Received: 12 December 2023 | Revised: 21 May 2024 | Accepted: 26 May 2024

DOI: 10.1111/jbg.12884

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics WILEY

Tag SNP selection for prediction of adaptation traits in Braford and Hereford cattle using Bayesian methods

Fernando A. Reimann[1] | **Gabriel S. Campos**[1] | **Vinícius S. Junqueira**[2,3] | **Helena B. Comin**[1] | **Bruna P. Sollero**[4] | **Leandro L. Cardoso**[4] | **Rodrigo F. da Costa**[1] | **Arione A. Boligon**[1] | **Marcos J. Yokoo**[4] | **Fernando F. Cardoso**[1,4]

[1]Departamento de Zootecnia, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil

[2]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil

[3]Breeding Research Department, Bayer Crop Science, Uberlândia, Minas Gerais, Brazil

[4]Embrapa Pecuária Sul, Bagé, Rio Grande do Sul, Brazil

Correspondence
Fernando A. Reimann, Departamento de Zootecnia, Universidade Federal de Pelotas, Pelotas, RS 96010-900, Brazil.
Email: fe_reimann@hotmail.com

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 305102/2018-4; Empresa Brasileira de Pesquisa Agropecuária, Grant/Award Number: 02.13.10.002

Abstract

This study utilized Bayesian inference in a genome-wide association study (GWAS) to identify genetic markers associated with traits relevant to the adaptation of Hereford and Braford cattle breeds. We focused on eye pigmentation (EP), weaning hair coat (WHC), yearling hair coat (YHC), and breeding standard (BS). Our dataset comprised 126,290 animals in the pedigree. Out of these, 233 sires were genotyped using high-density (HD) chips, and 3750 animals with medium-density (50 K) single-nucleotide polymorphism (SNP) chips. Employing the Bayes B method with a prior probability of $\pi = 0.99$, we identified and tagged single nucleotide polymorphisms (Tag SNPs), ranging from 18 to 117 SNPs depending on the trait. These Tag SNPs facilitated the construction of reduced SNP panels. We then evaluated the predictive accuracy of these panels in comparison to traditional medium-density SNP chips. The accuracy of genomic predictions using these reduced panels varied significantly depending on the clustering method, ranging from 0.13 to 0.65. Additionally, we conducted functional enrichment analysis that found genes associated with the most informative SNP markers in the current study, thereby providing biological insights into the genomic basis of these traits.

KEYWORDS

beef cattle, candidate genes, functional analysis, genomic selection, GWAS, low-density SNP panel

1 | INTRODUCTION

Efficient meat and dairy production in hot and tick-infested tropical and subtropical regions relies on animals well-adapted to their environment. A myriad of complex metabolic pathways determines the production potential of beef cattle. These pathways are activated differently in response to various factors, including nutrition, sanitary conditions, and weather temperature (Santos et al., 2024). Therefore, identifying genomic regions associated with the expression of adaptation traits is necessary for defining breeding goals for developing more adapted animals. Consequently, it significantly enhances production capacity and improves animals' welfare.

Genomic information is a powerful tool in research, enabling us to untangle genetic and environmental factors

associated with adaptation-related traits such as hair coat and eye pigmentation. In this regard, GWAS is a method that can significantly enhance our understanding of the genetic architecture of these traits. By capitalizing on the linkage disequilibrium between genes and markers, GWAS allows us to assess the effect of SNPs. This assessment of the association between SNPs and genes leads to the identification of informative markers potentially involved in metabolic pathways (Mountjoy et al., 2021). These SNP markers often tag candidate genes, and in some cases, custom-built low-density (LD) panels can be constructed. An LD panel is a collection of key SNP markers that can explain a portion of the genetic variation associated with traits of interest. If an LD panel can demonstrate a higher prediction ability than a higher-density SNP panel, it can serve as an economical and promising alternative for genomic selection.

This study aims to perform a GWAS using a Bayesian framework to identify the genomic regions and Tag SNPs linked with eye pigmentation, weaning and yearling hair coat, and breeding standards for the Hereford and Braford breeds. The practical application of the Tag SNP approach will be verified by custom-building an LD panel and assessing its predictive ability performance compared to the higher-density SNP panels. Furthermore, we aim to identify candidate genes by conducting a functional enrichment analysis for adaptation traits.

2 | MATERIALS AND METHODS

2.1 | Phenotype, genotype, and pedigree

The data used were provided by the Conexão Delta G Breeding Program, Rio Grande do Sul, Brazil, and derived from a sample of 126,290 animals (28,392 Hereford and 97,898 Braford), including pedigree ancestors without records. The traits investigated were hair coat at weaning (WHC; $n=81{,}043$), hair coat at yearling (YHC; $n=41{,}390$), and eye pigmentation (EP; $n=73{,}615$), and to the breed standard (BS; $n=28{,}186$), where n is the number of recorded animals. Animals were evaluated between 100 and 300 days of age at weaning and 360 and 670 days at yearling. The WHC and YHC are recorded as 1 (short), 2 (medium), and 3 (long); EP is recorded from 1 (absent), 2 (partial), and 3 (total); and BS is evaluated and scored at yearling, ranging from 1 to 5, with the highest score awarded to animals with better conformation.

