatibility of SNP arrays difficult. In this sense, the genetic improvement programs received genotypes obtained in several genotyping companies with different array density and percentage of common SNPs, and the major challenge for the genetic improvement programs is to standardize for a single density arrays to carry out the genomic evaluation trough the ssGBLUP method. Imputation is an attractive cost-saving to implementing genomic selection to infer the non-common SNPs and increase the marker density, however, combining different commercial SNPs arrays with a varying number of common SNPs is a challenge to attain a reliable imputation together with a low error rate. Additionally, imputation to higher density panels in some cattle populations, i.e., indicine cattle, is not sufficiently accurate using commercially available low-density panels (Carvalheiro et al., 2014; Bernardes et al., 2019). The results obtained in this study shown that moderate to low density *in silico* arrays, prioritizing informative markers, were capable to estimate genomic relationship coefficients allowing the estimation of reliable genomic predictions for growth and maternal related traits in Nelore cattle. The development of customized SNPs arrays for *Bos taurus indicus* beef cattle breeds is an interesting strategy to obtain reliable and unbiased GEBV predictions lowering the genotyping cost, decreasing the SNP imputation, and increasing the computational optimization.

The GEBV accuracy is obtained from the PEV based on the inverse of the left-hand side (LHS) and reflects the quality and quantity of the data to obtain the genomic predictions in breeding programs. In the current study, the GEBV accuracies of the validation dataset were overestimated in the complete and partial scenario as the SNP density arrays decreased, and the GEBV of young animals without phenotypic information and progeny records was overestimated as the

SNP marker density decayed. Therefore, the prediction ability evaluated through the correlation between the GEBV using all the available information and the GEBV obtained with the customized SNPs arrays and omitting the phenotypic information and progeny records is a more adequate approach to evaluate the impact of marker density with the ssGBLUP method than the genomic accuracies obtained from PEV.

CONCLUSION

The results of this study revealed that genomic selection with low-density customized arrays could be feasible and cost-effective in Nellore beef cattle. Therefore, there is an important margin to reduce high or moderate to low-density SNPs arrays without compromising the predictive capacity of the genomic information using the ssGBLUP method. This fact opens the opportunity for indicine beef cattle breeding programs develop *in-silico* SNPs markers arrays for GEBVs predictions, minimizing the number of SNPs to be imputed. At least 10,000 informative SNPs obtained from the Illumina BovineHD BeadChip SNPs are necessary to adequately predict the GEBVs for growth and maternal related traits of young candidates with the ssGBLUP method. Additionally, the criteria to select the SNPs to customize the arrays is non-essential if the markers are properly distributed in the genome and highly informative.

ACKNOWLEDGEMENTS

The authors thank National Association of Breeders and Researchers (ANCP), the Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias (FCAV/Unesp), the Universidad de la Republica, Facultad de Veterinaria (UdelaR), Departamento de Genética y Mejoramiento Animal and the Instituto Nacional de Investigación Agropecuaria of Uruguay (INIA).

FUNDING

This study was supported by the Programa Estudantes Convênio de Pós-Graduação da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (PECPG-CAPES: 88881.154576/2017-01).

REFERENCES

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & 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(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Aguilar, Ignacio, Misztal, I., Tsuruta, S., Legarra, A., Wang, H., Aguilar, I., … Wang, H. (2014). PREGSF90-POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. Retrieved from http://nce.ads.uga.edu/wiki/doku.php].

Barjasteh, S., Dashab, G. R., Rokouei, M., Shariati, M. M., & Vafaye Valleh, M. (2020). Comparing Different Marker Densities and Various Reference Populations Using Pedigree-Marker Best Linear Unbiased Prediction (BLUP) Model. In Iranian Journal of Applied Animal Science (Vol. 10). Islamic Azad University - Rasht Branch. Retrieved from Islamic Azad University - Rasht Branch website: www.ijas.ir

Bernardes, P. A., Nascimento, G. B. do, Savegnago, R. P., Buzanskas, M. E., Watanabe, R. N., de Almeida Regitano, L. C., … Munari, D. P. (2019). Evaluation of imputation accuracy using the combination of two high-density panels in Nelore beef cattle. Scientific Reports, 9(1), 1–10. https://doi.org/10.1038/s41598-019-54382-w

