UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS

CÂMPUS JABOTICABAL

Relatório Final

Pós-Doutorado FAPESP

Bolsa no País – Regular – Pós-Doutorado (Processo: 2020/03975-6)

Incorporando variantes de sequência para melhorar a predição genômica de características reprodutivas em bovinos Nelore

Pós-Doutorando: Gabriel Soares Campos
Supervisor: Profa. Dra. Lucia Galvão de Albuquerque

Vigência: 01/01/2021 a 31/12/2022
Período: 01/01/2021 a 31/12/2022

Jaboticabal – SP

2023

Sumário

1. Resumo do projeto proposto

Espera-se que as variantes causais selecionadas de dados de sequência aumentem a acurácia dos valores genéticos genômicos. O principal objetivo deste estudo será avaliar a viabilidade do uso de dados de sequência genômica completa (WGS) para predição genômica de características reprodutivas em bovinos Nelore usando diferentes métodos. Além disso, realizaremos um estudo de associação genômica ampla (GWAS) usando com o objetivo de selecionar variantes para características reprodutivas. As características reprodutivas a serem consideradas são: idade ao primeiro parto, probabilidade de prenhez aos 18 meses, reconcepção de novilhas, habilidade de permanência ou "stayability" e número de partos até os 53 meses de idade. Foram utilizados registros fenotípicos medidos em aproximadamente 250.000 animais Nelore provenientes do programas de melhoramento Aliança Nelore. Um conjunto de aproximadamente 42.788 animais genotipados com o ensaio Illumina Bovine HD, bem como a sequência do genoma completo de 150 touros importantes foram usadas para o estudo. A imputação de todos os animais genotipados de HD para WGS foi realizada utilizando o software FImpute. Depois disso, o painel imputado foi aplicado em um GWAS para selecionar variantes para uso posterior na predição genômica. Para predição genômica, testamos dois cenários: 1) usando painel HD SNP e 2) aplicando painel HD mais variantes de sequência selecionadas pelo GWAS. Além disso, a acurácia de predição foi comparada usando o método de passo único GBLUP (ssGBLUP) com matriz de parentesco genômica (**G**) não ponderada e ponderada, com pesos não lineares. Para avaliar a qualidade da predição genômica a partir de diferentes modelos, a validação foi realizada para fêmeas jovens e touros jovens, sem fenótipos nos dados reduzidos, mas com pelo menos um registro nos dados completos, usando um método baseado em regressões lineares denominado LR. O ganho de acurácia com métodos de predição genômica também será comparado ao BLUP tradicional de pedigree.

2. Realizações no período coberto pelo presente relatório

O presente relatório científico se refere ao projeto de pesquisa processo nº 2020/03975-6, realizado no período de 01 de janeiro de 2021 até 31 de dezembro de 2022 pelo bolsista Gabriel Soares Campos. O projeto de pesquisa foi realizado na Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Campus de Jaboticabal (FCAV/Unesp) sob a orientação da Profa. Dra. Lucia Galvão de Albuquerque e coorientação do pesquisador Dr. Roberto Carvalheiro. O objetivo deste item é descrever

as atividades científicas e acadêmicas realizadas durante o período, bem como relatar os resultados obtidos e apresentados em congressos.

Durante este período, o bolsista desenvolveu diversas atividades fundamentais ao andamento do projeto temático "Aspectos genéticos da qualidade, eficiência E sustentabilidade da produção de carne mm animais da raça Nelore" (Processo FAPESP: 2017/10630-2) no qual o presente projeto faz parte. No que se refere as participações em trabalhos em andamento, o pós-doutorando auxiliou na orientação dos trabalhos do grupo de pesquisa que utilizam sequência completa do genoma. No total, o bolsista ajudou dois alunos de pós-doutorado com processo já encerrado, auxiliou três alunos de doutorado e um aluno de mestrado nas quais as análises, dissertação e teses ainda estão em andamento. Além da participação direta nesses trabalhos, o pós-doutorando também atuou em trabalhos de todo o grupo, principalmente com relação ao desenvolvimento e orientação de análises de bioinformática, como alinhamento, montagem e análises das sequencias de DNA como "variant calling", imputação de animais genótipados com chip de alta densidade para a sequência completa e administração e manutenção dos servidores do laboratório de bioinformática. O bolsista também foi responsável, junto à sua supervisora, pela coordenação do funcionamento e dos trabalhos realizados no laboratório de bioinformática, nos quais as atividades relacionadas ao projeto temático 2017/10630-2 se desenvolvem.