A total of 3983 samples were used for the analysis, consisting of 3750 animals genotyped with the Illumina BovineSNP50 (50 K) Bead Chip (54,562 SNPs) and 233 sires genotyped with the Illumina High-Density (HD) Beef Chip Array (777,962 SNPs). The SnpStats R package version 1.39.0 (Clayton, 2023) was utilized to implement genotype quality control (QC). Samples with a call rate below 0.90, heterozygosity exceeding three standard deviations above or below the observed mean, mismatching sex, and duplicated records were removed. In addition, SNPs were discarded based on call rate (<0.98), minor allele frequencies (MAF) (<0.03), and Hardy–Weinberg equilibrium chi-square test ($P \leq 10{-}7$), genotypes highly correlated ($r > .98$), and SNPs in the same position. Markers in the HD chip were subset to only those SNPs in both panels (HD and 50 K). Missing genotypes were imputed across breeds using the FImpute software v3 (Sargolzaei et al., 2011). After QC, 41,011 SNPs and 3951 animals (3020 Braford and 934 Hereford) remained for further analysis.

2.2 | Bayesian GWAS

Phenotypes used in the GWAS were derived from estimated breeding values (EBV) for WHC, YHC, EP, and BS. These EBVs were obtained from a univariate threshold pedigree-based animal model using the THRGIBBS1F90 software (Misztal et al., 2002). The EBV for all genotyped animals was deregressed before the GWAS (Garrick et al., 2009). The GWAS analysis was done to estimate marker effects under the Bayes A and Bayes B models using the GenSel software version 4.0 (Fernando & Garrick, 2008). The statistical model used for the Bayesian analyses was as follows:

$$\mathbf{y} = \sum_{i=1}^{K} \delta_i \mathbf{z_i} a_i + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes (DEBV); $K = \{1, \ldots 41,011\}$ is the number of SNPs; δ_i indicates whether SNP i is included in ($\delta_i = 1$) or excluded ($\delta_i = 0$) from the model for a given MCMC iteration; $\mathbf{z}_i$ is a vector of genotypes of the fitted SNP i, coded −10/0/10; a_i is the random substitution effect of the fitted SNP i with its own variance $\sigma^2_{a_i}$ and an a priori zero effect with probability π or a non-zero effect with probability $1-\pi$, and $\mathbf{e}$ is the vector of normally distributed random residuals. Each SNP is assumed to have a locus-specific variance obtained from an inverted chi-square distribution in Bayes A and Bayes B. For Bayes A, δ_i is always 1. For Bayes B, the parameter π was assumed to be equal to 0.99. We utilized the default values from Gensel, where the degrees of freedom (v) for the SNP variance were set to 4. The prior distribution for the residual variance had $v = 10$ degrees of freedom. A chain size of 42,000 samples was used, and the first 2000 were discarded as burn-in. The convergence of MCMC was accessed by the Geweke test using the software R in package boa (Smith, 2007).

2.3 | Top windows and tag SNP selection

Identifying the most informative 1 Mb genomic windows (Top Windows) for the studied traits utilized the method described in Sollero et al. (2017). Firstly, we selected windows that accounted for the target proportion of genetic variance for each trait in the Bayes B analysis, with a set threshold of $\pi = 0.99$. Assuming an equal contribution of all genomic regions, the selection criterion was established based on the percentage of genetic variance explained by each window. The selection threshold was set by considering windows that explained genetic variance exceeding five times the expected average genetic variance for each window. The expected average genetic variance was calculated as 100% divided by the number of windows analyzed (2519), multiplied by five (0.2%). Therefore, windows exhibiting a genetic variance higher than 0.2% were deemed putative Quantitative Trait Loci (QTL) and were subsequently chosen for further analysis.

To create a custom LD panel, we utilized four parameters: model frequency (MF), *t*-like statistic (TL), linkage disequilibrium, and minor allele frequency (MAF). We selected the top SNPs with the highest MF within each window and added additional SNPs with MF values above the minimum observed MF value for the top SNPs from the top windows. Based on MF, we also chose SNPs with TL values exceeding the minimum TL value within the pre-selected SNPs. To avoid redundancy, we removed SNPs with high linkage-disequilibrium by retaining only the SNP with the highest MAF when two SNPs had an r^2 value higher than 0.4.

2.4 | Prediction accuracy of tag SNP panels

The custom-built LD panel was used to assess its prediction accuracy in genomic modeling. A cross-validation strategy was designed to evaluate the prediction accuracy of the selected tag SNP panels for the studied traits and, therefore, their utility for selection. We used two approaches to create clusters to represent two scenarios of genomic selection in which animals are either more genetically distant or closely related to the reference population. Both clustering strategies used in this study divided the genotyped animals with phenotypes into four groups to ensure an adequate number of individuals in each group. The first approach used K-means clustering, which relied on a genetic dissimilarity matrix based on the number of non-identical-by-state SNP markers between animals. This approach yielded unbalanced groups with high relationships within clusters but low relationships between clusters. This strategy resulted in minimum/maximum animals in each cluster being 414/1438, 255/839, 358/890, and 145/520 for WHC, YHC, EP, and BS, respectively. The second strategy used random clustering. In this case, groups had balanced numbers with 807 animals for WHC, 546 for YHC, 639 for EP, and 378 for BS. Each one of these groups was used as a validation set, while the three others were used to compose the training set.