Boddhireddy, P., Prayaga, K., Barros, P., Lôbo, R., & Denise, S. (2014). Proceedings, 10 th World Congress of Genetics Applied to Livestock Production Genomic Predictions of Economically Important Traits in Nelore Cattle of Brazil.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., … Wiggans, G. R. (2012). Design of a bovine low-density snp array optimized for imputation. PLoS ONE, 7(3). https://doi.org/10.1371/journal.pone.0034130

Cardoso, D. F., de Albuquerque, L. G., Reimer, C., Qanbari, S., Erbe, M., do Nascimento, A. v., … Tonhati, H. (2018). Genome-wide scan reveals

population stratification and footprints of recent selection in Nelore cattle. Genetics Selection Evolution, 50(1), 22. https://doi.org/10.1186/s12711-018-0381-2

Carvalheiro, R., Boison, S. A., Neves, H. H. R., Sargolzaei, M., Schenkel, F. S., Utsunomiya, Y. T., … Garcia, J. F. (2014). Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution, 46(1), 1–11. https://doi.org/10.1186/s12711-014-0069-1

Cesarani, A., Hidalgo, J., Garcia, A., Degano, L., Vicario, D., Masuda, Y., … Lourenco, D. (2020). Beef trait genetic parameters based on old and recent data and its implications for genomic predictions in Italian Simmental cattle. Journal of Animal Science, 98(8). https://doi.org/10.1093/jas/skaa242

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science, 89(9), 2673–2679. https://doi.org/10.2527/jas.2010-3555

Chiaia, H. L. J., Peripolli, E., de Oliveira Silva, R. M., Feitosa, F. L. B., de Lemos, M. V. A., Berton, M. P., … Baldi, F. (2018). Genomic prediction ability for beef fatty acid profile in Nelore cattle using different pseudo-phenotypes. Journal of Applied Genetics, 59(4), 493–501. https://doi.org/10.1007/s13353-018-0470-5

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics, Vol. 193, pp. 347–365. Genetics. https://doi.org/10.1534/genetics.112.147983

de Lima, L. G., de Souza, N. O. B., Rios, R. R., de Melo, B. A., dos Santos, L. T. A., Silva, K. de M., … Fraga, A. B. (2020, January 1). Advances in molecular genetic techniques applied to selection for litter size in goats (Capra hircus): a review. Journal of Applied Animal Research, Vol. 48, pp. 38–44. Taylor and Francis Ltd. https://doi.org/10.1080/09712119.2020.1717497

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013, February 1). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, Vol. 193, pp. 327–345. Genetics. https://doi.org/10.1534/genetics.112.143313

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., … Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science, 95, 4114–4129. https://doi.org/10.3168/jds.2011-5019

Forni, S., Aguilar, I., & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution, 43(1), 1–7. https://doi.org/10.1186/1297-9686-43-1

Gao, H., Madsan, P., Nielsen, U. S., Aamand, G. P., Su, G., Byskov, K., & Jensen, J. (2015). Including different groups of genotyped females for genomic prediction in a Nordic Jersey population. Journal of Dairy Science, 98(12), 9051–9059. https://doi.org/10.3168/jds.2015-9947

Garcia, A. L. S., Bosworth, B., Waldbieser, G., Misztal, I., Tsuruta, S., & Lourenco, D. A. L. (2018). Development of genomic predictions for harvest and carcass weight in channel catfish 06 Biological Sciences 0604 Genetics. Genetics Selection Evolution, 50(1), 66. https://doi.org/10.1186/s12711-018-0435-5

Gipson, T. A. (2019). — Special Issue — Recent advances in breeding and genetics for dairy goats. Asian-Australasian Journal of Animal Sciences, 32(8), 1275–1283. https://doi.org/10.5713/ajas.19.0381

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2), 245–257. https://doi.org/10.1007/s10709-008-9308-0

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics, 128(6), 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Gunia, M., Saintilan, R., Venot, E., Hozé, C., Fouilloux, M. N., & Phocas, F. (2014). Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. Journal of Animal Science, 92(8), 3258–3269. https://doi.org/10.2527/jas.2013-7478

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. Genetics, 182(1), 343–353. https://doi.org/10.1534/genetics.108.100289

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., & van der Werf, J. H. J. (2012). Accuracy of genotype imputation in sheep breeds. Animal Genetics, 43(1), 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x

Hayes, Ben J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research, 13(4), 635–643. https://doi.org/10.1101/gr.387103