Em termos de produção e divulgação de trabalhos científicos, o bolsista publicou um artigo em revista científica indexada e outros quatro artigos no qual o bolsista auxiliou estão em fase de preparação, pois os alunos estão trabalhando com os dados de sequenciamento completo. Um resumo foi publicado em anais de congresso internacional e outro em congresso nacional como primeiro autor. Outros dois resumos foram publicados em anais de congressos nacionais e internacionais. Além disso, o bolsista participou de um evento científico internacional e outro nacional, ambos com a apresentação de trabalhos.

Abaixo, estão listados, os trabalhos desenvolvidos durante o período da bolsa:

2.1 PRODUÇÃO CIENTIFICA

2.1.1 Participação em evento científico

18 e 19 de Outubro de 2021. Participou do XIV Simpósio Brasileiro de Melhoramento Animal (SBMA), promovido de forma remota Universidade Federal de Santa Catarina, realizado em Florianópolis, Santa Catarina (Documento 1 e 2).

01 a 08 de Julho de 2022: Participou presencialmente do "12th World Congress on Genetics Applied to Livestock Production" (WCGALP) promovido pela Wageningen University realizado em Rotterdam, Netherlands (Documento 3).

2.1.2 LISTA DAS PUBLICAÇÕES RESULTANTES DA BOLSA

Artigos publicados em revistas cientificas indexadas

Fernandes Júnior, G. A.; Peripolli, E.; Schmidt, P. I.; Campos, G. S.; Mota, L. F. M.; Mercadante, M. E. Z.; Baldi, F.; Carvalheiro, R.; de Albuquerque, L. G. (2022). Current applications and perspectives of genomic selection in Bos indicus (Nellore) cattle. *Livestock Science*, 105001. (Documento 4).

Trabalhos apresentados em conferências internacionais

Campos, G. S., D. A. Silva, H. H. R. Neves, D. Lourenco, G. A. F. Júnior, L. F. S. Fonseca, L. G. Albuquerque and R. Carvalheiro (2022, July). Including selected sequence variants in genomic predictions for age at first calving in Nellore cattle. In Proceedings of the 12th World Congress on Genetics Applied to Livestock Production. (Documento 5).

Carvalho Filho, I.; Schenkel, F. S.; *Campos, G. S.*; Silva, D. A.; Teixeira, C. S.; Silva, T. L.; Lourenco, D.; Albuquerque, L. G.; Carvalheiro, R. Genomic prediction considering genotype by environment interaction for post-weaning weight gain in Nellore cattle. 1st Department of Animal Biosciences Graduate Student Symposium, University of Guelph, Guelph, Canada, 2022. (Documento 6).

Trabalhos apresentados em conferências nacionais

Campos, G. S.; Silva, D. A.; Carvalho Filho, I.; Neves, H. H. R..; Albuquerque, L. G.; Carvalheiro, R. Avaliação genômica utilizando o método de passo único para características reprodutivas em bovinos Nelore. XIV Simpósio Brasileiro de Melhoramento Animal (XIV SBMA). Outubro, 2021 (Documento 7).

Silva, D. A.; Melo, T. P.; *Campos, G. S.*; Silva, A. A.; Carvalho Filho, I.; Albuquerque, L. G.; Carvalheiro, R. Identificação de polimorfismos pleiotrópicos para precocidade sexual em bovinos da raça Nelore. XIV Simpósio Brasileiro de Melhoramento Animal (XIV SBMA). Outubro, 2021 (Documento 8).

2.1.3 Participação em banca

Carvalheiro, Roberto; Oliveira, H. N.; Mota, L. F. M.; Silva, D. A.; **Campos, G. S**. Participação em banca de Thales Lima da Silva. Estudo de associação genômica ampla e análise funcional com dados de sequenciamento completo para hipoplasia testicular e conformação de pernas e pés em bovinos da raça Nelore. Qualificação (Doutorado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias - UNESP. (Documento 9).