The SNP effects were estimated in each training set and used to predict direct genomic values (DGV) for the excluded animals in the cluster. These DGVs were also estimated using the same cross-validation strategy (K-means and random) with Bayes B ($\pi = 0.99$), considering all 41,011 markers to compare the relative efficiency of reduced Tag SNP panels. Prediction accuracies were measured as the genetic correlation between EBV and DGV:

$$\begin{bmatrix} \mathbf{L} \\ \text{DGV} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_1} & 0 \\ \mathbf{0} & \boldsymbol{Z_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha_1} \\ \boldsymbol{\alpha_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{m_1} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \end{bmatrix},$$

where $\mathbf{L}$ is the vector of EBV liabilities, an underlying unobservable normal variable related to WHC, YHC, EP, and BS. Also, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, and $\boldsymbol{W_1}$ are known incidence matrices for the $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, and $\boldsymbol{m_1}$ effects, respectively. Above, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are vectors of systematic effects. The vector in $\boldsymbol{\beta_1}$ included effects of contemporary groups farm, year and season of birth, sex, and management groups, except for EP, for which season of birth and management group were not included. In addition, linear regression was included for the additive and maternal breed composition, individual and maternal heterozygosity, and linear and quadratic regression for animal age. Breed composition was derived from the pedigree. For DGV, $\boldsymbol{\beta_2}$ is considered only a mean effect. Furthermore, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are vectors of random direct additive genetic effects for the two traits, $\boldsymbol{m_1}$ is the vector of random maternal additive genetic effects (included only for WHC), and $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$ are vectors of random errors for the two traits.

A normal prior distribution with a zero mean and a large variance σ^2_β was assigned to $\boldsymbol{\beta} = [\boldsymbol{\beta_1} \ \boldsymbol{\beta_2}]' \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_\beta\right)$, the systematic effects. The prior distributions for direct and maternal genetic effects and the error distribution were:

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha_1} & \boldsymbol{\alpha_2} \end{bmatrix}' \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{\alpha_1} & \sigma_{\alpha_1\alpha_2} \\ \sigma_{\alpha_1\alpha_2} & \sigma^2_{\alpha_2} \end{bmatrix} \otimes \mathbf{A}^*\right),$$

$$\mathbf{m_1} \sim N\left(\mathbf{0}, \mathbf{A}^*\sigma^2_{m_1}\right), \text{and}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} & \boldsymbol{\varepsilon_2} \end{bmatrix}' \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{\varepsilon_1} & 0 \\ 0 & \sigma^2_{\varepsilon_2} \end{bmatrix} \otimes \mathbf{I}\right).$$

where $\boldsymbol{A}$ is a pedigree numerator relationship matrix, in which covariances between individuals in different clusters were set to zero (Saatchi et al., 2011):

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{A}_{44} \end{bmatrix}$$

where $\boldsymbol{A}_{ii}$ ($i=1, 2, 3, 4$) is the relationship numerator matrix for the cluster i, the covariance matrices for additive genetic $\begin{bmatrix} \sigma^2_{\alpha_1} & \sigma_{\alpha_1\alpha_2} \\ \sigma_{\alpha_1\alpha_2} & \sigma^2_{\alpha_2} \end{bmatrix}$ and residual $\begin{bmatrix} \sigma^2_{\varepsilon_1} & 0 \\ 0 & \sigma^2_{\varepsilon_2} \end{bmatrix}$ effects were assumed to be distributed, a priori, as inverted Wishart, while the maternal genetic variance $\sigma^2_{\boldsymbol{m}_1}$ as an inverted scaled chi-square random variable.

All model parameters, including components of (co) variance, were estimated by Bayesian inference using the THRGIBBS1F90 software with a length of MCMC 60,000 cycles, with the first 10,000 cycles discarded (burn-in) and a thinning interval of 10 cycles. Therefore, 5000 samples were used for inference. Convergence was evaluated using graphical inspection (trace plots) and the Geweke criterion.

The posterior mean of genetic correlation was calculated as $r_{\alpha_1\alpha_2} = \sigma_{\alpha_1\alpha_2} / \sqrt{\sigma^2_{\alpha_1} \times \sigma^2_{\alpha_2}}$, and it was used to assess the prediction accuracy of a given SNP panel.

2.5 | Functional enrichment analysis

SNPs included in the potential QTL windows were considered more informative, having a greater probability of being within or near coding regions. According to genome annotations at the National Center for Biotechnology Information (NCBI) and the Medical Subject Headings (MeSH) database, these regions were explored as candidate genes. Genes potentially related to BS and adaptive traits (WHC, YHC, and EP) were subjected to functional analyses using the *Meshr* package available in the R environment. Using the lists of more informative SNPs, the Bovine ARS-UCD1.2 genome assembly was adapted as a reference to map annotated genes within ±100 kb from the location of each SNP (gene list).

The association of a particular MeSH term was analyzed using Fisher's exact test, and the degree of enrichment of a given MeSH descriptor was calculated as the probability that terms (and genes) were detected in the gene list more often than expected by chance. The MeSH terms with a *p-value* < 0.05 were considered statistically significant in the overrepresentation analysis (ORA). The ORA was conducted using the MeSH package in the R software. The MeSH categories evaluated were (1) anatomy, (2) diseases, (3) chemicals and drugs, and (4) phenomena and processes.

3 | RESULTS AND DISCUSSION

3.1 | Genome-wide association study (GWAS)

The GWAS, using all 40,011 SNPs with Bayes B ($\pi = 0.99$), allowed the identification of top windows and Tag SNPs for the adaptation traits. The parameter $\pi = 0.99$ was used, corresponding to 1% of SNPs (on average) fitted in the model, for example, around 410 SNPs at each iteration. The SNP heritability (h^2) estimated using Bayes B values for WHC, YHC, EP, and BS were, respectively, 0.56, 0.65, 0.80, and 0.39.