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009, January 1). Genomic selection for crop improvement. Crop Science, Vol. 49, pp. 1–12. John Wiley & Sons, Ltd. https://doi.org/10.2135/cropsci2008.08.0512

Joshi, R., Skaarud, A., de Vera, M., Alvarez, A. T., & Ødegård, J. (2020). Genomic prediction for commercial traits using univariate and multivariate approaches in Nile tilapia (Oreochromis niloticus). Aquaculture, 516, 734641. https://doi.org/10.1016/j.aquaculture.2019.734641

Kluska, S., Olivieri, B. F., Bonamy, M., Chiaia, H. L. J., Feitosa, F. L. B., Berton, M. P., … 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. https://doi.org/10.1016/j.livsci.2018.08.015

Lee, S. H., Clark, S., & van der Werf, J. H. J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLoS ONE, 12(12). https://doi.org/10.1371/journal.pone.0189775

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. Livestock Science, 166(1), 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method 01 Mathematical Sciences 0104 Statistics. Genetics Selection Evolution, 50(1), 53. https://doi.org/10.1186/s12711-018-0426-6

Liu, Y., Xu, L., Wang, Z., Xu, L., Chen, Y., Zhang, L., … Li, J. (2019). Genomic prediction and association analysis with models including dominance effects for important traits in Chinese simmental beef cattle. Animals, 9(12). https://doi.org/10.3390/ani9121055

Lourenco, D. A. L., Misztal, I., Tsuruta, S., Aguilar, I., Ezra, E., Ron, M., … Weller, J. I. (2014). Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. Journal of Dairy Science, 97(3), 1742–1752. https://doi.org/10.3168/jds.2013-6916

Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., & Misztal, I. (2020). Single-step genomic evaluations from theory to practice: using snp chips and sequence data in blupf90. Genes, 11(7), 1–32. https://doi.org/10.3390/genes11070790

Lu, D., Sargolzaei, M., Kelly, M., Li, C., Voort, G. vander, Wang, Z., … Miller, S. P. (2012). Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Frontiers in Genetics, 3(AUG). https://doi.org/10.3389/fgene.2012.00152

Ma, P., Lund, M. S., Nielsen, U. S., Aamand, G. P., & Su, G. (2015). Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. Journal of Dairy Science, 98(12), 9026–9034. https://doi.org/10.3168/jds.2015-9703

Macedo, F. L., Reverter, A., & Legarra, A. (2020). Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. Journal of Dairy Science, 103(1), 529–544. https://doi.org/10.3168/jds.2019-16603

Meuwissen, T. (2009). Genetic management of small populations: A review. Acta Agriculturae Scandinavica A: Animal Sciences, 59(2), 71–79. https://doi.org/10.1080/09064700903118148

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics, 157(4).

Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science, 92(9), 4648–4655. https://doi.org/10.3168/jds.2009-2064

Misztal, I, Tsuruta, S., Strabel, T., Auvray, B., Druet, T., & Lee, D. H. (2002). BLUPF90 AND RELATED PROGRAMS (BGF90). Retrieved from http://www.ozemail.com.au/~milleraj.

Misztal, Ignacy. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics, 202(2), 401–409. https://doi.org/10.1534/genetics.115.182089

Misztal, Ignacy, Fragomeni, B., Lourenco, D. A. L., Tsuruta, S., Masuda, Y., Aguilar, I., … Lawlor, T. J. (2015). Efficient inversion of genomic relationship matrix by the algorithm for proven and young (APY). Retrieved from https://hal.inrae.fr/hal-02743591

Misztal, Ignacy, Lourenco, D., & Legarra, A. (2020). Current status of genomic evaluation. Journal of Animal Science, 98(4), 1–14. https://doi.org/10.1093/jas/skaa101

Nayee, N., Sahana, G., Gajjar, S., Sudhakar, A., Trivedi, K., Lund, M. S., & Guldbrandtsen, B. (2018). Suitability of existing commercial single nucleotide polymorphism chips for genomic studies in Bos indicus cattle breeds and their

Bos taurus crosses. Journal of Animal Breeding and Genetics, 135(6), 432–441. https://doi.org/10.1111/jbg.12356

Pocrnic, I., Lourenco, D. A. L., Masuda, Y., Legarra, A., & Misztal, I. (2016). The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics, 203(1), 573–581. https://doi.org/10.1534/genetics.116.187013