Baldi, F. S.; Maiorano, A. M.; **Campos, G. S**. Participação em banca de Miller de Jesus Teodoro. Estudo quantitativo e genômico de regiões codificantes e regulatórias para características ponderais no cromossomo 14 de bovinos Nelore. Qualificação (Mestrado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias - UNESP. (Documento 10).

Estão faltando os artigos em preparação. Pode colocar só o título

2.1.4 OUTROS TRABALHOS DESENVOLVIDOS NO PERÍODO (ou coisa parecida)

Artigos publicados em revistas cientificas indexadas

Comin, H. B.; *Campos, G. S.*; Domingues, R.; Gaspar, E. B.; Sollero, B. P.; Cardoso, F. F. (2022). Genetic parameters and accuracy of traditional and genomic breeding values for resistance to infectious bovine keratoconjunctivitis in Hereford. *Livestock Science*, 264, 105078. (Documento 14).

Trabalhos apresentados em conferências internacionais

Junqueira, V. S.; *Campos, G. S.*, Cardoso, F. F. (2022, July). INTERGEN: An efficient and flexible tool for large-scale genetic evaluation of complex animal and plant populations. In Proceedings of the 12th World Congress on Genetics Applied to Livestock Production. (Documento 13).

Trabalhos apresentados em conferências nacionais

Tyska, D. U.; Viana, A. F. P.; de Souza, K. A. R.; *Campos, G. S.*; Cardoso, F. F.; Braccini Neto, J. Caracterização das corridas de homozigose no genoma de bovinos da raça Hereford. XIV Simpósio Brasileiro de Melhoramento Animal (XIV SBMA). Outubro, 2021 (Documento 11).

Tyska, D. U.; Viana, A. F. P.; de Souza, K. A. R.; *Campos, G. S.*; Cardoso, F. F.; Braccini Neto, J. Identificação da distribuição de ilhas de homozigose em bovinos de corte da raça Braford. XIV Simpósio Brasileiro de Melhoramento Animal (XIV SBMA). Outubro, 2021 (Documento 12).

3. Desenvolvimento do projeto de pesquisa

Neste tópico se encontra o manuscrito que está sendo preparado referente ao presente projeto. O manuscrito será submetido para o Journal of Animal Science. Alguns resultados do projeto foram publicados como resumo no simpósio da SBMA 2021 e no 12th World Congress on Genetics Applied to Livestock Production conforme já mencionado anteriormente. Abaixo, encontra-se o manuscrito em preparação.

3.1 Manuscrito em preparação

Including selected sequence variants in genomic predictions for reproductive traits in Nellore cattle

Gabriel S. Campos[1*], Delvan A. Silva[1], Haroldo H.R. Neves[2], Daniela Lourenco[3], Ivan C. Filho[1], Gerardo A.F. Júnior[1], Larissa F.S. Fonseca[1], Lucia G. Albuquerque[1] and Roberto Carvalheiro[1]

[1]Departamento de Zootecnia, Universidade Estadual Paulista "Júlio de Mesquita Filho", Via de Acesso Professor Paulo Donato Castellane S/N - Vila Industrial, 14884-900, Jaboticabal, SP, Brazil;
[2]GenSys Consultores Associados S/S, R. Guilherme Alves, 170 - Conj. 304 - Jardim Botânico, 90680-000, Porto Alegre,RS, Brazil;
[3]University of Georgia, Department of Animal and Dairy Science, 425 River Rd, GA 30602, Athens, USA.
[*]gabrielsoarescampos@hotmail.com

Introduction

Improvements in herds' reproductive performance have a direct impact on the profitability of the production systems by influencing the longevity of cows and the number of animals for sale and selection (Krupa et al., 2005). On the other hand, the incorporation of reproductive traits in breeding programs has been slow due to low heritability, difficulty recording such traits' phenotypes, and a long generation interval to obtain selection response (Shiotsuki et al., 2009).