Figure 1 presents the genetic variance percentage explained by the 2417 1-Mb windows using all 40,011 markers for WHC, YHC, EP, and BS. Only windows that explained more than 0.2% of the genetic variance were used for further analysis. For WHC, 77 windows with 1362 SNPs were selected, representing 32% of the total variance. The top 1% of 1-Mb windows, mapped on BTA 3, 5, and 20, explained 4.78% of the total variance of the trait (Table 1). For YHC, 74 windows containing 1296 SNPs were selected, accounting for 37.79% of the genetic variance. The top 1% of 1-Mb windows for YHC, mapped on BTA 1, 7, 11, 15, 18, and 22, explained 9.78% of the total variance (Table 1). For WHC, in the present study, we found informative windows on BTA20. This window explains more than 1% of the genetic variation. WHC and YHC were found to have a genetic correlation of 0.62 in our population, but only two windows were shared between these traits. These windows in common were identified at the 23 Mb and 38 Mb positions of BTA chromosomes 18 and 20, respectively.

A total of 102 windows containing 1804 SNPs were identified for EP, and these windows explain 50.75% of the total variance. The top 1% 1-Mb windows, as listed in Table 1, are located on BTA 5, 6, 9, 14, 18, and 22. Interestingly, on BTA22, five windows explain a target proportion of the total variance, indicating that this chromosome may contain putative candidate genes associated with the phenotypic expression of this trait.

We identified 37 windows with 652 SNPs that accounted for 20.12% of the total variance and were selected as the top windows for BS. While the top 1% of these

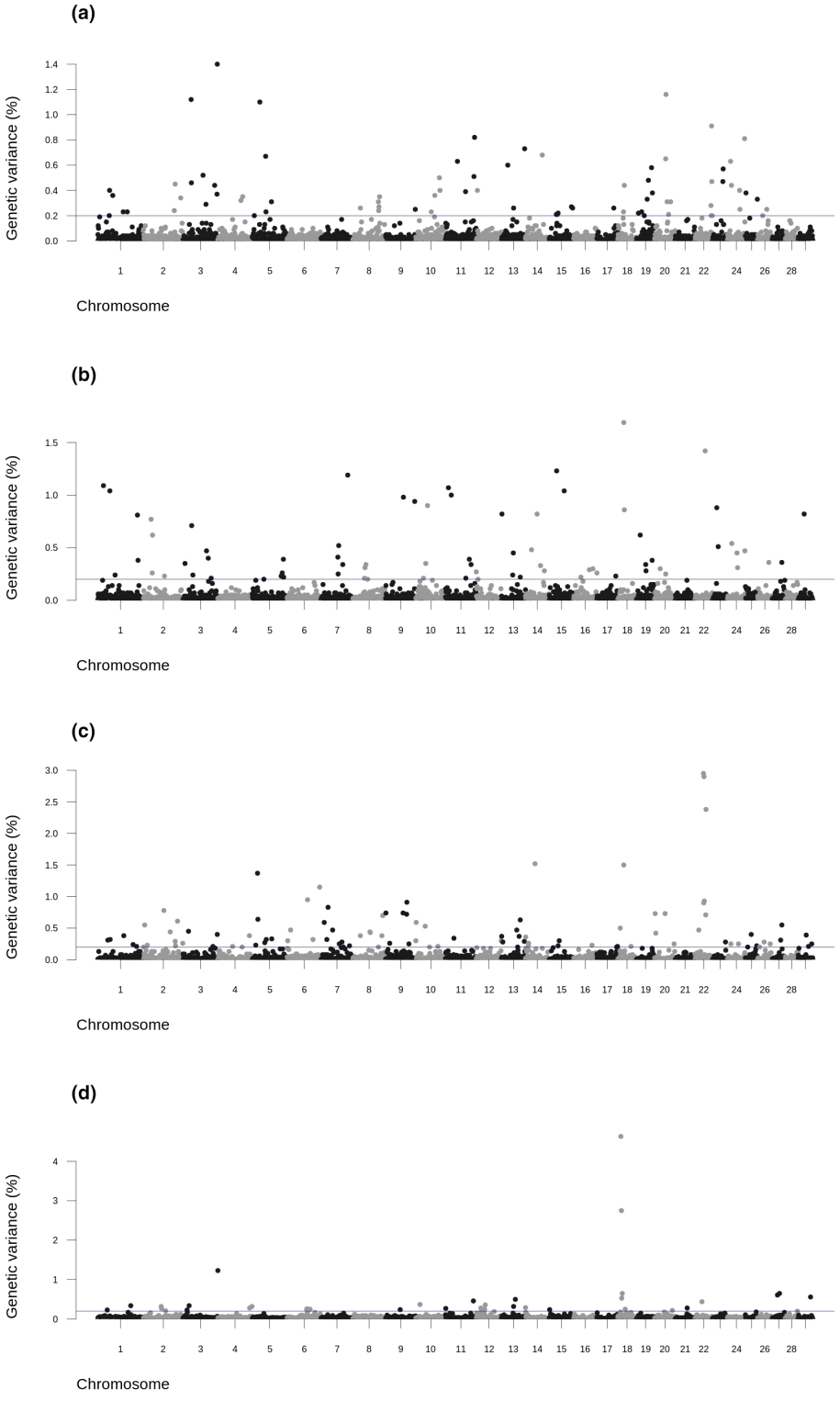

FIGURE 1 Manhattan plot for the genome-wide association study (BayesB $\pi = 0.99$) for hair coat weaning (a), hair coat yearling (b), eye pigmentation (c), and breed standard (d). The Y-axis represents the proportion of genetic variance explained by 1 Mb windows, and the X-axis represents the chromosomes where windows are located.