Resende, J. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., … Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics, 190(4), 1503–1510. https://doi.org/10.1534/genetics.111.137026

Rezende, F. M., Pablo Nani, J., & Peñagaricano, F. (2019). Genomic prediction of bull fertility in US Jersey dairy cattle. Journal of Dairy Science, 102, 3230–3240. https://doi.org/10.3168/jds.2018-15810

Rolf, M. M., Taylor, J. F., Schnabel, R. D., McKay, S. D., McClure, M. C., Northcutt, S. L., … Weaber, R. L. (2010). Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genetics, 11. https://doi.org/10.1186/1471-2156-11-24

Salvian, M., Costa, G., Moreira, M., Spangler, M. L., & Mourão, G. B. (2020). Estimation of Breeding Values Using Different Densities of Snp to Inform Kinship in Broiler Chickens. https://doi.org/10.21203/rs.3.rs-32429/v1

Sargolzaei, M. (2014) SNP1101 user guide. 1.0.

Silva, R. M. O., Fragomeni, B. O., Lourenco, D. A. L., Magalhães, A. F. B., Irano, N., Carvalheiro, R., … Albuquerque, L. G. (2016). Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. Journal of Animal Science, 94(9), 3613–3623. https://doi.org/10.2527/jas.2016-0401

Silveira, L. S., Lima, L. P., Nascimento, M., Nascimento, A. C. C., & Silva, F. F. (2020). Regression trees in genomic selection for carcass traits in pigs. Genetics and Molecular Research, 19(1). https://doi.org/10.4238/gmr18498

Sousa, T. V., Caixeta, E. T., Alkimim, E. R., Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., … Resende, M. D. V. (2019). Early selection enabled by the implementation of genomic selection in coffea arabica breeding. Frontiers in Plant Science, 9. https://doi.org/10.3389/fpls.2018.01934

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology, 2(2), 125–141. https://doi.org/10.1016/0040-5809(71)90011-6

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Misztal, I., & Lawlor, T. J. (2019). Controlling bias in genomic breeding values for young genotyped bulls. Journal of Dairy Science, 102(11), 9956–9970. https://doi.org/10.3168/jds.2019-16789

Tsuruta, S., Misztal, I., & 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(5), 3332–3335. https://doi.org/10.3168/jds.2012-6272

van Grevenhof, E. M., van Arendonk, J. A., & Bijma, P. (2012). Response to genomic selection: The Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genetics Selection Evolution, 44(1). https://doi.org/10.1186/1297-9686-44-26

VanRaden, P. M. (2008a). Efficient methods to compute genomic predictions. Journal of Dairy Science, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M. (2008b). Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Waide, E. H., Tuggle, C. K., Serão, N. V. L., Schroyen, M., Hess, A., Rowland, R. R. R., … Dekkers, J. C. M. (2018). Genomic prediction of piglet response to infection with one of two porcine reproductive and respiratory syndrome virus isolates. Genetics Selection Evolution, 50(1), 3. https://doi.org/10.1186/s12711-018-0371-4

Wang, J. (2014). Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. Journal of Evolutionary Biology, 27(3), 518–530. https://doi.org/10.1111/jeb.12315

Wang, Q., Yu, Y., Li, F., Zhang, X., & Xiang, J. (2017). Predictive ability of genomic selection models for breeding value estimation on growth traits of Pacific white shrimp Litopenaeus vannamei. Chinese Journal of Oceanology and Limnology, 35(5), 1221–1229. https://doi.org/10.1007/s00343-017-6038-0

Weigel, K. A., van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., & Wiggans, G. R. (2010). Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science, 93(5), 2229–2238. https://doi.org/10.3168/jds.2009-2849

Wiggans, G. R., Cooper, T. A., VanRaden, P. M., & Cole, J. B. (2011). Technical note: Adjustment of traditional cow evaluations to improve accuracy of

genomic predictions. Journal of Dairy Science, 94(12), 6188–6193. https://doi.org/10.3168/jds.2011-4481

Wongpom, B., Koonawootrittriron, S., Elzo, M. A., Suwanasopee, T., & Jattawa, D. (2019). Accuracy of genomic-polygenic estimated breeding value for milk yield and fat yield in the Thai multibreed dairy population with five single nucleotide polymorphism sets. Asian-Australasian Journal of Animal Sciences, 32(9), 1340–1348. https://doi.org/10.5713/ajas.18.0816

Wu, X. L., Xu, J., Feng, G., Wiggans, G. R., Taylor, J. F., He, J., … Bauck, S. (2016). Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. PLoS ONE, 11(9), e0161719. https://doi.org/10.1371/journal.pone.0161719

Wu, X.-L., Li, H., Xu, J., Ferraz, J. B. S., Silva, L. R., Garcia, J. F., … Bauck, & S. (2018). Evaluation of genomic prediction accuracies of growth and reproduction traits in Nellore cattle using the new GGP® indicus low density SNP chip.