Using genomic information can help increase the prediction accuracy and, consequently, genetic gains for reproductive traits. The prediction of genomic estimated breeding values (GEBV) based on large sets of single nucleotide polymorphisms (SNP) has been performed in animal breeding programs without knowledge about genes or quantitative trait loci (QTL) affecting the traits. However, advances in sequencing technologies have allowed the discovery and inclusion of variants in genomic prediction.

Research interest in whole-genome sequence (WGS) data has recently increased. This data harbors causal mutations, which can be identified and included directly in genomic models without relying on linkage disequilibrium between markers and QTL. Pérez-Enciso et al. (2015) showed, in a simulation study, that if all causative variants were known, the accuracy of genomic prediction would be perfect (i.e., equal to 1). Nevertheless, including millions of genome sequence variants into genetic evaluation routines is not advantageous. A practical strategy could be variant selection to identify ones associated with traits of interest (VanRaden et al., 2017).

One suitable method to consider genotyped and ungenotyped animals in genomic evaluation is ssGBLUP (Aguilar et al., 2010). As ssGBLUP is an extension of GBLUP,

which in turn is equivalent to SNP-BLUP, SNP effects can be backsolved from ssGBLUP GEBVs, and variance explained by SNPs can be calculated based on those effects. The variance explained by SNPs can then be used as weights in constructing the genomic relationship matrix (**G**) to account for the fact that SNPs may explain different proportions of variance. This approach may be helpful when causative variants from WGS data are included in the SNP chips used for genomic evaluations (Fragomeni et al., 2017).

The main objective of this study was to evaluate the gain in accuracy of genomic predictions using weighted or regular ssGBLUP when adding selected sequence variants to a high-density SNP panel for reproductive traits in Nellore cattle. Additionally, a whole genome association using imputed sequencing data was carried out to pre-select variants for inclusion in the genomic predictions. Acho que isso não é objetivo do seu trabalho. Você não vai discutir esse resultado, certo? Deveria estar só no material e métodos. Mesmo porque é objetivo de uma tese de doutorado

Material and Methods

Animal Care and Use Committee approval was not needed because information was obtained from pre-existing databases.

Phenotype, pedigree, genotype, and sequence data

The Alliance Nellore breeding program provided the phenotypes for the reproductive traits for 401,583 females used in this study. This database contains records of animals raised on tropical pastures in different commercial herds in southeastern and central-western regions of Brazil and Paraguay. Animals were born from 1974 to 2021 and had phenotypes for the following reproductive traits: age at first calving (AFC), heifer pregnancy at 18 months at age (HP), heifer rebreeding (HR), number of calves at 53 months of age (NC53) and stayability (STAY).

The AFC was defined as the number of days from birth to first calving; HP was classified as success (1) for heifers that calved up to 31 months of age, i.e., heifers that became pregnant in the first mating season, or failure (0) for heifers with age at first calving was greater than 31 months of age; HR, also a binary trait, was defined as success (1) or failure (0) for females that calved or not, respectively, given that they had their first calving; NC53, which evaluates the ability of females to remain productive in the herd until 53 months of age, ranging from 0 to 3 and STAY, defined as a binary trait, with the status of “1” for females that remained in the herd until 76 months of age and “0” for

females that were excluded from the herd before this age. Table 1 shows the number of females with records for each trait.

From 1,348,164 animals in the pedigree data, 42,788 were either genotyped with the high-density Illumina 777k BovineHD bead chip (HD) or a medium-density SNP chip (30k-80k SNPs). All genotyped animals were imputed to HD panel sing FImpute software (Sargolzaei et al., 2014). After that, quality control of genotypes (QC) was implemented using the preGSf90 software (Misztal et al., 2014). Samples with genotyping call rates (CR) < 0.90, heterozygosities three SD above or below the observed mean, mismatched sex, and duplicate records were removed. Only SNPs mapped to autosomes with CR > 0.98, minor allele frequencies (MAF) > 0.02, and with a probability of deviation from Hardy–Weinberg Equilibrium < 0.4 (Wiggans et al., 2009) were considered in the analyses. Lastly, samples with a call rate less than 0.90 were also excluded. After editing, 425,567 SNPs and 42,788 samples remained to carry out the analysis.

