scientific reports

OPEN

Variable selection strategies for genomic prediction of growth and carcass related traits in experimental Nellore cattle herds under different selection criteria

Lucio F. M. Mota[1]✉, Leonardo M. Arikawa[1], Júlia P. S. Valente[1,2], Larissa F. S. Fonseca[1], Maria E. Z. Mercadante[2,3], Joslaine N. S. G. Cyrillo[2], Henrique N. Oliveira[1] & Lucia G. Albuquerque[1,3]

Genomic selection (GS) has become a widely used tool in breeding programs, enhancing selection accuracy and leading to faster genetic progress. However, in small populations, GS faces challenges due to limited data and a large number of markers potentially leading to biased predictions. Implementing feature selection strategies is essential to improve prediction accuracy and avoid overfitting. Hence, we compared the predictive ability of genomic best linear unbiased prediction (GBLUP), Bayesian B (BayesB), and elastic net (ENet) models, using all markers and feature selection via GWAS and fixation index (FST) to reduce marker numbers, for growth and ultrasound carcass traits in three Nellore cattle populations differentially selected for yearling body weight (YBW). The populations evaluated included: Nellore Control (NeC), selected for YBW; Nellore Selection (NeS), selected for maximum YBW; and Nellore Traditional (NeT), selected for maximum YBW and lower residual feed intake (RFI) since 2013. Comparing the statistical approaches using GBLUP as the reference, ENet improved prediction accuracy by 10% for growth traits and 12% for carcass traits, while BayesB showed no improvement for growth traits but achieved a 3% gain for carcass traits. When comparing models using all markers to those with variable selection, both GWAS and FST improved prediction accuracy across models, with FST outperforming GWAS in stratified populations. A stricter GWAS threshold (> 1.0% explained variance), compared to a less conservative criterion (> 0.5%), reduced BayesB prediction accuracy (6.8%), while slightly increasing accuracy for GBLUP (1.3%) and ENet (2.4%). Similarly, a more restrictive FST threshold (> 0.2) against a less conservative (> 0.1) resulted in smaller gains for GBLUP (4%) and ENet (5%), but reduced BayesB accuracy (− 4%). Overall, selecting markers through GWAS and FST improves prediction accuracy for both growth and carcass traits, particularly in stratified populations. However, stricter thresholds can negatively impact accuracy, highlighting the need for optimized marker selection strategies.

Keywords Genomic prediction, Genome-wide association, Fixation index, Small population, SNP selection method

Using genomic information in animal breeding has become a standard approach for evaluating and selecting animals. Including this information in genetic evaluations has improved selection accuracy, especially for young animals, compared to traditional pedigree-based evaluation[1,2]. Single-nucleotide polymorphism (SNP) information has provided a new opportunity to predict complex phenotypes accurately by offering a better match of genetic architecture and Mendelian sampling (MS) than traditional pedigree-based methods[3]. SNP markers allow the estimation of genomic-based breeding values (GEBVs), achieving greater genetic gains and

[1]Department of Animal Science, School of Agricultural and Veterinary Sciences, São Paulo State University (UNESP), Jaboticabal 14884-900, SP, Brazil. [2]Institute of Animal Science, Beef Cattle Research Center, Sertãozinho, SP 14174-000, Brazil. [3]National Council for Science and Technological Development, Brasilia, DF 71605-001, Brazil. ✉email: flaviommota.zoo@gmail.com

reducing costs in progeny testing[4]. This cost reduction occurs because genomic information enables early and accurate identification of superior animals, eliminating the need to wait for phenotypic data from offspring. Consequently, fewer animals must be required for expensive and time-consuming progeny performance recording, thereby minimizing the resources necessary for data collection and analysis.

Genomic prediction (GP) relies on a large number of widely distributed markers across chromosomes to ensure that quantitative trait loci (QTLs), which have both small and large effects, can be captured[5,6]. A high density of markers provides a more comprehensive representation of the genetic background of traits, enabling markers to capture a significant proportion of the trait variation and accurately predict breeding value (EBV)[7]. For GS, the genomic best linear unbiased prediction (GBLUP) and single-step GBLUP (ssGBLUP) are traditionally used, based on the assumption that trait variability results from the influence of multiple loci with additive contributions distributed across the genome[8]. However, ssGBLUP has become more widely adopted in recent years because it combines pedigree, phenotypic, and genomic data into a single model, whereas standard GBLUP includes only genotyped animals[9]. This integration raises predictive ability, especially when many animals lack genotypes. Additionally, penalized regressions such as Elastic Net (ENet) and Bayesian B (BayesB) that assign differential weights to genetic markers aim for a better alignment of the SNP contributions to the genetic architecture of the trait[10,11]. However, only a minor fraction of markers explain the additive genetic variance of the trait, and a large number of markers (p) increases computational costs and can create problems due to collinearity among markers[12]. Maintaining a balanced set of predictors on GP is crucial to avoid a negative impact on the models' statistical power[12].

Dealing with high dimensionality is particularly challenging in small populations, such as those of many local breeds, where the number of molecular markers often exceeds the number of individuals. In this context, implementing robust strategies to reduce the number of predictors is essential for improving model stability and predictive performance[13]. Such strategies play a key role in preventing model overfitting, which is commonly associated with high-dimensional predictor information, reducing the number of predictors while ensuring that the model remains reliable and effective[14]. Using a preselected subset of SNPs in GBLUP has shown enhanced prediction accuracy compared to using all genomic information[12,15]. This highlights the importance of the strategy for accurately identifying and selecting relevant markers that match the target trait's genetic architecture, reducing the noise in model[16]. Usually, this process is carried out based on GWAS results[17], where SNPs are ranked according to their association with the phenotype of interest. A common application of GWAS results in GP involves ranking SNPs based on their statistical significance or the proportion of additive genetic variance they explain, and top-ranked markers are then selected and incorporated into prediction models as a reduced SNP panel. However, there are limitations in selecting top markers from GWAS approaches, such as the difficulty in determining the appropriate number of markers required to achieve high GP accuracy for each trait[18], especially under varying genetic architectures[19,20]. Additionally, this approach can lead to biased predictions due to false-positive rates[21].

In populations submitted to different selection criteria, genetic differences arise from increased diversity in the genomic regions directly related to the traits involved in the selection process. These differences can be mapped through the fixation index (FST), which measures the changes in the allele frequencies of divergent markers between populations[22]. The FST is a helpful tool for screening markers under selection by evaluating population differentiation and can provide an efficient criterion for identifying relevant markers for GS[23]. The FST is able to capture the direction of the Mendelian sampling of the QTL in populations under different selection criteria[16,24]. Using prioritized genetic markers from FST for genomic prediction, the prediction accuracy would improve the genomic evaluation power. Hence, we aimed to compare the predictive ability of GS using all the genomic information and preselecting strategies to reduce the number of markers through GWAS and the FST index, three statistical approaches: GBLUP, the penalized regression method Elastic Net (ENet) and Bayesian B (BayesB) for growth and ultrasound carcass traits in Nellore cattle.

Materials and methods

Ethical approval

The animal handling procedure followed ARRIVE (Animal Research: Reporting of In Vivo Experiments) and was approved by the ethical committee procedures Animal Care and Ethical Committee recommendations of São Paulo State University (UNESP), School of Agricultural and Veterinary Sciences (protocol number 18.340/16). All methods were carried out in accordance with relevant guidelines and regulations.

Phenotypic and genotypic information

Animals are from an experimental breeding program at the Beef Cattle Research Center at the Institute of Animal Science (IZ) in Sertãozinho, São Paulo, Brazil. Since the 1980s, three selection herds have been maintained in the IZ: Nellore control (NeC), Nellore Selection (NeS), and Nellore Traditional (NeT). In the NeC herd, animals are selected for yearling body weight (YBW), which is close to the average of the contemporary group, while in the NeS and NeT herds, animals are selected for the maximum yearling body weight (YBW). Since 2013, the NeT herd has also been selected for lower residual feed intake (RFI)[13,25]. In the NeT herd, sires from NeT and NeS herds can be used during the breeding season. In the NeC and NeS, only sires from the same herd are used during the breeding season, with a controlled inbreeding rate. The current inbreeding coefficients are 7.0% in NeC, 4.0% in NeS, and 3.5% in NeT. In the three herds, animal selection is based on body weight (BW_{sel}) measured at 378 days in young bulls after feedlot performance testing (168 days) and at 550 days in heifers raised on pasture. The average daily gain (ADG) was obtained during the feed efficiency performance testing and estimated as the linear regression slope of body weight (BW) on days in the trial. During the trials, animals were weighed without fasting at the beginning and the end of the feeding trial, as well as every 14 days during the experimental period.

The growth and carcass-related traits were evaluated on 1621 animals born between 2004 and 2020, with 262 animals from the NeC herd (221 young bulls and 41 heifers), 458 from the NeS herd (359 young bulls and 99 heifers), and 901 from the NeT herd (697 young bulls and 204 heifers). The carcass traits were measured by ultrasound at yearling in young bulls (n. 1277; 381.5 ± 31.05 days) and heifers (n. 344, 389.7 ± 34.69 days) using the PIE MEDICAL Aquila equipment with a probe of 7 inches and 3.5 MHz. The image analysis was performed through Echo Image Viewer 1.0 (Pie Medical Equipment B.V., Maastricht, The Netherlands), following the criteria of the Ultrasound Guidelines Council. The ribeye area (REA, cm^2) was measured by positioning the ultrasound probe between the 12th and 13th ribs (*longissimus thoracis* muscle) on the left side. The back fat thickness (BF, mm) was also measured on the *longissimus thoracis* muscle at a point three-quarters of the ventral length of the REA. The rump fat thickness (RF, mm) was measured by placing the transducer at the intersection of the *gluteus Medius* and *biceps femoris* muscles between the hook and pin bone.