Yang, H., & Su, G. (2016). Impact of phenotypic information of previous generations and depth of pedigree on estimates of genetic parameters and breeding values. Livestock Science, 187, 61–67. https://doi.org/10.1016/j.livsci.2016.03.001

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565–569. https://doi.org/10.1038/ng.608

Zhu, B., Zhang, J. jing, Niu, H., Guan, L., GUO, P., XU, L. yang, … LI, J. ya. (2017). Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture, 16(4), 911–920. https://doi.org/10.1016/S2095-3119(16)61474-0

APPENDIX

Appendix. Parameters applied to customize the marker arrays with different densities.

BTA[1]	BTA length size (kb)	Criteria	Customized marker density arrays										
			40k		20k		10k		5k		2k		427k
			N SNPs selected	SNP distance (kb/SNP)	N SNPs selected	SNP distance (kb/SNP)	N SNPs selected	SNP distance (kb/SNP)	N SNPs selected	SNP distance (kb/SNP)	N SNPs selected	SNP distance (kb/SNP)	N SNPs selected
1	158395.2	**H-I**	2948	53.73	1424	111.23	693	228.56	342	463.14	169	937.25	28586
		E-D	2699	58.69	1349	117.42	674	235.01	337	470.02	168	942.83	
2	136702.6	**H-I**	2536	53.90	1239	110.33	601	227.46	297	460.28	146	936.32	22587
		E-D	2330	58.67	1165	117.34	582	234.88	291	469.77	145	942.78	
3	121358.4	**H-I**	2218	54.72	1103	110.03	536	226.41	262	463.20	129	940.76	20680
		E-D	2068	58.68	1034	117.37	517	234.74	258	470.38	129	940.76	
4	120592.4	**H-I**	2217	54.39	1098	109.83	533	226.25	261	462.04	129	934.82	19146
		E-D	2055	58.68	1027	117.42	513	235.07	256	471.06	128	942.13	
5	121051.2	**H-I**	2204	54.92	1096	110.45	536	225.84	261	463.80	129	938.38	18921
		E-D	2063	58.68	1031	117.41	515	235.05	257	471.02	128	945.71	
6	119321.9	**H-I**	2201	54.21	1064	112.14	520	229.47	257	464.29	128	932.20	23275
		E-D	2033	58.69	1016	117.44	508	234.89	254	469.77	127	939.54	

7	112576.5	**H-I**	2017	55.81	990	113.71	491	229.28	244	461.38	120	938.14	19552
		E-D	1918	58.69	959	117.39	479	235.02	239	471.03	119	946.02	
8	113299.9	**H-I**	2075	54.60	1009	112.29	495	228.89	246	460.57	121	936.36	20107
		E-D	1931	58.67	965	117.41	482	235.06	241	470.12	120	944.17	
9	105666.7	**H-I**	1971	53.61	950	111.23	463	228.22	228	463.45	112	943.45	19519
		E-D	1801	58.67	900	117.41	450	234.81	225	469.63	112	943.45	
10	104163.2	**H-I**	1902	54.77	930	112.00	455	228.93	227	458.87	111	938.41	17277
		E-D	1775	58.68	887	117.43	443	235.13	221	471.33	110	946.94	
11	107238.9	**H-I**	2002	53.57	979	109.54	473	226.72	232	462.24	115	932.51	17753
		E-D	1827	58.70	913	117.46	456	235.17	228	470.35	114	940.69	
12	91099.4	**H-I**	1591	57.26	795	114.59	399	228.32	197	462.43	98	929.59	14482
		E-D	1552	58.70	776	117.40	388	234.79	194	469.58	97	939.17	
13	84027.6	**H-I**	1556	54.00	759	110.71	369	227.72	182	461.69	89	944.13	13168
		E-D	1432	58.68	716	117.36	358	234.71	179	469.43	89	944.13	
14	84465.7	**H-I**	1514	55.79	736	114.76	361	233.98	179	471.88	89	949.05	15989
		E-D	1439	58.70	719	117.48	359	235.28	179	471.88	89	949.05	
15	85203.9	**H-I**	1592	53.52	777	109.66	378	225.41	184	463.06	91	936.31	14196
		E-D	1452	58.68	726	117.36	363	234.72	181	470.74	90	946.71	