In addition, 151 influential Nellore sires were sequenced with the Illumina sequencing-by-synthesis technology at an overall average sequence coverage of 14.5x (Fernandes Júnior et al., 2021). A quality control filtering step was implemented according to Fernandes Júnior et al., (2021), using the following exclusion criteria: quality by depth < 2.0; Fisher Strand test > 60.0; root mean square of the mapping quality score < 40.0; ranked sum test for the distance of alleles from the end of the reads < −8.0; mapping qualities of reads < −12.5; and SOR > 3.0. After this step, we excluded non-biallelic markers and also those with a minor allele frequency lower than 0.01. Marker genotypes with a phred-scaled confidence (a genotype quality score) less than 15 were treated as missing and those SNPs with missing values for more than 40 individuals (26.5% of the total population) were removed from the analyses. A total of 30,394,484 sequence SNPs in autosomal regions remained after quality control. All genotyped individuals were imputed to whole genome sequence (WGS) using those 151 sires as a reference population. After imputation with the software FImpute3 (Sargolzaei et al., 2014), we filtered out variants with minor allele frequencies lower than 0.001 and linkage disequilibrium values (r^2) greater than 0.9 using PLINK 2.0 (Chang et al., 2015). After this step, 4,266,603 SNPs were retained for subsequent analysis.

Statistical analysis

Contemporary groups (CG) for AFC and NC53 were formed by animals from the same farm, sex, year and season of birth, weaning and yearling management group, and

yearling date. For HR, HP and STAY, the CG of the cow's first progeny at weaning formed by farm, season and year of birth, sex, weaning management group and weaning date were considered. For AFC, CG with less than five animals and data exceeding 3.5 SD above or below the CG mean were excluded. For the other reproductive traits, CG with less than five animals and without variability were eliminated (Table 1).

Variance components based on phenotypes and pedigree were estimated using the restricted maximum likelihood method with the AIREMLF90 (Misztal et al., 2014) which uses the average information algorithm for AFC. A threshold model with the THRGIBBS1F90 program (Misztal et al., 2014) was used for the categorical reproductive traits. The THRGIBBS1F90 analyses used a single chain of 400,000 cycles, a burn-in of 100,000 cycles, and a thinning interval of 10 cycles. Posterior estimates were obtained using the POSTGIBBSF90 program (Misztal et al., 2014). Convergence was tested using the criterion proposed by Geweke (1992), with the "boa" package in R (Smith, 2007).

The model included the systematic effects of CG, calving classes (heifers with calving < 15 months; heifers with calving between 15 and 18 months; and heifers with calving higher than 18 months at age) and sex of calf for HR; CG and calving classes for STAY; and only CG for AFC, HP and NC53. Additionally, direct additive genetic and residual were included as random effects in a single trait animal model, which can be represented in matrix notation as follows:

$$y = X\beta + Za + e$$

where, **y** is the vector of observations; **β** is the vector of systematic effects; **a** is the vector of direct additive genetic effects; and **e** is the vector of random residuals. The **X** and **Z** are incidence matrices relating observations in vector **y** to effects in vectors **β** and **a**, respectively.

It was assumed that E[**y**] = **Xβ**; Var(**a**) = $\sigma_a^2 \otimes \Lambda$ and Var(**e**) = $\sigma_e^2 \otimes \mathbf{I}_n$ wherein Λ is the numerator of the relationship matrix obtained from pedigree information; **I** is an identity matrix of order l, n (number of animals with available records), ⊗denotes the direct product between the matrices, σ_a^2 is the additive genetic variance and σ_e^2 is the residual variance.

The full Bayesian threshold model for categorical reproductive traits contained systematic, additive and residual effects. A linear model using Markov-Chain Monte Carlo (MCMC) with Gibbs sampling was used for analysis of the underlying liability and it was defined as following data distribution:

$L \vee \beta, a, R\ N(X\beta + Za, R)$,

where L is a vector of unobserved liabilities of all animals. For the systematic effects, it was assumed the following prior distribution: $\beta \sim N(0,V_\beta)$ in which V_β is the non-informative diagonal variance matrix, assuming $V_\beta \rightarrow 10^{12}$. For the genetic effects, the prior distributions were previously defined. The model assumed existence of an underlying unobservable normal variable (Li), to analyze HR, HP and STAY with two scores, and for NC53 with tree scores, respectively:

$$f(y|L) = \Pi_{i=1}^{n} f(y|L_i) = \Pi_{i=1}^{n} I(L_i < t_1) I(y_i = 1) + I(L_i > t_2) I(y_i = 2),$$

$$f(y|L) = \Pi_{i=1}^{n} f(y|L_i) = \quad \Pi_{i=1}^{n} I(L_i < t_1) I(y_i = 1) + I(t_1 < L_i < t_2) I(y_i = 2) + I(L_i > t_2) I(y_i = 3),$$

where, y is the observed score for categorical reproductive traits; t_1, t_2 and t_3 are the thresholds that categorize the two levels of response for HP, HR and STAY; and three levels of response for NC53. The I is an indicator function that assumes value 1 if evaluated expression is true and 0 otherwise.

Whole genome association analysis

The EBVs calculated with the same models explained previously were used as response variables for the whole-genome association analysis. The software GCTA (Yang et al. 2014) was applied to estimate SNP effects via linear mixed models with a random polygenic effect based on G. The variance components explained by SNPs on all the chromosomes were estimated via restricted maximum likelihood (REML) was also estimated. The following statistical model was used:

$$y = 1\mu + Xb + Zu + e,$$

where y is the vector of EBVs; 1 is a vector of ones; M is the overall mean; **b** is the fixed effect of the SNP tested for association with each trait; **X** is a matrix containing the genotype score for the tested SNP; **u** is a vector of polygenic effects with $u\ N(0, G\sigma_u^2)$, where **G** is the genomic relationship matrix; σ_u^2 and is the additive genetic variance of polygenic effects; **Z** is the incidence matrix of **u**; and **e** is a vector of residual effects with $e\ N(0, I\sigma_e^2)$, where **I** is an identity matrix and σ_e^2 is the residual variance. We also performed a principal component analysis and included the first eight principal components as covariates in the model to account for possible population stratification. To prevent SNP double-counting, we utilized the option LOCO (leaving-one-chromosome-out) that excludes the chromosome on which the candidate SNP is located (Yang et al. 2014). The significant variants were selected by using a false discovery rate lower than 5% for multiple test correction. The selected variants (SV) not present in the

HD panel were included for genomic prediction. In addition to that, we also checked the quantile-quantile (Q-Q) plots and calculated the genomic inflation factor λ for each whole genome association analysis. The genomic inflation factor was defined as the median of the observed chi-squared test statistics calculated from the p-values divided by the expected median of the corresponding chi-squared distribution.

Genomic prediction

The genomic estimated breeding values (GEBV) for all traits were computed using ssGBLUP with the same models used to estimate variance components. The inverse of the realized relationship matrix (H^{-1}) was obtained as in Aguilar et al., (2010):

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

where A^{-1} is the inverse of the pedigree relationship matrix, A_{22}^{-1} is the inverse of the pedigree relationship matrix for genotyped animals, and G^{-1} is the inverse of the genomic relationship matrix (**G**) constructed using the first method of (VanRaden, 2008), with current allele frequencies as follows:

$$G = \frac{MDM\prime}{2\sum p_j(1-p_j)},$$

where **M** is a matrix of genotypes centered by twice the current allele frequencies (p); j is the jth locus; **D** is a diagonal matrix of SNP weights with a dimension equal to the number of SNPs. All SNPs were assumed to have homogeneous weights in regular ssGBLUP, meaning that **D** was an identity matrix (I). To ensure that the inverse of **G** exists, it was constructed as $0.95\mathbf{G} + 0.05A_{22}$. Parameters like τ and ω that help make G^{-1} and A_{22}^{-1} compatible, avoiding inflation of GEBV, were not changed from the default values of 1. This is because including inbreeding in the calculation of all the relationship matrices aids in matrix compatibility (Tsuruta et al., 2019). The pedigree included animals up to 10 generations back from animals with phenotypes and/or genotypes. Inbreeding was considered in the construction of all relationship matrices.

Two scenarios were tested to compute GEBV as follows: (a) using only HD SNPs, and (b) using HD SNPs and selected variants (HD+SV).