windows were located on BTA 3 and 18, they only explained 8.61% of the total variance. Additionally, we found that the KIT gene on chromosome 6 is associated with the typical Hereford coat coloration, which is considered when assigning scores for the BS trait. This region may be a potential QTL for BS. We also identified three windows on chromosome 6 related to BS but explained less than 0.26% of the trait's genetic variance.

TABLE 1 Top 1-Mb windows[a] that explain >1.0% of genetic variance for adaptation traits of Hereford and Braford breeds.

Traits	Chromosomes	Chromosome_Mb	1 Mb window number	Number of SNPs	% Genetic Variance
WHC	3	3_118	414	17	1.40
	20	20_39	1978	27	1.16
	3	3_28	324	5	1.12
	5	5_26	563	14	1.10
YHC	18	18_21	1830	18	1.69
	22	22_35	2116	20	1.42
	15	15_26	1595	23	1.23
	7	7_92	868	10	1.19
	1	1_20	20	11	1.09
	11	11_10	1217	22	1.07
	1	1_42	42	21	1.04
	15	15_52	1621	14	1.04
EP	22	22_29	2110	13	2.95
	22	22_31	2112	14	2.90
	22	22_38	2119	15	2.38
	14	14_33	1519	15	1.52
	18	18_21	1830	18	1.50
	5	5_18	555	7	1.37
	6	6_112	770	13	1.15
BS	18	18_11	1820	18	4.63
	18	18_13	1822	12	2.75
	3	3_120	416	19	1.23

Abbreviations: BS, breed standard; EP, eye pigmentation; WHC, hair coat at weaning; YHC, hair coat at yearling.

[a]Each chromosome was divided into one megabase non-overlapping windows, containing a variable number of SNPs with the BayesB method and $\pi = 0.99$; Chromosome_Mb = position the chromosome; 1 Mb window number = the number of windows; Number of SNPs = Number of SNPs in the window; % Genetic Variance = percentage of genetic variance explained by the window.

3.2 | Tag SNP selection

In Table 2, filtering Tag SNPs within the top windows resulted in subsets of markers with consistently larger MF and TL parameters. For WHC, the first step involved identifying all 1362 SNPs in the top windows. In step 2, we used the maximum MF value within the 77 top windows to identify the most informative SNP. The lowest MF threshold was set at 0.097, and we selected 91 SNPs across all top windows. In step 3, we used the TL parameter and its minimum value (0.898), considering only the SNPs selected from the previous step to exclude 15 SNPs due to redundancy ($r^2 > 0.4$) with another marker with higher MAF within the same top window (step 4). The final Tag SNP list for WHC had 105 markers. We followed the same procedure to select Tag SNPs for YHC, EP, and BS.

In selecting Tag SNPs, the MF parameter plays a vital role. This parameter indicates the proportion of posterior samples that include the SNP in the model. Markers with larger MF values are more informative as they correlate highly with the variance the marker explains. SNPs with higher MF are likely to be associated with the trait of interest. Several Bayesian GWAS studies show that SNPs with MF values ≥0.90 should be considered relevant, while MF values ≤0.10 suggest a high false positive rate. In our study, most of the selected Tag SNPs fell in the MF range between 0.10 and 0.90 (Table 2). We found five Tag SNPs with an MF value greater than 0.90 for EP, which were included in more than 90.0% of the MCMC samples.

3.3 | Accuracy of genomic predictions of tag SNP panels

The number of Tag SNPs varied between traits and clusters (Table 3). The SNP heritability with Bayes A using

TABLE 2 Number of selected markers (N. SNP), minimum, mean, and maximum values obtained for the *Model Frequency* and *t.like* parameters at each selection step (steps 1 to 4) of the most informative markers after GWAS for adaptation traits.

Trait	SNP selection	N. SNP	*Model frequency*			*t.like* (TL)		
			Min	Mean	Max	Min	Mean	Max
WHC	Step1	1362	0.005	0.031	0.904	0.001	0.544	1.950
	Step2	91	0.097	0.292	0.904	0.898	0.995	1.950
	Step3	120	0.051	0.401	0.904	0.898	0.997	1.950
	Step4	105	0.051	0.257	0.904	0.898	0.982	1.950
YHC	Step1	1296	0.005	0.035	0.892	0.001	0.526	1.860
	Step2	104	0.081	0.293	0.892	0.861	1.006	1.860
	Step3	153	0.028	0.215	0.892	0.861	0.965	1.860
	Step4	129	0.028	0.236	0.325	0.861	0.977	1.860
EP	Step1	1804	0.004	0.043	0.988	0.001	0.571	3.643
	Step2	132	0.146	0.407	0.988	0.894	1.132	3.643
	Step3	224	0.035	0.209	0.988	0.894	1.04	3.643
	Step4	188	0.035	0.300	0.988	0.894	1.06	3.643
BS	Step1	652	0.005	0.020	0.555	0.001	0.525	1.104
	Step2	70	0.032	0.108	0.555	0.842	0.910	1.104
	Step3	92	0.017	0.088	0.555	0.842	0.897	1.104
	Step4	63	0.018	0.097	0.555	0.842	0.903	1.104

Abbreviations: BS, breed standard; EP, eye pigmentation; WHC, hair coat at weaning; YHC, hair coat at yearling.

the Tag SNPs panel selected in each clustering strategy was lower than with Bayes B ($\pi = 0.99$) considering the full-marker panel. Conversely, in Bayes B ($\pi = 0.99$), only 1% of the markers are fitted in the model, which may be more suitable for QTL discovery, as applied in the present study.