16	81657.8	**H-I**	1479	55.21	735	111.10	357	228.73	177	461.34	86	949.51	13794
		E-D	1391	58.70	695	117.49	347	235.33	173	472.01	86	949.51	
17	75088.4	**H-I**	1365	55.01	672	111.74	326	230.33	161	466.39	79	950.49	12849
		E-D	1279	58.71	639	117.51	319	235.39	159	472.25	79	950.49	
18	65870.9	**H-I**	1181	55.78	593	111.08	295	223.29	142	463.88	70	941.01	11279
		E-D	1122	58.71	561	117.42	280	235.25	140	470.51	70	941.01	
19	63990.7	**H-I**	1160	55.16	582	109.95	285	224.53	139	460.36	68	941.04	10047
		E-D	1090	58.71	545	117.41	272	235.26	136	470.52	68	941.04	
20	71943.0	**H-I**	1358	52.98	659	109.17	316	227.67	155	464.15	76	946.62	12355
		E-D	1226	58.68	613	117.36	306	235.11	153	470.22	76	946.62	
21	71557.1	**H-I**	1306	54.79	638	112.16	308	232.33	153	467.69	77	929.31	12056
		E-D	1219	58.70	609	117.50	304	235.39	152	470.77	76	941.54	
22	61232.6	**H-I**	1114	54.97	551	111.13	270	226.79	132	463.88	65	942.04	10476
		E-D	1043	58.71	521	117.53	260	235.51	130	471.02	65	942.04	
23	52450.7	**H-I**	945	55.50	468	112.07	230	228.05	113	464.17	56	936.62	9210
		E-D	894	58.67	447	117.34	223	235.21	111	472.53	55	953.65	
24	62012.4	**H-I**	1147	54.06	561	110.54	275	225.50	134	462.78	66	939.58	10635
		E-D	1056	58.72	528	117.45	264	234.90	132	469.79	66	939.58	

25	42744.1	**H-I**	782	54.66	387	110.45	188	227.36	92	464.61	45	949.87	7082
		E-D	728	58.71	364	117.43	182	234.86	91	469.72	45	949.87	
26	51452.6	**H-I**	946	54.39	462	111.37	228	225.67	110	467.75	54	952.83	9168
		E-D	876	58.74	438	117.47	219	234.94	109	472.04	54	952.83	
27	45606.4	**H-I**	816	55.89	400	114.02	197	231.50	96	475.07	48	950.13	7746
		E-D	777	58.70	388	117.54	194	235.08	97	470.17	48	950.13	
28	46243.4	**H-I**	850	54.40	423	109.32	208	222.32	100	462.43	49	943.74	7368
		E-D	788	58.68	394	117.37	197	234.74	98	471.87	49	943.74	
29	51082.1	**H-I**	937	54.52	458	111.53	224	228.05	109	468.64	54	945.96	8192
		E-D	870	58.72	435	117.43	217	235.40	108	472.98	54	945.96	
Total	2508095.77	**H-I**	45930	54.69	22538	111.32	11010	227.71	5412	463.98	2669	940.91	427495
		E-D	42734	58.69	21360	117.43	10671	235.06	5329	470.78	2656	944.90	
	MAF	**H-I**		0.4982		0.4993		0.4995		0.4998		0.4995	0.4592
		E-D		0.4617		0.4620		0.4597		0.4598		0.4640	

[1]**BTA=** *Bos Taurus* autosome; **40k=** SNP array with more than 40,000 markers; **20k=** SNP array with more than 20,000 markers; **10k=** SNP array with more than 10,000 markers; **5k=** SNP array with more than 5,000 markers; **2k=** SNP array with more than 2,000 markers; **427k=** SNP array with 427,495 markers; **MAF=** average of minor allele frequency; **H-I**=criteria to select markers with high informativity; **E-D**=criteria to select markers evenly distributed.