The contemporary groups (CG) for all evaluated traits were defined by birth year, sex, and selection herd. Quality control for growth and carcass traits was performed by removing values outside the interval of ± 3.5 standard deviations below and above the CG mean. Additionally, only CGs with at least five animals were retained in the dataset, and the number of animals per CG ranged from 5 to 78.

The genotyped population encompass animals from three experimental herds (NeC, NeS, and NeT) with both males and females represented. A total of 780 animals were genotyped using the Illumina Bovine HD BeadChip (770 k, Illumina Inc., San Diego, CA, USA), including 48 males and 25 females from NeC, 109 males and 65 females from NeS, and 255 males and 186 females from NeT. Additionally, 881 animals were genotyped using the GeneSeek Genomic Profiler Indicus (GGP Indicus 50 K), encompassing 163 males and 14 females from NeC, 234 males and 28 females from NeS, and 434 males and 8 females from NeT. Markers situated in non-autosomal regions or having the same genomic coordinates were removed, followed by the application of a quality control (QC) filter to exclude autosomal SNPs with a GenCall score of less than 0.6, thereby removing genotyping problems.

The medium-density genotyped animals were imputed to the HD panel (777k markers) using FImpute v3[26], considering a reference population of 780 animals and an imputation accuracy of 0.98[13]. The genomic QC for imputed animals was performed using the qcf90 programs[27] to remove genetic markers: (a) located on sex and mitochondrial chromosomes, (b) with a call rate < 0.90, (c) with a minor allele frequency (MAF) < 0.05, (d) with deviation from HWE ($P \leq 10^{-5}$) and (e) monomorphic markers. In addition, samples with a call rate < 0.90 and Mendelian conflict were also removed. After quality control of genomic and phenotypic information, a total of 1569 genotyped animals with growth and carcass information genotyped with 384,519 SNP markers remained for GS analyses. Descriptive statistics for growth and carcass-related traits are shown in Table 1. Population structure was evaluated through principal component analysis (PCA) based on the genomic relationship matrix, using the ade4 R package[28] (Additional File 1: Fig. S1).

Genetic parameters and adjusted phenotype

The variance components and phenotypes adjusted for the fixed effects for growth and carcass traits were estimated considering a genomic BLUP (GBLUP) method as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic information for BW$_{\mathrm{Sel}}$, ADG, REA, BF or RF; $\mathbf{b}$ is the vector of fixed effects of CG and as covariate: (1) cow age (linear and quadratic effect) and (2) age at measurement (linear effect) except for ADG; $\mathbf{a}$ represents the additive genetic effect of animals, and $\mathbf{e}$ is the residual effect. $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices associating fixed (**b**) and random effects (**a**) with the phenotypic information (**y**), respectively.

The GBLUP was fitted considering the following assumptions for random effects: $\mathbf{a} \sim (0, \mathbf{G}\boldsymbol{\sigma}^2_{\mathbf{a}})$ and $\mathbf{e} \sim (0, \mathbf{I}\boldsymbol{\sigma}^2_{\mathbf{e}})$, where $\boldsymbol{\sigma}^2_{\mathbf{a}}$ is the additive genetic variance, $\boldsymbol{\sigma}^2_{\mathbf{e}}$ is the residual variance, $\mathbf{G}$ represents the genomic relationship matrix (GRM) obtained according to VanRaden[3]: $\mathbf{G} = \frac{\mathrm{MM'}}{2\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{p_j}\left(1-\mathrm{p_j}\right)}$, where $\mathbf{M}$ is the SNP marker matrix coded as 0, 1, and 2 for genotypes AA, AB, and BB adjusted for allele frequency (2 $\mathrm{p_j}$), $\mathrm{p_j}$ is the second allele frequency of the jth SNP marker, and $\mathbf{I}$ is an identity matrix.

Trait[1]	Mean	min	max	$\boldsymbol{\sigma}^2_{a}$	$\boldsymbol{\sigma}^2_{e}$	$\boldsymbol{\sigma}^2_{p}$	h^2
BW$_{\mathrm{Sel}}$, kg	353.69 ± 50.6	199.27	493.95	555.6 ± 52.91	711.88 ± 41.89	1,267.48 ± 45.02	0.44 ± 0.025
ADG, g/d	0.88 ± 0.18	0.38	1.310	0.025 ± 0.002	0.041 ± 0.003	0.066 ± 0.003	0.38 ± 0.038
REA, cm^2	53.68 ± 8.52	29.00	79.80	15.47 ± 1.84	22.77 ± 1.36	38.24 ± 1.52	0.40 ± 0.031
BF, mm	2.23 ± 1.01	0.40	7.60	0.23 ± 0.029	0.47 ± 0.025	0.70 ± 0.026	0.33 ± 0.033
RF, mm	5.01 ± 1.7	1.60	12.00	0.72 ± 0.083	1.28 ± 0.069	2 ± 0.071	0.36 ± 0.032

Table 1. Descriptive statistics (n. 1569), variance components, and heritability (h^2) estimates for growth and carcass-related traits in Nellore cattle using single-trait genomic-based analyses. [1]BW$_{\mathrm{Sel}}$, Body weight at selection (378 days for young bulls and 550 days for heifers); ADG, average daily gain during the feed trial; REA, rib eye area obtained by ultrasound; BF, subcutaneous backfat thickness obtained by ultrasound; and RF, rump fat thickness obtained by ultrasound.

The model was implemented via Bayesian inference using gibbsf90 + software from the blupf90 family[29]. The Gibbs sample covered a chain of 500,000 cycles, considering a burn-in the first 100,000 iterations and samples stored every five iterations. The posterior means of the genetic parameters were obtained from 80,000 samples. Convergence was evaluated through visual inspection of the trace plot using the BOA package in R[30]and all traits were found to converge ($P > 0.12$) according to the Geweke test[31].

The phenotypic data adjusted for fixed effects ($y^{*} = y - Xb$) was determined using the predictf90 software[29] and was used as the response variable in the genomic prediction. This step was essential to remove systematic environmental effects and ensure comparability across methods. The adjusted phenotypes were used directly as the response variable in GBLUP, BayesB, and ENet. Moreover, in the GBLUP model, variance components were re-estimated to ensure alignment with the variance structure inherent to the adjusted phenotypes.

Validation scenarios

A forward validation scheme was applied to assess the prediction accuracy, splitting the dataset based on birth year, with animals born between 2004 and 2018 assigned as the training population ($n = 1263$) and those born in 2019 and 2020 ($n = 306$) as the validation set. In this study, GP analyses were performed using three different methods: GBLUP, BayesB, and ENet. Adjusted phenotypic records were used as the response variable in the GP for all statistical approaches. Additionally, the same validation design was consistently applied across all approaches to ensure comparability of predictive performance under identical training and testing conditions (Fig. 1).

Genomic prediction (GP) analyses

GBLUP

Genomic prediction for growth and carcass traits considering the GBLUP can be described as follows:

$$\mathbf{y}^{*} = \mu + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}^{*}$ is the vector of adjusted phenotypic information, μ is the overall mean, $\mathbf{Z}$ is the incidence matrix relating observations to GEBV, and $\mathbf{g}$ is a vector of additive genetic effects assumed to follow a normal distribution $N(0, \mathbf{G}\sigma_{g}^{2})$, where σ_{g}^{2} is the genetic variance, and $\mathbf{G}$ is the GRM. The residual effect ($\mathbf{e}$) followed a normal distribution ($N(0, \mathbf{I}\sigma_{e}^{2})$), with σ_{e}^{2} representing the residual variance and $\mathbf{I}$ the identity matrix. The genetic parameters were recalculated in the GBLUP model, considering adjusted phenotypes to align with the variance structure of those adjusted phenotypes. The GBLUP analyses were conducted using the blupf90 + program[29].

BayesB

BayesB considers a linear regression, assuming that a known proportion of SNP markers does not contribute to the genetic variation (i.e., a point of mass at zero) with probability π and a probability of 1-π markers affecting the trait following an univariate t-distribution[8,11]. The adjusted phenotype ($\mathbf{y}_{i}^{*}$) of the ith individual is expressed as $y_{i}^{*} = \mu + \sum_{w=1}^{p} x_{iw}u_{w} + e_{i}$, where μ is the unknown average; x_{iw} is the SNP marker w (coded as 0, 1, and 2) in animal i; u_{w} is the SNP marker effect (additive) of the wth SNP ($p = 384{,}519$); and e_{i} is a residual effect assumed to be normally distributed as $\mathbf{e} \sim N(0, \mathbf{I}\sigma_{e}^{2})$. A priori, the distribution of u_{w} is:

$$p\left(u_{w}|df, \pi, S_{B}\right) = \pi * \left(u_{w} = 0\right) + \left(1 - \pi\right) * t\left(u_{w}|df, S_{B}\right),$$

where π is the known prior probability of the SNP having a null effect, $1 - \pi$ is the probability of the SNP marker having a nonnull effect, and $t\left(u_{w}|df, S_{B}\right)$ is a scaled t distribution with df = 5 degrees of freedom and scale parameter S_B[32]. BayesB was implemented using the R package BGLR version 1.09[32], considering a Gibbs chain of 200,000 iterations, with a burn-in of the first 50,000 iterations and a sampling interval of 10 cycles.