The cross-validation accuracies for all traits were consistently greater when clusters were formed randomly than with K-means. These results are summarized in Table 4. We expected higher accuracy with random clustering due to a closer genomic relationship between training and validation sets. However, this method could result in overestimated accuracy since individuals from the same family could simultaneously be in the training and testing sets. On the other hand, K-means accuracies may be more consistent because clusters were created to reduce the degree of genetic distance between training and testing sets. Lower accuracies are expected for such predictions. The accuracies for WHC and YHC varied between 0.18 and 0.26 for K-means clustering and between 0.33 and 0.55 for random groups with Bayes A and Tag SNP panel. When using the HD panel, the accuracies were 0.33 (WHC) and 0.39 (YHC) with K-means and 0.66 (WHC) and 0.68 (YHC) for randomly formed groups. The accuracies for EP were found to be 0.42 and 0.61 using K-means and random clustering methods, respectively, for the Tag SNP panel. The HD panel with Bayes B showed higher accuracies of 0.75 for K-means and 0.78 for random groups. Among the traits studied, BS had the lowest accuracy, ranging from 0.13 to 0.23. All traits showed higher accuracy when using SNP panels of higher density than Tag SNPs. It is recommended that further investigations be conducted to explore alternative options to improve the prediction accuracy through LD panels. One such alternative could be the single-step GBLUP method, designed to leverage all pedigree, phenotypic, and genomic data simultaneously (Aguilar et al., 2010). Another alternative worth considering is exploring the single-step Bayesian regression method (SSBR), which can assign different weights to marker effects, which could prove advantageous in this case.

3.4 | Functional enrichment analysis

A total of 341 genes were identified and mapped to the most informative SNPs for WHC and YHC. Additionally, 517 genes were identified for the EP, which included the top windows (potential QTLs). The aim was to investigate functional associations by considering four MeSH categories (anatomy, diseases, chemicals and drugs, phenomena, and processes). Detailed information about SNP windows, chromosomes, positions, and top SNPs associated with candidate genes from functional analysis for WHC, YHC, and EP can be found in Tables S1–S3. The PRLR gene (prolactin receptor) in the Bovine Chromosome 20 (BTA20) was identified within one top window and significant MeSH terms for hair coat, such

TABLE 3 Groups of animals according to the cross-validation strategy, K-means (cl) or random (ran), mean of proportion variance explained by markers (h^2), number of top windows (TopW), number of SNPs in each window (TopSNPs) and number of Tag SNPs for each adaptation trait.

		Groups							
Trait	**Parameters**[a]	**cl1**	**cl2**	**cl3**	**cl4**	**ran1**	**ran2**	**ran3**	**ran4**
WHC	h^2_BB99	0.52	0.56	0.60	0.54	0.55	0.54	0.57	0.56
	h^2_BA	0.52	0.54	0.50	0.48	0.51	0.49	0.53	0.55
	TopW	71	51	48	55	54	56	54	66
	TopSNPs	1231	916	835	1007	933	1038	931	1232
	TagSNPs	94	66	48	55	73	74	68	85
YHC	h^2_BB99	0.65	0.52	0.51	0.57	0.53	0.56	0.57	0.57
	h^2_BA	0.48	0.50	0.54	0.59	0.52	0.54	0.54	0.53
	TopW	40	47	42	59	45	44	50	46
	TopSNPs	704	853	773	1082	786	813	881	833
	TagSNPs	78	65	82	94	70	61	64	51
EP	h^2_BB99	0.84	0.65	0.70	0.71	0.73	0.74	0.73	0.72
	h^2_BA	0.77	0.59	0.62	0.66	0.66	0.68	0.65	0.68
	TopW	62	48	60	65	62	44	50	65
	TopSNPs	1055	848	1044	1186	1063	813	881	1136
	TagSNPs	81	78	75	117	77	91	84	92
BS	h^2_BB99	0.69	0.65	0.67	0.67	0.69	0.67	0.68	0.66
	h^2_BA	0.51	0.37	0.54	0.57	0.52	0.51	0.47	0.49
	TopW	13	18	25	22	18	17	12	19
	TopSNPs	881	346	457	422	367	343	229	340
	TagSNPs	22	20	28	38	18	20	20	22

Abbreviations: BS, breed standard; EP, eye pigmentation; WHC, hair coat at weaning; YHC, hair coat at yearling.

[a]h^2_BB99: mean of proportion variance explained by markers using BayesB ($\pi = 0.99$); h^2_BA: mean of proportion variance explained by markers using BayesA; TopW: number of windows that explained above 0.2% of the genetic variance with BayesB ($\pi = 0.99$) GWAS analysis in each group; TopSNPs: number of SNPs included in each top window; TagSNPs: number of SNPs selected according to the parameters *model frequency* and *t-like* statistics, linkage disequilibrium and minor allele frequency.

as Receptor-cell surface, climate, and photoperiod associated. This candidate gene is close to the slick locus on chromosome 20, which is related to the slick coat phenotype of the Senepol breed of cattle and is associated with topcoat length in Brangus (Mariasegaram et al., 2007). The p.Leu462* mutation at this gene may confer additional thermotolerance to cattle beyond its effects on short coat length. The *EIF2AK4* gene was also pointed out as influencing hair coat traits (Edea et al., 2018). Comparative genome-wide analyses have detected positive selection signatures in Ethiopian and Asian *Bos indicus* and *Bos taurus* cattle, indicating that the *EIF2AK4* is associated with cellar stress/thermal tolerance and DNA damage repair. The *OLR1* gene was cited as necessary for adaptation in Dall sheep.

Several genes, including *GUCA1A*, *HTR4*, *PDE6A*, and *RHO*, have been associated with significant MeSH terms related to hair coat absorption, physicochemical properties, photolysis photoreceptor cells, photochemical processes, protein transport, scattering, radiation, and ultraviolet rays (Gautier et al., 2009; Huang et al., 1995; Kamenarova et al., 2013). Mutations in the *PDE6A* gene have been identified in several families with segregating autosomal recessive retinitis pigmentosa. Rhodopsin (RHO) is a biological pigment found in the retina rods and is essential in rod and cone visual phototransduction. Additionally, two novel *GUCA1A* mutations have been identified in two Spanish families with an autosomal dominant retinal degeneration with cone and rod involvement. The *HTR4* gene, a serotonin receptor that participates in the serotonergic system, has also been associated with these genes.

The top EP windows are associated with genes like *KITLG* and *MITF*. These highlighted genes may play central roles in the migration of melanoblast cells and melanocyte development. Gene categories such as muscle, smooth, vascular, intramolecular oxidoreductases, and organ specificity are associated

TABLE 4 Estimated genetic correlations for adaptation traits between liability breeding values and direct genomic value predictions from cross-validation using different methods.

Traits	Methods[a]			Tag SNP vs. all SNP[b]
	Cluster	BayesA_TagSNP	BayesB	
WHC	K-means	0.18 ± 0.03	0.33 ± 0.04	55%
	random	0.33 ± 0.06	0.66 ± 0.06	50%
YHC	K-means	0.26 ± 0.10	0.39 ± 0.15	67%
	random	0.55 ± 0.09	0.68 ± 0.11	81%
EP	K-means	0.42 ± 0.11	0.75 ± 0.10	56%
	random	0.61 ± 0.12	0.78 ± 0.10	78%
BS	K-means	0.21 ± 0.04	0.22 ± 0.05	95%
	random	0.13 ± 0.05	0.23 ± 0.05	56%

Abbreviations: BS, breed standard; EP, eye pigmentation; WHC, hair coat at weaning; YHC, hair coat at yearling.

[a]BayesA_TagSNP: Bayesian model with *t*-distribution and probability ($\pi = 0$), using only the more informative markers; BayesB: Bayesian mixture of *t*-distribution and point mass on zero with probability ($\pi = 0.99$), using all markers (41,011).

[b]Tag SNP vs. all SNP: % of the full SNP panel accuracy retained by the tag SNP panel.

with *HSP90B*, *LPAR1*, and *SLC2A1*. Heat shock proteins (Hsps) play an essential role in the immunity of organisms, particularly in heat resistance (Chen et al., 2013). Lysophosphatidic acid (LPA) is a bioactive lipid that may act in an autocrine manner, binding its specific receptors, particularly *LPAR1*, to modulate downstream signaling pathways and cellular functions. Some results support the important role of LPA in the integrity of retinal pigment epithelium (RPE). The bovine ocular squamous cell carcinoma primarily occurs in cattle lacking eye pigmentation. The *MIF*, *PRKG1, STAR, PLCB1*, and *TWIST2* genes were also suggested in the functional analysis for eye pigmentation. These genes are involved with the physiological pathways of the vitiligo condition, mechanisms of smooth muscle contraction, including vasoconstriction and vasodilation, and processes related to the reproductive functions of cattle (Cardona et al., 2014; Khan et al., 2020).

As discussed earlier, some genes are related to hair coat and eye pigmentation traits. These genes have physiological and behavioral interactions that are associated with adaptation. Although more research is needed to understand better the genetic architecture of these complex quantitative traits, the present functional analysis has identified candidate genes for eye pigmentation and hair coat. These genes can guide more efficient selection for animals less susceptible to serious eye diseases and better adapted to the tropics.

4 | CONCLUSION

Bayesian GWAS analyses have proven useful in selecting Tag SNPs for adaptation-related traits in Hereford and Braford cattle breeds. By employing a 50 K SNP panel, it is possible to achieve a sufficient level of predictability, which enables breeders to select more suitably adapted Hereford and Braford cattle for tropical environments. The Bayesian methods are particularly effective in identifying the most informative genomic regions and Tag SNP markers, which can assist in discovering candidate genes through MeSH terms. Our function analysis suggested that top informative SNP regions are associated with candidate genes annotated in genes reported in the literature.

AUTHOR CONTRIBUTIONS

FR analyzed the data and wrote the manuscript; GC analyzed the data, wrote the manuscript, and reviewed the final version; VJ contributed to the design of the initial analyses, wrote the manuscript, and reviewed the final version; HC performed the functional analysis; BS reviewed the manuscript; LC reviewed the manuscript; RC reviewed the manuscript; AB reviewed the manuscript; MY reviewed the manuscript; FC designed the analyses, reviewed the final manuscript. All authors approved the final manuscript.