Elastic-net (ENet)

The elastic net is a robust penalized regression that effectively controls the strong collinearity between predictor variables through the combination of two regularization terms: $l_1 = \sum \left|\beta_{j}\right|$(least absolute shrinkage and selection operator – LASSO) and $l_2 = \sum \beta_{j}^{2}$(ridge regression – RR)[10]. The l_1 and l_2 penalty terms are controlled by the alpha parameter (α), providing a balance between selection (LASSO) and shrinkage (RR) of predictor variables (SNP markers). The optimum weight values for λ and α in the ENet are considered in the loss function as follows:

$$L\left(\lambda, \alpha, \beta\right) = \min\left[\frac{1}{2N} \sum_{i=1}^{N} \{y_{i} - \left(\sum_{w=1}^{p} x_{iw}\beta_{w}\right)\}^{2} + \lambda\left(\left(1 - \alpha\right)\beta_{w}^{2} + \alpha\left|\beta_{w}\right|\right)\right],$$

where N is the number of animals, α is the value between 0 (RR penalty) and 1 (LASSO penalty), and λ is a regularization parameter that controls the variable shrinkage. The adjusted phenotype ($\mathbf{y}_{i}^{*}$) was directly used as response variable as $y_{i}^{*} = \mu + \sum_{w=1}^{p} x_{iw}u_{w} + e_{i}$, where μ is the unknown average; x_{iw} is the SNP marker w (coded as 0, 1, and 2) in animal i; u_{w} is the additive SNP marker effect ($p = 384{,}519$); and $e_{i}$is a residual effect.

The ENet model was performed using the h2o R package (https://github.com/h2oai/h2o-3). We performed a random grid search using the h2o.grid function with cross-validation that splits the training subset into five folds for training and testing to find optimal values for α and λ ranging from 0.0 to 1.0 with an interval of 0.1 for each parameter. Finally, the trained model with the highest accuracy and lowest mean square error (MSE) was applied to a disjointed validation set.

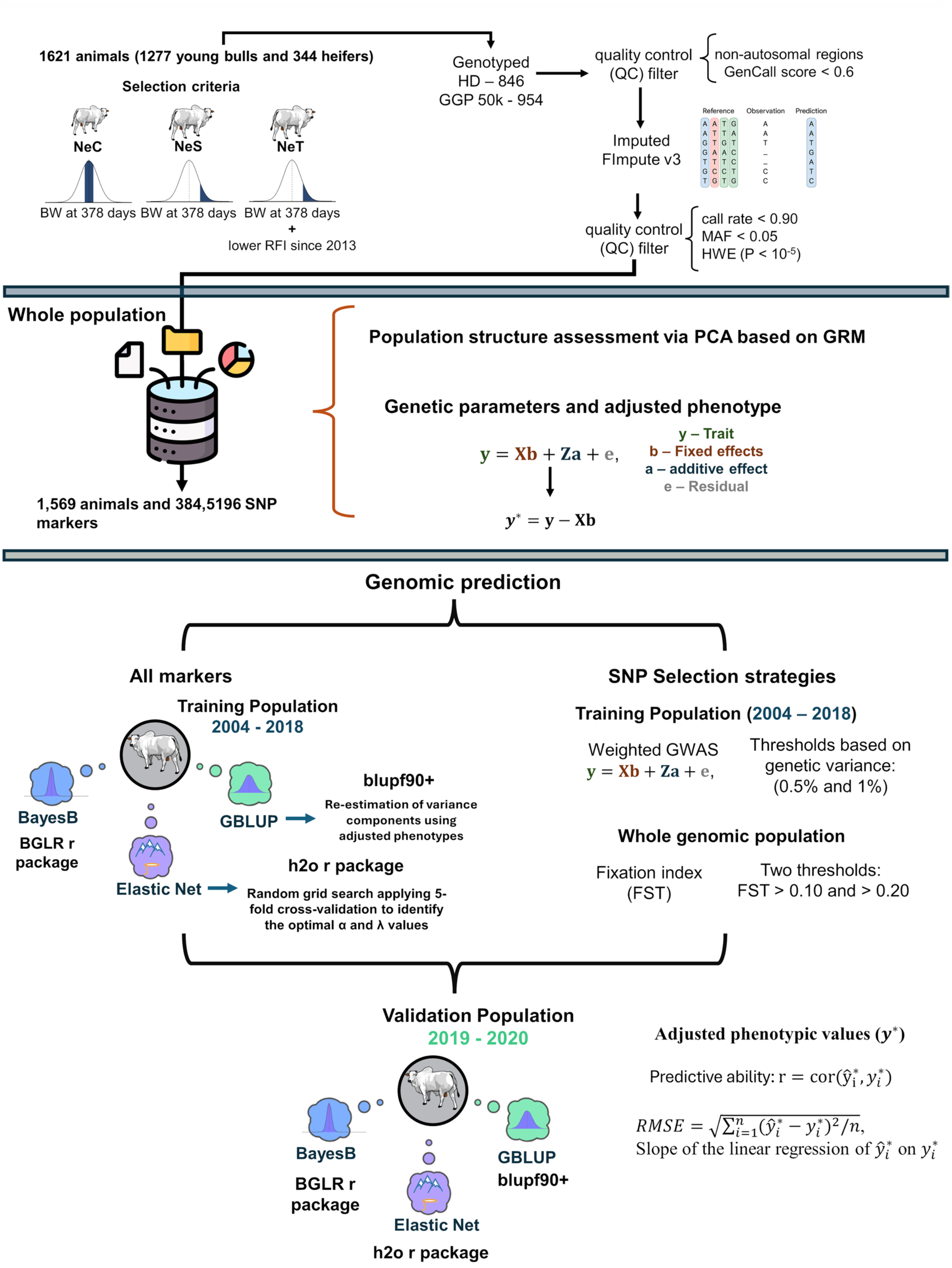

Fig. 1. Schematic representation of the analytical workflow used in this study, from raw data to genomic prediction. Three selection herds were defined: NeC (selection for average yearling body weight, YBW), NeS (selection for maximum YBW), and NeT (selection for maximum YBW and, since 2013, for lower residual feed intake, RFI). A forward validation scheme was applied, splitting the dataset into training (2004–2018) and validation (2019–2020) populations. Genomic prediction models (GBLUP, BayesB, and Elastic Net) were trained using the training set, and SNP selection strategies were applied within it. Predictive performance was then evaluated on the independent validation set.

SNP Selection strategies.

SNP selection based on weighted GWAS

To evaluate the effectiveness of reducing dimensionality and improving prediction accuracy, we preselected SNP markers from weighted GWAS (wGWAS) performed in the training population (t). We considered two thresholds based on genetic variance ($\sigma\,^2_a$) explained by markers for the target trait (0.5% and 1% of $\sigma\,^2_a$). These thresholds were selected because they captured a considerable proportion of the trait's total genetic variance, allowing the identification of the most informative markers while maintaining biological relevance. The wGWAS was performed considering an animal model applied to the training population (t), ($\mathbf{y_t} = \mathbf{Xb_t} + \mathbf{Za_t} + \mathbf{e_t}$). The SNP markers' effects and weights for wGWAS were estimated using the steps proposed by Wang[33], running two iterations to estimate the genetic variance explained by the markers[25,34].

The percentage of genetic variance explained by SNP markers ($\sigma^2_{\widehat{ut}}$) in the training population was calculated as follows: $\sigma^2_{\widehat{ut}} = \frac{Var\left(\sum_{j=1}^{100} Z_j \widehat{u_j}\right)}{\sigma\,^2_{at}} \times 100$, where $\sigma\,^2_{at}$ is the genetic variance for each trait in the training population, $\boldsymbol{Z_j}$ is the vector of the *jth* SNP marker and $\widehat{\boldsymbol{u}}_{\boldsymbol{j}}$ represents the SNP effect of the *jth* SNP within the window with 100 markers. The wGWAS analyses were performed using the postGSf90 program from the BLUPF90 family[29].

SNP selection based on FST

The FST quantifies allele differentiation among herds by identifying genomic segments under selection pressure, and it was used as a strategy for the prioritization of SNP markers in genomic prediction methods. The genetic differentiation among the NeC, NeT, and NeS herds was assessed using the FST methodology as described by Weir and Cockerham[35]using the hierfstat R package[36]. The global FST was calculated using the varcomp.glob function, which estimates variance components between and within populations, with the FST value obtained by the ratio[35]:

$$FST = \frac{\sigma\,^2_{between}}{\sigma\,^2_{Total}},$$

where $\sigma\,^2_{between}$ represents the genetic variance between populations and $\sigma\,^2_{Total}$ the total genetic variance.

Pairwise genetic differences between populations were estimated using the pairwise.WCfst function. The significance of the FST values was assessed by bootstrapping with 1,000 resamplings (boot.ppfst). The FST values were interpreted according to Wright[37]where values greater than 0.15 suggest significant differentiation between populations. We considered two thresholds based on FST (0.1 and 0.2).

Model performance

The predictive ability of the different methods was assessed by Pearson's correlation ($\mathrm{r} = \mathrm{cor}(\widehat{\mathrm{y}_\mathrm{i}}^*, y_i^*)$) between the between phenotypes adjusted for fixed effects (y_i^*) and predicted adjusted phenotype in the validation set ($\widehat{\mathrm{y}_\mathrm{i}}^*$). The predictive root mean squared error (RMSE) was $RMSE = \sqrt{\sum\,_{i=1}^{n}(\widehat{y}_i^* - y_i^*)^2/n}$, where *n* represents the number of animals in the validation set. The slope of the linear regression of $\widehat{y_i^*}$ on y_i^* was also used to assess prediction bias.

Results and discussion

Genetic parameters

Variance components and heritability estimates for growth and carcass-related traits are shown in Table 1. Heritability estimates for carcass-related traits ranged from 0.33 for BF to 0.40 for REA and agreed with those reported by other studies in Nellore cattle, with values varying from 0.18 to 0.50[38–42]. Genomic selection has been widely applied to improve quantitative traits, particularly those that are difficult or costly to measure, and it may be especially helpful for enhancing animal selection for ultrasound carcass traits. The heritabilities observed for $\mathrm{BW}_{\mathrm{Sel}}$ (0.44) and ADG (0.33) were similar to those reported by Benfica et al.[43] (0.43 for $\mathrm{BW}_{\mathrm{sel}}$ and 0.31 for ADG) and Mota et al.[25] of 0.44 for BW at 455 days.