ACKNOWLEDGEMENTS

F.F. Cardoso is a research fellow of CNPq. The authors acknowledge the Conexão Delta G (Hereford and Braford) for providing data for this research.

FUNDING INFORMATION

Research funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number 305102/2018–4 and Empresa Brasileira de Pesquisa Agropecuária, Grant/Award Number 02.13.10.002.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support this study's findings are available from the Conexao Delta G breeding program. Restrictions apply to the availability of these data, which were used under license for this study. Data may be obtained from the authors upon request with the permission of the owner breeders.

ORCID

Fernando A. Reimann https://orcid.org/0000-0002-4464-5295

Gabriel S. Campos https://orcid.org/0000-0002-7459-824X
Vinicius S. Junqueira https://orcid.org/0000-0001-7883-1902
Marcos J. Yokoo https://orcid.org/0000-0003-3821-1826

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

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. A., Yngvadottir, B., Shwe, M. T. T., Wee, J., Romero, I. G., & Raj, S. (2014). Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS One*, *9*(5), e98076.

Chen, Z.-Y., Gan, J.-K., Xiao, X., Jiang, L.-Y., Zhang, X.-Q., & Luo, Q.-B. (2013). The association of SNPs in Hsp90*β* gene 5′ flanking region with thermo tolerance traits and tissue mRNA expression in two chicken breeds. *Molecular Biology Reports*, *40*, 5295–5306.

Clayton, D. (2023). *snpStats: SnpMatrix and XSnpMatrix classes and methods*. Bioconductor.

Edea, Z., Dadi, H., Dessie, T., Uzzaman, M., Rothschild, M. F., Kim, E. S., Sonstegard, T., & Kim, K. S. (2018). Genome-wide scan reveals divergent selection among taurine and zebu cattle populations from different regions. *Animal Genetics*, *49*(6), 550–563.

Fernando, R. L., & Garrick, D. J. (2008). *GenSel-user manual for a portfolio of genomic selection related analyses*. Animal Breeding and Genetics, Iowa State University, Ames. bigs.ansci.iastate.edu/bigsgui

Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, *41*, 1–8.

Gautier, M., Flori, L., Riebler, A., Jaffrézic, F., Laloé, D., Gut, I., Moazami-Goudarzi, K., & Foulley, J.-L. (2009). A whole genome Bayesian scan for adaptive genetic divergence in west African cattle. *BMC Genomics*, *10*, 1–18.

Huang, S. H., Pittler, S. J., Huang, X., Oliveira, L., Berson, E. L., & Dryja, T. P. (1995). Autosomal recessive retinitis pigmentosa caused by mutations in the *α* subunit of rod cGMP phosphodiesterase. *Nature Genetics*, *11*(4), 468–471.

Kamenarova, K., Corton, M., García-Sandoval, B., Jose, F.-S., Panchev, V., Ávila-Fernández, A., López-Molina, M. I., Chakarova, C., Ayuso, C., & Bhattacharya, S. S. (2013). *Novel GUCA1A mutations suggesting possible mechanisms of pathogenesis in cone, cone-rod, and macular dystrophy patients*. BioMed Research International, 2013.

Khan, A., Khan, M. Z., Umer, S., Khan, I. M., Xu, H., Zhu, H., & Wang, Y. (2020). Cellular and molecular adaptation of bovine granulosa cells and oocytes under heat stress. *Animals*, *10*(1), 110.

Mariasegaram, M., Chase, C., Jr., Chaparro, J., Olson, T., Brenneman, R., & Niedz, R. (2007). The slick hair coat locus maps to chromosome 20 in Senepol-derived cattle. *Animal Genetics*, *38*(1), 54–59.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., & Lee, D. (2002). *BLUPF90 and related programs (BGF90)*. Paper Presented at the Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., & Karim, M. A. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, *53*(11), 1527–1533.

Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., Taxis, T. M., Chapple, R. H., Ramey, H. R., & Northcutt, S. L. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, *43*, 1–16.

Santos, C. A. D., Eler, J. P., Oliveira, E. C. D. M., Espigolan, R., Giacomini, G., Ferraz, J. B. S., & Paim, T. D. P. (2024). Selective signatures in composite Montana tropical beef cattle reveal potential genomic regions for tropical adaptation. *PLoS One*, *19*(4), e0301937.

Sargolzaei, M., Chesnais, J., & Schenkel, F. (2011). FImpute-an efficient imputation algorithm for dairy cattle populations. *Journal of Dairy Science*, *94*(1), 421.

Smith, B. J. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, *21*, 1–37.

Sollero, B. P., Junqueira, V. S., Gomes, C. C., Caetano, A. R., & Cardoso, F. F. (2017). Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. *Genetics Selection Evolution*, *49*, 1–15.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Reimann, F. A., Campos, G. S., Junqueira, V. S., Comin, H. B., Sollero, B. P., Cardoso, L. L., da Costa, R. F., Boligon, A. A., Yokoo, M. J., & Cardoso, F. F. (2024). Tag SNP selection for prediction of adaptation traits in Braford and Hereford cattle using Bayesian methods. *Journal of Animal Breeding and Genetics*, *00*, 1–10. https://doi.org/10.1111/jbg.12884