Model performance

Predictive ability was compared with GBLUP as the benchmark against BayesB and ENet for growth and carcass traits, considering a forward validation using all genetic markers (Table 2). The highest predictive ability was obtained with ENet (0.74 to 0.83) followed by BayesB (0.66 to 0.77) and GBLUP (0.65 to 0.74) (Table 2). The values obtained in the present study were greater than those observed by Mehrban et al.[44] for REA (0.45) and BF (0.47) obtained by ultrasound in Hanwoo cattle and were comparable to those reported by Silva et al.[45] in Nellore cattle for REA (0.55 to 0.62), BF (0.57 to 0.60) and RF (0.54 to 0.61). However, the predictive ability obtained with GBLUP and BayesB were lower than those reported by Lopes et al.[46]who evaluated Bayesian regression methods for REA (0.75), BF (0.95) and RF (0.85) in Nellore cattle.

Despite the polygenic nature of the traits evaluated (Additional File : Fig. S2 and S3), BayesB slightly increased predictive ability compared to GBLUP, with improvements of 3% for REA, 1% for BF, and 5% for RF. Similar differences in predictive ability between BayesB and GBLUP have been reported in other studies, where GBLUP or single-step GBLUP were used as benchmarks for ultrasound carcass traits[45,46]. The improvement observed with the BayesB method may be due to its ability to better capture the genetic architecture of ultrasound carcass traits in small populations by applying different weights to markers, thereby identifying the most significant markers associated with each trait[8]. Additionally, as reported by Gao et al.[47]BayesB enhanced predictive ability

Trait[1]	Predictive ability [2]					Root mean square error			Prediction bias[3]		
	GBLUP	BayesB	Enet	RD (%) BayesB	RD (%) Enet	GBLUP	BayesB	Enet	GBLUP	BayesB	Enet
BW_{Sel}	0.69	0.69	0.75	0.0%	8.7%	12.90	18.94	10.1	1.08	0.95	1.01
ADG	0.679	0.679	0.761	0.0%	12.1%	0.03	0.05	0.02	0.97	1.07	1.01
REA	0.728	0.753	0.828	3.4%	13.7%	2.51	3.16	1.97	0.98	1.06	0.99
BF	0.651	0.66	0.745	1.4%	14.4%	0.37	0.5	0.29	0.97	0.95	0.99
RF	0.742	0.777	0.8	4.7%	7.8%	0.22	0.37	0.17	1.06	0.94	1.01

Table 2. Predictive ability, root mean square error (RMSE), and prediction bias for growth (BWSel and ADG) and carcass-related (REA, BF, and RF) traits in Nellore cattle. [1] BW_{Sel}, Body weight at selection; ADG, average daily gain; REA, rib eye area obtained by ultrasound; BF, subcutaneous backfat thickness obtained by ultrasound; and RF, rump fat thickness obtained by ultrasound. [2]Predictive ability was assessed by Pearson's correlation (r) between phenotypes adjusted for fixed effects (y_i^*) and predicted adjusted phenotype ($\widehat{y}_i^*$). RD - relative difference (RD) in prediction accuracy assessed as $RD\ (\%) = \frac{(r_{m1} - r_{GBLUP})}{r_{GBLUP}} \times 100$, where r_{m1} is the predictive ability using BayesB or ENet and r_{GBLUP} is the predictive ability using the GBLUP approach. [3]Slope of linear regression of phenotypes adjusted for fixed effects (y_i^*) on predicted adjusted phenotype ($\widehat{y}_i^*$).

Trait[1]	GBLUP				BayesB				ENet			
	0.50%		1.00%		0.50%		1.00%		0.50%		1.00%	
	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope
BW_{Sel}	0.72	0.97	0.70	0.90	0.70	0.95	0.68	1.12	0.80	0.97	0.78	1.04
ADG	0.70	0.97	0.67	0.89	0.69	0.94	0.65	1.11	0.79	1.01	0.76	1.09
REA	0.74	0.97	0.72	0.92	0.72	1.04	0.67	1.14	0.83	1.02	0.81	1.09
BF	0.66	0.98	0.65	0.94	0.64	1.06	0.62	1.13	0.79	0.99	0.75	1.03
RF	0.75	0.97	0.75	0.92	0.72	0.95	0.69	1.13	0.83	0.99	0.82	1.06

Table 3. Predictive ability considering SNP markers selected from GWAS, explaining more than 0.5% and more than 1.0% of the genetic variance for growth (BW_{Sel} and ADG) and carcass-related (REA, BF, and RF) traits using GBLUP, bayesb, and ENet approaches. Predictive ability, cor ($\widehat{y}_i^*$, y_i^*)) Pearsons' correlation between phenotypes adjusted for fixed effects (y_i^*) and predicted adjusted phenotype ($\widehat{y}_i^*$). BWSel, Body weight at selection; ADG, average daily gain during the feed trial; REA, rib eye area obtained by ultrasound; BF, subcutaneous backfat thickness obtained by ultrasound; and RF, rump fat thickness obtained by ultrasound.

due to its superior capacity to estimate marker effects in smaller datasets. Therefore, possibly the improvements in predictive ability for ultrasound carcass traits (REA, BF, and RF) using BayesB could be attributed to its accurate estimation of marker effects in small datasets, as well as for handling SNPs in high linkage disequilibrium (LD)[48].

The ENet method outperformed GBLUP, improving predictive ability from 8.7% for BW_{Sel} to 14.4% for BF (Table 2). Probably, ENet was able to improve predictive ability because it combines two different types of penalty terms, l_1 (Lasso) and l_2 (ridge), on the predictor variables. The l_1 term reduces non-informative SNP effects exactly to zero, removing noise and performing variable selection[49], while the l_2 term shrinks the remaining, often highly correlated effects toward zero without fully removing them, reducing multicollinearity and stabilizing estimates[50]. Combining these terms in Enet helps to reduce overfitting in the training population by regulating the degree of shrinkage and applying flexible penalties on the predictors, allowing them to reduce to zero or close to zero, reducing the problem of dimensionality[51,52]. This factor enhances the prediction accuracy and simplifies the model, increasing interpretability. In Chinese Simmental beef cattle, Wang et al.[53] reported that the ENet approach slightly increased prediction accuracy for ADG and carcass weight compared to GBLUP and BayesB but decreased accuracy for REA (2.0%) and marbling score (2.17%). The authors concluded that GBLUP is more effective for traits controlled by many genetic markers with minor effects, while ENet can be adaptable and perform well in various circumstances by adjusting penalty values (l_1 and l_2) to the genetic structure of the target trait[20].

	GBLUP				BayesB				Enet			
	0.1		0.2		0.1		0.2		0.1		0.2	
Trait	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope	Acc	Slope
BW_{Sel}	0.75	0.98	0.73	0.97	0.73	0.91	0.7	0.94	0.82	0.99	0.8	1.01
ADG	0.73	0.98	0.71	0.97	0.69	0.92	0.66	0.93	0.82	0.98	0.79	1.01
REA	0.76	0.97	0.74	0.97	0.74	0.93	0.7	0.93	0.87	0.98	0.85	1.00
BF	0.7	0.99	0.68	0.98	0.66	0.93	0.62	0.91	0.81	1.02	0.8	0.99
RF	0.78	0.97	0.76	0.96	0.75	0.93	0.73	0.90	0.86	1.01	0.84	0.98

Table 4. Prediction accuracy considering the SNP markers with a fixation index (FST) higher than 0.1 or 0.2 for growth (BW_{Sel} and ADG) and carcass-related (REA, BF, and RF) traits using GBLUP, bayesb, and enet. Predictive ability, cor $(\widehat{y}_i^*, y_i^*)$) Pearsons' correlation between phenotypes adjusted for fixed effects (y_i^*) and predicted adjusted phenotype ($\widehat{y}_i^*$). BWSel, Body weight at selection; ADG, average daily gain during the feed trial; REA, rib eye area obtained by ultrasound; BF, subcutaneous backfat thickness obtained by ultrasound; and RF, rump fat thickness obtained by ultrasound.

Impact of variable selection from GWAS and FST on prediction accuracy

Selecting informative SNP subsets markedly improved prediction (Tables 3 and 4). The prediction accuracy ranged from moderate to high, with values between 0.62 and 0.83 when markers were selected based on GWAS (Table 3) and between 0.62 and 0.87 when selecting markers based on FST (Table 4) for growth and carcass-related traits. Among the evaluated traits, the highest prediction accuracies were consistently obtained, varying from 0.75 to 0.83 (GWAS) and 0.76 to 0.87 (FST) for RF, and 0.72 to 0.83 (GWAS) and 0.74 to 0.87 (FST) for REA. Overall, the prediction accuracies obtained using selected markers were higher for GBLUP and ENet compared to those using the full set of markers. ENet consistently outperformed the other methods when using marker selection strategy, showing relative improvements of approximately 12.9% and 13% over GBLUP, and 18% and 19% over BayesB, using markers selected based on GWAS and FST, respectively. However, both BayesB and ENet methods assign different weights to markers attempting to match their contributions to the variation in a target trait. The superior performance of the ENet method in both variable selection strategies may be attributed to its greater flexibility in capturing complex relationships between genomic data and target phenotypes. On the other hand, BayesB employs more restricted differential shrinkage, allowing only a few genomic regions to have a large effect on the trait. This can result in reduced accuracy when considering selected markers compared to using all markers[20].

Figure 2a and 3a show the relative gains in prediction accuracy achieved by GBLUP, BayesB, and ENet when using variables selected from GWAS (Additional File 1: Figs. S2 and S3) and FST (Additional File 1: Figs. S4), respectively, compared to applying these statistical methods to the full set of SNP markers. The patterns of relative gains varied depending on the target trait and the model employed. For growth traits, selecting markers that explained more than 0.5% of the genetic variance in GWAS led to increases in prediction accuracy of 3.7% with GBLUP, 1.5% with BayesB, and 5.2% with ENet (Fig. 2a). In the case of ultrasound carcass traits, prediction accuracy improved by 1.4% with GBLUP and 3.3% with ENet (Fig. 2a). In comparison, BayesB exhibited a reduction of approximately 4.9% in prediction accuracy (Fig. 2a).

The observed decreased prediction accuracy with BayesB when employing selected markers can be attributed to its assumption that only a small number of markers significantly affect the evaluated traits. In contrast, most markers have minor or null effects[8]. When selected markers are used, the dataset becomes smaller, potentially excluding markers that may individually have small effects but jointly contribute significantly to the trait's genetic variation. Removing these markers can limit BayesB's ability to capture the entire genetic architecture of the trait accurately and may introduce bias into the model[54,55].

Considering FST-selected markers (>0.1; Additional File 1: Fig. S4), the prediction accuracy increased by 7% with GBLUP and 8% with ENet for all traits, whereas BayesB showed improvements only for growth traits (Fig. 3a). These improvements in prediction accuracy with SNPs selected from GWAS and FST agree with the findings of other studies with different species[15,23,24,56,57]. Akbarzadeh et al.[57] reported that using subsets of SNPs from a standard GWAS analysis in humans (1%, 5%, 10%, and 50% of significant SNPs) slightly improved accuracy with the top 10% and 50% of SNPs but showed reduced accuracy with the top 1% and 5%. In Rendena cattle, applying different SNP weighting strategies improved prediction accuracy for ADG, in vivo carcass fleshiness and dressing percentage than traditional single-step GBLUP, although extreme shrinkage decreased prediction accuracy and led to biased predictions[58]. In addition, Meuwissen et al.[59] reported up to a 10% improvement in prediction accuracy over GBLUP for milk traits and somatic cell count by adjusting weights in the genomic matrix based on GWAS results.

There was a greater improvement in prediction accuracy for growth traits (BW_{Sel} and ADG) and carcass traits (REA, BF and RF), using selected markers by the FST, compared to GWAS, with improvements of 4% with GBLUP, 3% with BayesB, and 4% with ENet (Additional File 1: Fig. S5a,b). This superiority can be attributed to FST's effectiveness in capturing genetic variation (Additional File 1: Fig. S1) and considering allele frequency differentiation across the population under different selection criteria[22]. Combining Yorkshire pig populations, Ye et al.[56] observed greater prediction accuracy for GBLUP when markers were selected by the FST index than when using all the markers, demonstrating their usefulness for multi-population GS. Chang et al.[23] observed improvements in prediction accuracy when the G matrix was weighted based on the FST index, achieving a

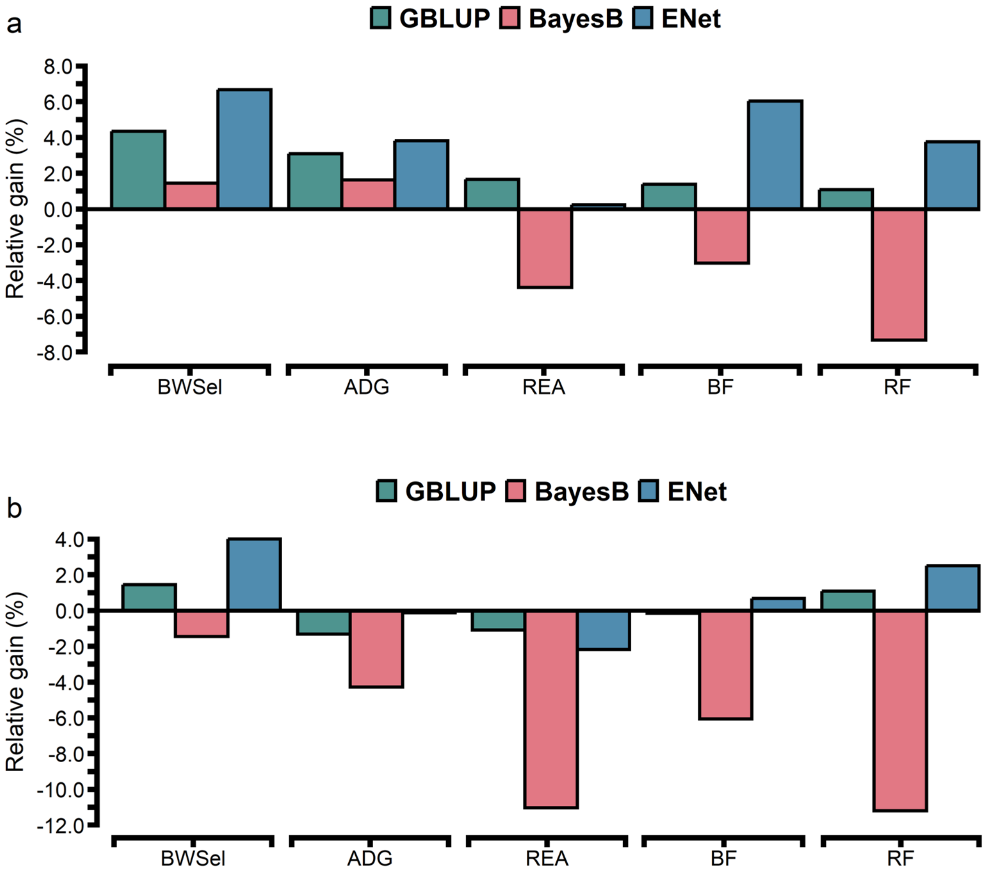

Fig. 2. Relative difference (RD) of the prediction accuracy, considering the SNP markers explaining more than 0.5% (**a**) and 1% (**b**) of genetic variance against all SNPs for growth (BW$_{Sel}$ and ADG) and carcass-related (REA, BF, and RF) traits using GBLUP, BayesB, and Enet.

5% improvement for GBLUP. This variable selection strategy has the power to address heterogeneous and large datasets, providing improvements in prediction accuracy by using the subset of genomic regions that impact the relationships between genotype and phenotype[12,15].

Efficient variable selection from GWAS allows for the identification of SNP markers that are biologically relevant to the target trait (Additional File 1: Fig. S2, S3). On the other hand, selecting variables based on FST helps to identify the major QTLs that differentiate the herds in the study, reflecting the distinct selection strategies applied in each herd (Additional File 1: Fig. S4). Using a GWAS threshold of > 1% of genetic variance (Fig. 2b), only minor improvements in prediction accuracy were observed with GBLUP (1.4% for BWsel and 1.1% for RF) and ENet (4.0% for BWsel, 0.7% for BF, and 0.5% for RF). BayesB exhibited reductions in prediction accuracy ranging from − 1.4% for BWsel to − 11% for REA and RF, while GBLUP showed reductions from − 0.2% for RF to − 1.3% for ADG (Fig. 2b).

Considering a more conservative FST threshold (> 0.2) resulted in slight improvements in prediction accuracy considering the methods GBLUP (1–6%) and ENet (2–7%), whereas BayesB exhibited reductions (3–6%), except for BWsel (Fig. 3b). Moreover, when comparing the thresholds of 0.5% vs. 1.0% for GWAS (Additional File 1: Fig. S6a) and 0.1 vs. 0.2 for FST (Additional File 1: Fig. S6b), the less restrictive threshold provided superior prediction accuracy: approximately 2.3% higher for GWAS and 2.8% higher for FST with GBLUP, 4.8% higher for both GWAS and FST with BayesB, and 3.1% higher for GWAS and 2.5% higher for FST with ENet. These findings suggest that selecting significant markers is more likely to be successful when a less conservative threshold is applied (Figs. 2 and 3). Lopes et al.[46], using a Markov blanket al.gorithm to select markers associated with carcass traits, reported a reduction in prediction accuracy in Bayesian regression approaches, which could be attributed to the reduced degrees of freedom caused by the limited number of markers included in the model. Mota et al.[19] observed that preselecting genetic markers with a less restrictive threshold from GWAS resulted in better performance than considering all markers, and more restrictive thresholds led to a negligible improvement in prediction accuracy.

Decreasing the SNP density for genomic predictions, particularly by applying more conservative thresholds in GWAS (> 1%) and FST (> 0.2), results in decreased predictive ability and increased bias (Tables 2 and 3). This increased bias was particularly notable with BayesB, where the bias ranged from 1.11 to 1.14 for GWAS using markers that explained > 1% of the genetic variance and from 0.90 to 0.94 for FST (> 0.2). In summary, selecting markers using less restrictive thresholds (GWAS > 0.5% and FST > 0.1) can enhance the predictive

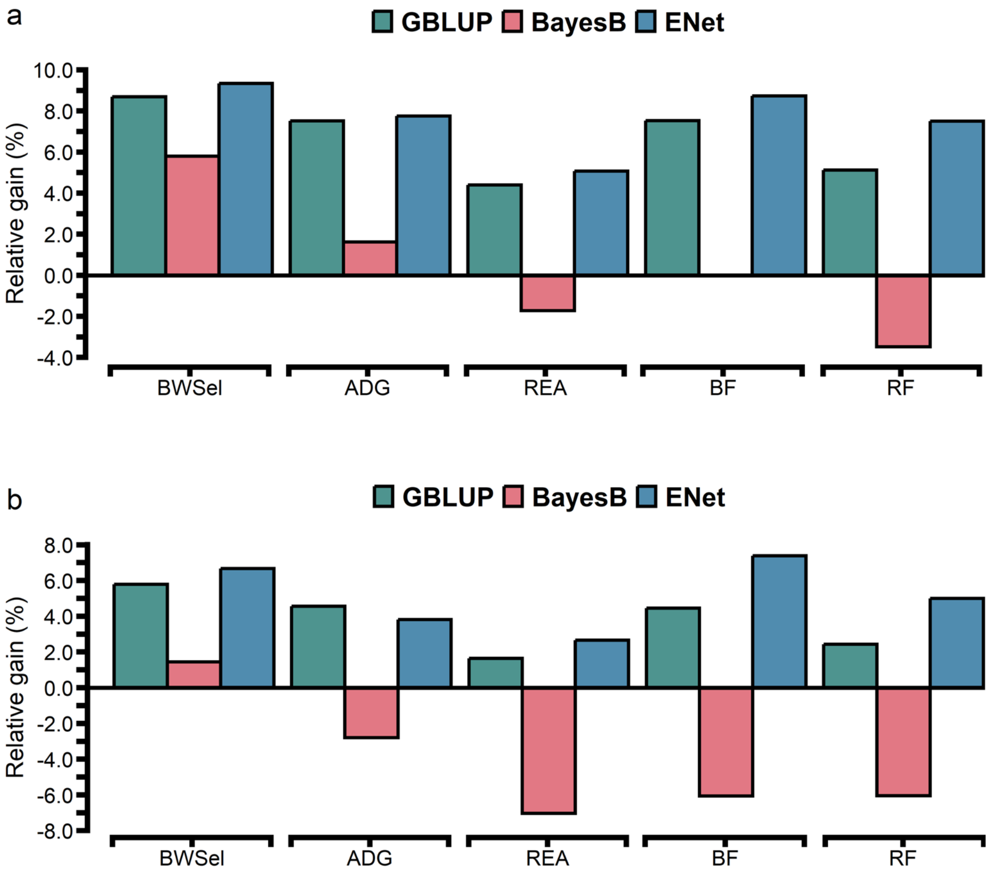

Fig. 3. Relative difference (RD) of the prediction accuracy, considering the SNP markers with a fixation index (FST) higher than 0.1 (**a**) and 0.2 (**b**) against all SNPs for growth (BW and ADG) and carcass-related (REA, BF, and RF) traits using GBLUP, BayesB, and Enet.

ability for growth and carcass traits for GBLUP and ENet. Removing non-informative SNPs improves the resolution of potential causative variants affecting the target trait by removing noise information and, thus, increasing prediction accuracy. Mainly, populations under selection are dynamic, and changes in breeding cycles lead to changes in allele frequency, LD level, and introgression of new alleles caused by selection, genetic drift, and unequal parental contributions to progenies that differentiate populations[60]. Such changes significantly challenge GP across populations, which explains the superiority of selecting markers based on FST against GWAS (Additional File: Fig. S5). Selecting markers based on the FST may be more effective in identifying markers that segregate across the population.

Given the dynamic nature of populations subject to selection, marked by changes in allele frequencies, LD patterns and the introduction of new alleles due to selective pressures and unequal parental contributions, it's essential to assess and consider population structure when using FST-based marker selection. To avoid bias from population structure and accurately capture meaningful genetic differentiation, the presence of subpopulations within the dataset must be evaluated through PCA (Fig. S1a). This analysis aims to uncover genetic differences linked to various selection strategies (NeC, NeS, and NeT), validating genetic stratification across herds. The observed structure supports the application of FST as a method to identify loci that contributes to divergence between groups under different selection criteria. Therefore, an effective workflow for incorporating FST into genomic prediction should include (1) identifying potential subpopulations via PCA or Bayesian clustering (e.g., Admixture), (2) calculating FST between these groups, (3) selecting SNPs that exceed a specific differentiation threshold (e.g., FST > 0.1) or weighing the G matrix, and (4) using these markers in genomic prediction models. This systematic approach aids in prioritizing genomic regions that are both highly differentiated and likely subject to divergent selection, ultimately enhancing predictive ability in structured populations. The improvements observed with FST-selected SNPs across various traits in our study highlight the advantages of this strategy for capturing population-specific genetic signals, especially in breeding programs involving multiple herds or regional subdivisions.

Impact of selected SNP markers on Mendelian sampling

Selecting markers based on different GWAS and FST thresholds impacted the genetic relationship between the training and validation sets (Figs. 4 and 5). The conservative threshold led to the exclusion of markers that, although individually contributing small effects, collectively have a significant impact on the trait's genetic architecture. As a result, the genomic relationship matrix (GRM) built from this limited marker set may not

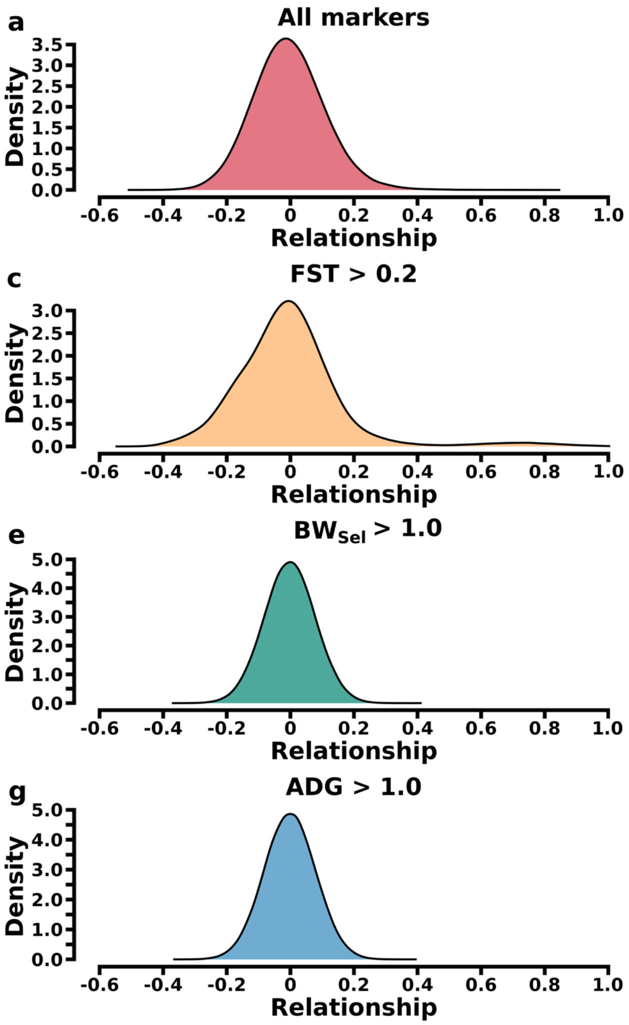

Fig. 4. Genomic relationship between training and validation set considering all genomic information (**a**), selecting markers by fixation index (FST) of 0.1 (**b**) and 0.2 (**c**) and GWAS considering a threshold of 0.5% of genetic variance for BW_{Sel} (**d**) and ADG (**f**), and considering a threshold of 1.0% of genetic variance for BW_{Sel} (**e**) and ADG (**g**). BW_{Sel}, Body weight at selection (378 days for young bulls and 550 days for heifers) and ADG, average daily gain.

adequately capture the genetic relationships between individuals (Figs. 4 and 5), which results in increased prediction bias and decreased accuracy. Selecting markers from FST resulted in a smaller reduction in genetic relationships due to the selection and utilization of a greater number of markers to build the GRM compared to GWAS.

Differences in the prediction accuracy using a selected subset of markers depend on how well it matches the genetics architecture of the trait (Fig. 6) and captures the mendelian sampling (MS), between the training and validation sets (Figs. 4 and 5). Including genomic information into genetic evaluations, compared to pedigree-only approaches, enhances both MS between animals and the accuracy of QTL mapping[24]. While the standard procedure for using genomic markers for GS of complex traits involves considering all SNP markers from chip assay as predictors, focusing on markers directly associated with the target trait could improve predictive performance, as growth and carcass traits are influenced by diverse biological processes[25,61]. Marker selection strategies should aim to preserve the genetic relationship between training and validation sets, as captured by MS, to avoid reductions in prediction accuracy[24].

Multiple factors, such as marker densities, heritability, models, and interaction, seem to impact the prediction accuracy in GS. The heritability estimates for the evaluated traits were influenced by the number of markers selected at each FST and GWAS threshold (Fig. 6). Using selected markers from GWAS, the heritability estimates for carcass traits were lower compared to using all genetic markers, as GWAS tends to capture only the most significant variants associated with the trait, potentially excluding smaller-effect loci that contributes to the overall traits' genetic variation. On the other hand, selecting markers based on FST, which measure population divergence, had heritability estimates closer to those obtained using the full set of markers. This is because FST

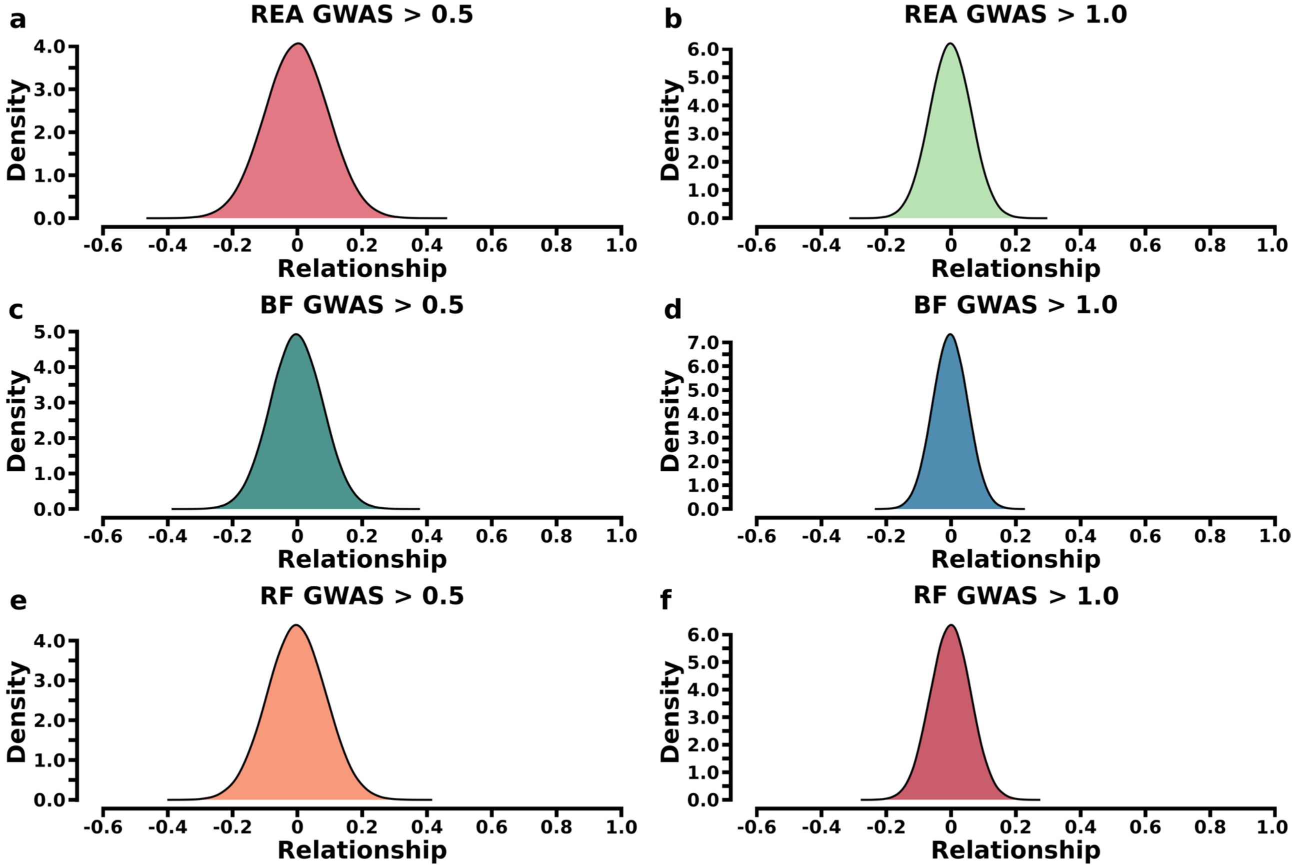

Fig. 5. Genomic relationship between training and validation set selecting markers GWAS considering a threshold of 0.5% of genetic variance for REA (**a**), BF (**c**), and RF (**e**), and considering a threshold of 1.0% of genetic variance REA (**b**), BF (**d**), and RF (**f**). REA, rib eye area obtained by ultrasound; BF, subcutaneous backfat thickness obtained by ultrasound; and RF, rump fat thickness obtained by ultrasound.

captures a broader spectrum of genetic variation, including population-specific differences, which help retain polygenic contributions to the trait.

Applying less conservative thresholds in GWAS (> 0.5% of genetic variance) and FST (> 0.1) improved predictive ability by 7% with GBLUP and 8% with ENet. These gains can occur due to the selection of informative SNPs that capture similar levels of genetic variability and heritability as using all markers. Although medium- to high-density SNP panels are commonly used in GP for beef cattle, our findings suggest that a strategic reduction in marker density through the selection of highly informative SNPs can maintain or even enhance predictive ability. Selected SNPs were also effective at capturing within-family variation and MS effects[19,62]. The results indicated that when properly designed, low-density SNP panels can enhance the predictive ability for growth and carcass traits while substantially reducing genotyping costs. Consequently, selecting informative SNPs to reduce marker density emerges as a viable strategy for cost-effective genomic evaluations in beef cattle. However, the effectiveness of this approach may vary depending on population structure, genetic architecture of the trait, and specific thresholds used for SNP selection. These findings highlight the need for context-dependent strategies when implementing marker reduction protocols in genomic prediction pipelines.

Conclusions

The ENet approach exhibited better performance for genomic prediction of complex traits in Nellore cattle herds, outperforming GBLUP and BayesB approaches across different selection criteria. Preselecting SNPs based on GWAS and FST proved advantageous in filtering out non-informative genetic markers, thereby enhancing prediction accuracy with GBLUP and ENet. However, no benefits were observed with the BayesB method. The FST has been shown to be an effective criterion for selecting markers for genomic prediction, as it better captured QTL similarities between individuals in the training and validation sets, resulting in higher and more stable prediction accuracies compared to GWAS. We showed that preselecting genetic markers with more restrictive thresholds for GWAS (> 1% of genetic variance) and FST (> 0.2) led to a small improvement in the prediction accuracy of growth and carcass traits. These results suggest that careful marker selection, particularly through FST, can optimize genomic predictions, especially in the context of complex trait analyses.

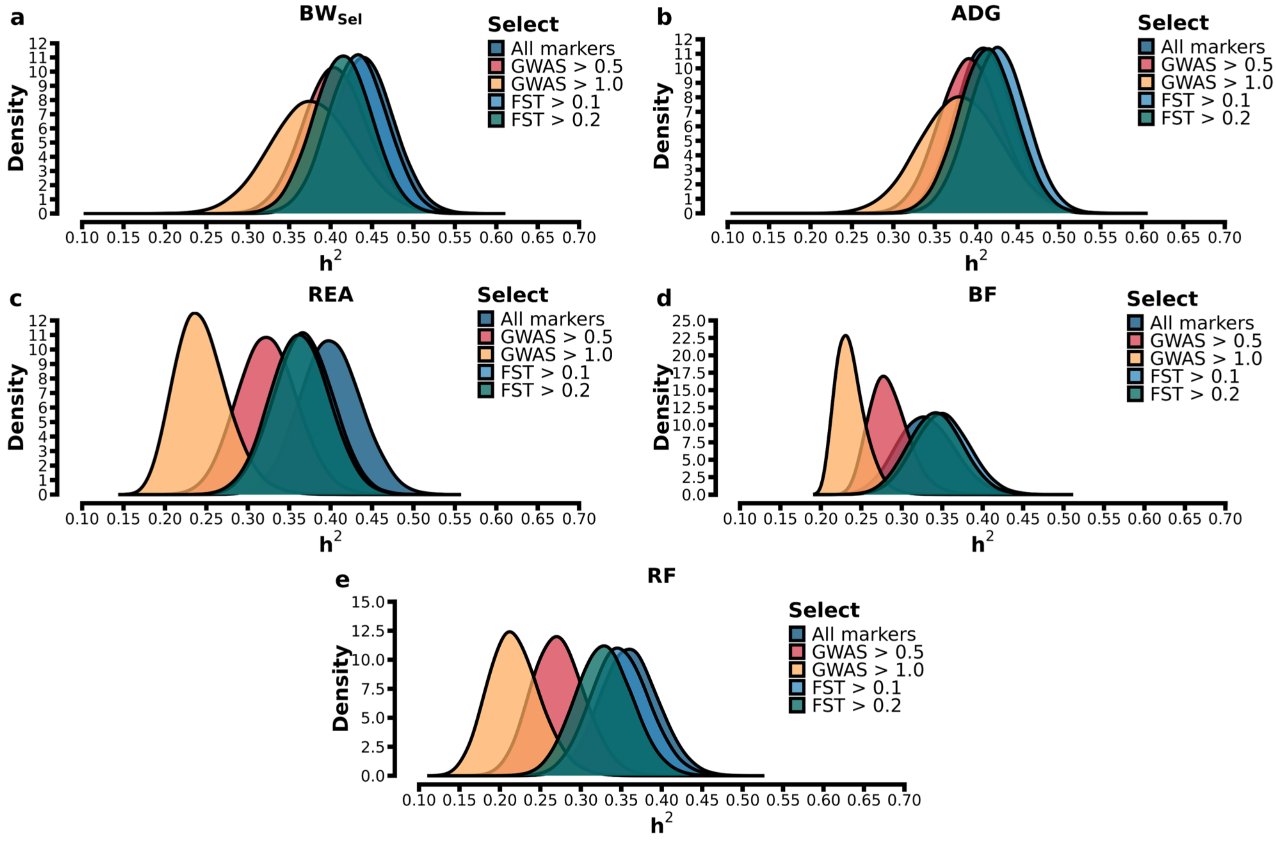

Fig. 6. Estimation of the heritability for BW$_{Sel}$ (Body weight at selection; **a**) and ADG (average daily gain during the feed trial; (**b**), REA (rib eye area obtained by ultrasound; (**c**), BF (subcutaneous backfat thickness obtained by ultrasound; (**d**), and RF (rump fat thickness obtained by ultrasound; (**e**) considering all SNP markers (384,519), and selecting strategies based on GWAS (> 0.5 and > 1.0) and FST index (> 0.1 and > 0.2).

Data availability

The phenotypic and genotypic information related to this study is available for academic purposes upon reasonable request to the authors. The data used in this study belong to experimental breeding programs, and therefore, restrictions apply to their availability. However, the data can be made available by contacting the corresponding authors and obtaining permission from the experimental breeding program. To contact the researchers, please use the email address provided: mezmercadante@gmail.com.

Received: 17 February 2025; Accepted: 11 June 2025
Published online: 01 July 2025

References

1. Fernandes Júnior, G. A. et al. Current applications and perspectives of genomic selection in Bos indicus (Nellore) cattle. *Livest. Sci.* **263**, 105001 (2022).
2. Mrode, R., Ojango, J. M. K., Okeyo, A. M. & Mwacharo, J. M. Genomic selection and use of molecular tools in breeding programs for Indigenous and crossbred cattle in developing countries: current status and future prospects. *Front. Genet* **10**, (2019).
3. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy. Sci.* **91**, 4414–4423 (2008).
4. Zheng, X. et al. Long-term impact of genomic selection on genetic gain using different SNP density. *Agriculture*. **12**, 1463 (2022).
5. Moser, G., Khatkar, M. S., Hayes, B. J. & Raadsma, H. W. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.* **42**, 37 (2010).
6. Zhang, H., Yin, L., Wang, M., Yuan, X. & Liu, X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* **10**, 1–10 (2019).
7. Solberg, T. R., Sonesson, A. K., Woolliams, J. A. & Meuwissen, T. H. E. Genomic selection using different marker types and densities. *J. Anim. Sci.* **86**, 2447–2454 (2008).
8. Gianola, D. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*. **194**, 573–596 (2013).
9. Lourenco, D. et al. Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. *Genes (Basel)*. **11**, 790 (2020).
10. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
11. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157**, 1819–1829 (2001).
12. Piles, M., Bergsma, R., Gianola, D., Gilbert, H. & Tusell, L. Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. *Front. Genet.* **12**, 611506 (2021).

13. Mota, L. F. M. et al. Benchmarking machine learning and parametric methods for genomic prediction of feed efficiency-related traits in Nellore cattle. *Sci. Rep.* **14**, 6404 (2024).
14. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. **23**, 2507–2517 (2007).
15. Mancin, E. et al. Improvement of genomic predictions in small breeds by construction of genomic relationship matrix through variable selection. *Front. Genet.* **13**, 1–25 (2022).
16. Chang, L. Y., Toghiani, S., Aggrey, S. E. & Rekaya, R. Increasing accuracy of genomic selection in presence of high density marker panels through the prioritization of relevant polymorphisms. *BMC Genet.* **20**, 21 (2019).
17. Raymond, B., Bouwman, A. C., Schrooten, C., Houwing-Duistermaat, J. & Veerkamp, R. F. Utility of whole-genome sequence data for across-breed genomic prediction. *Genet. Sel. Evol.* **50**, 1–12 (2018).
18. Jeong, S., Kim, J. Y. & Kim, N. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. *Sci. Rep.* **10**, 1–12 (2020).
19. Mota, L. F. M. et al. Combining genetic markers, on-farm information and infrared data for the in-line prediction of blood biomarkers of metabolic disorders in Holstein cattle. *J. Anim. Sci. Biotechnol.* **15**, 83 (2024).
20. Mota, L. F. M. et al. Genomic prediction of blood biomarkers of metabolic disorders in Holstein cattle using parametric and nonparametric models. *Genet. Sel. Evol.* **56**, 31 (2024).
21. Zhou, G. L. et al. E-GWAS: an ensemble-like GWAS strategy that provides effective control over false positive rates without decreasing true positives. *Genet. Sel. Evol.* **55**, 1–17 (2023).
22. Wright, S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evol. (N Y)*. **19**, 395–420 (1965).
23. Chang, L. Y., Toghiani, S., Hay, E. H., Aggrey, S. E. & Rekaya, R. A weighted genomic relationship matrix based on fixation index (FST) prioritized SNPs for genomic selection. *Genes (Basel)*. **10**, 1–10 (2019).
24. Ling, A. S., Hay, E. H., Aggrey, S. E. & Rekaya, R. Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection1. *BMC Genomic Data*. **22**, 26 (2021).
25. Mota, L. F. M. et al. Meta-analysis across Nellore cattle populations identifies common metabolic mechanisms that regulate feed efficiency-related traits. *BMC Genom.* **23**, 424 (2022).
26. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genom.* **15**, 1–12 (2014).
27. Masuda, Y., Legarra, A., Aguilar, I. & Misztal, I. Efficient quality control methods for genomic and pedigree data used in routine genomic evaluation. *J. Anim. Sci.* **97**, 50–51 (2019).
28. Dray, S. & Dufour, A. B. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
29. Misztal, I. et al. *Manual for BLUPF90 Family of Programs* (University of Georgia, 2018).
30. Smith, B. J. Boa: an R package for MCMC output convergence assessment and posterior inference. *J. Stat. Softw.* **21**, 1–37 (2007).
31. Geweke, J. Clarendon Press,. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. in *Bayesian Statistics* 169–193 (1992).
32. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**, 483–495 (2014).
33. Wang, H. et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Front. Genet.* **5**, 134 (2014).
34. Mota, L. F. M. et al. Integrating genome-wide association study and pathway analysis reveals physiological aspects affecting heifer early calving defined at different ages in Nelore cattle. *Genomics* **114**, 110395 (2022).
35. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
36. Goudet, J. hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol. Ecol. Notes*. **5**, 184–186 (2005).
37. Wright, S. *Evolution and the Genetics of Populations. Vol. 4. Variability Within and among Natural Populations*, vol. 4 (University of Chicago Press, 1978).
38. Caetano, S. L. et al. Estimates of genetic parameters for carcass, growth and reproductive traits in Nellore cattle. *Livest. Sci.* **155**, 1–7 (2013).
39. Bonamy, M. et al. Genetic association between different criteria to define sexual precocious heifers with growth, carcass, reproductive and feed efficiency indicator traits in Nellore cattle using genomic information. *J. Anim. Breed. Genet.* **136**, 15–22 (2019).
40. da Silva Neto, J. B. et al. Weighted genomic prediction for growth and carcass-related traits in Nelore cattle. *Anim. Genet.* **54**, 271–283 (2023).
41. Ceacero, T. M. et al. Phenotypic and genetic correlations of feed efficiency traits with growth and carcass traits in Nellore cattle selected for postweaning weight. *PLoS One*. **11**, e0161366 (2016).
42. de Silva, R. M. Genome-wide association study for carcass traits in an experimental Nelore cattle population. *PLoS One*. **12**, e0169860 (2017).
43. Benfica, L. F. et al. Genetic association among feeding behavior, feed efficiency, and growth traits in growing indicine cattle. *J. Anim. Sci.* **98**, skaa350 (2020).
44. Mehrban, H., Naserkheil, M. & Lee, D. Multi-Trait Single-Step GBLUP improves accuracy of genomic prediction for carcass traits using yearling weight and ultrasound traits in Hanwoo. *Front. Genet.* **12**, 692356 (2021).
45. Silva, R. P. et al. Genomic prediction ability for carcass composition indicator traits in Nellore cattle. *Livest. Sci.* **245**, 104421 (2021).
46. Lopes, F. B. et al. Genomic prediction for meat and carcass traits in Nellore cattle using a Markov blanket algorithm. *J. Anim. Breed. Genet.* **140**, 1–12 (2023).
47. Gao, N. et al. Improving accuracy of genomic prediction by genetic architecture based priors in a bayesian model. *BMC Genet.* **16**, 120 (2015).
48. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning. Springer Series in Statistics* (Springer, 2009). https://doi.org/10.1007/978-0-387-84858-7.
49. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*. **58**, 267–288 (1996).
50. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. **12**, 55–67 (1970).
51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
52. Mota, L. F. M. et al. Real-time milk analysis integrated with stacking ensemble learning as a tool for the daily prediction of cheese-making traits in Holstein cattle. *J. Dairy. Sci.* **105**, 4237–4255 (2022).
53. Wang, X. et al. Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese simmental beef cattle. *PLoS One*. **14**, e0210442 (2019).
54. Mota, L. F. M. et al. Integrating on-farm and genomic information improves the predictive ability of milk infrared prediction of blood indicators of metabolic disorders in dairy cows. *Genet. Sel. Evol.* **55**, 23 (2023).
55. Shi, S. et al. Genomic prediction using bayesian regression models with global–local prior. *Front. Genet.* **12**, 628205 (2021).
56. Ye, S., Song, H., Ding, X., Zhang, Z. & Li, J. Pre-selecting markers based on fixation index scores improved the power of genomic evaluations in a combined Yorkshire pig population. *Animal*. **14**, 1555–1564 (2020).
57. Akbarzadeh, M. et al. GWAS findings improved genomic prediction accuracy of lipid profile traits: Tehran cardiometabolic genetic study. *Sci. Rep.* **11**, 5780 (2021).

58. Mancin, E., Tuliozi, B., Sartori, C., Guzzo, N. & Mantovani, R. Genomic prediction in local breeds: the Rendena cattle as a case study. *Anim. (Basel).* **11**, 1–19 (2021).
59. Meuwissen, T., Eikje, L. S. & Gjuvsland, A. B. GWABLUP: genome-wide association assisted best linear unbiased prediction of genetic values. *Genet. Sel. Evol.* **56**, 17 (2024).
60. Barton, N. H. & Otto, S. P. Evolution of recombination due to random drift. *Genetics.* **169**, 2353–2370 (2005).
61. Arikawa, L. M. et al. Genome-wide scans identify biological and metabolic pathways regulating carcass and meat quality traits in beef cattle. *Meat Sci.* **209**, 109402 (2024).
62. Chen, Z. Q., Klingberg, A., Hallingbäck, H. R. & Wu, H. X. Preselection of QTL markers enhances accuracy of genomic selection in Norway Spruce. *BMC Genom.* **24**, 1–16 (2023).

Acknowledgements

The authors acknowledge the experimental breeding program at IZ-Sertãozinho for providing the dataset used in this research and thank the São Paulo Research Foundation (FAPESP) for their financial support.

Author contributions

L.G.A., M.E.Z.M., and L.F.M.M. conceived and coordinated the study. L.F.M.M. performed the study design. L.F.M.M., J.P.S.V., and L.M.A. contributed to the statistical analysis. L.F.M.M. performed the genotype imputation and led the data analysis and manuscript preparation. M.E.Z.M., J.A.V.S., and H.N.O. contributed to data preparation and analysis. All authors read and approved the final manuscript version.

Funding

The phenotyping and genotyping of the animals used in the research were funded by the Foundation for Research Support of the State of São Paulo (FAPESP 2006/58092-4; 2009/16118–5; 2017/10630–2; 2017/50339-5 and 2018/20026-8) and the scholarship granted to LFMM by FAPESP (2022/11852-7).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-06949-z.

Correspondence and requests for materials should be addressed to L.F.M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.