Furthermore, the ssGBLUP was performed with SNP weights iteratively, where now **D** is a matrix of weights, and each diagonal element is defined as below:

$$d_i = \sigma_{u,i}^2 \frac{\sum_j^m 2p_j q_j}{\sigma_a^2},$$

where $\sigma^2_{u,i}$ is the SNP prior variance, however, this is unknown, and the individual SNP variance can be assessed from estimates of SNP effects. In the ssGBLUP method, the SNP effects can be calculated applying a backsolving process as demonstrated by Wang et al., (2012):

$$\hat{u} = \frac{1}{\sum_j^m 2p_j q_j} DM' G_w^{-1} \hat{a},$$

where $\hat{u}$ is the vector of SNP effects, and $\hat{a}$ is a vector of GEBV. Then, after calculated SNP effects, the variances $\sigma^2_{u,i}$ were estimated with non-linear A weights as described by VanRaden et al., (2008):

$$\sigma^2_{u,i} = \frac{\sigma^2_a}{\sum_j^m 2p_j q_j} CT^{\frac{|\hat{u}_i|}{sd(\hat{u})}-2},$$

where CT is a constant of value equal to 1.125, proposed by VANRADEN (2008) which explain how much the distribution of SNP effects departures from the normal distribution; $\hat{u}_i \vee$ is the absolute estimated SNP effect for marker *i*, and *sd (û)* is the standard deviation of the vector of estimated SNP effects. The weighted ssGBLUP (WssGBLUP) was performed considering HD or HD + SV. We chose five iterations for the weighted ssGBLUP as suggested by Fragomeni et al. (2019), and results are shown only for the last iteration. In addition, we tested the regular ssGBLUP with the variants filtered out by LD to investigate the use of the sequence data. The pedigree BLUP was also applied to assess gains in prediction accuracy when using genomic information.

Validation

Validation was performed by the linear regression (LR) method proposed by Legarra and Reverter (2018), where young animals had no phenotypes in the reduced dataset (r) and complete records in the whole dataset (w). The estimators of the LR were the prediction accuracy), calculated as follows: $\hat{acc} = \sqrt{cov}$, where cov is the sample covariance, $\overline{F}$ is the average inbreeding coefficient of validation animals (i.e. young animals) and σ^2_a is the additive genetic variance computed with the whole data set. The correlation between (G)EBV$_w$ and GEBV$_r$ which assesses the association between (G)EBVs obtained with the complete and reduced datasets, was used as a measure of consistency between subsequent evaluations. The regression coefficient (b_1) of (G)EBV$_w$ on (G)EBV$_r$ was used to assess the degree of inflation/deflation of pedigree and genomic predictions. The model performance was also evaluated by the bias, calculated as $\mu = (G)EBV_r - (G)EBV_w$, where μ has an expected value of zero if the evaluation is unbiased. Statistics were calculated

for genotyped and for all young heifers (genotyped and non-genotyped) and for genotyped young bulls that were at most two years old.

Results and Discussion

The variance components and heritabilities are presented in Table 2. The heritabilities were low, ranging from 0.08 to 0.12 for all reproductive traits. These values usually agree with the ones reported in the literature for Nellore cattle in Brazil (Schmidt et al., 2019; 2020). Despite the low additive variability for female reproductive traits, the use of genomic information can help achieve higher accuracies than the traditional genetic evaluation for female reproductive traits, as shown in this paper.

Whole genome association analysis

We performed the GWAS using the mixed model leave-one-chromosome-out (LOCO) method, which prevents the tested SNP from being double-fitted in the model. This method excludes the chromosome on which the candidate SNP is located (Yang et al. 2014). The LOCO approach has been used in humans to improve the power of GWAS (Listgarten et al., 2012; Yang et al., 2014) and with sequence data in livestock to be an extremely efficient method. The Manhattan plots, Q-Q plots and genomic inflation factors are presented in Figure 1 for all female reproductive traits.

The number of significant SNPs that were different for each trait ranged between 185 and 4,957. AFC had the greatest number of significant SNPs (4,957 SNPs), accounted for 1.59% of the total genetic variance. The trait HP had 2,208 significant associated markers and 1,029 SNPs for STAY, which explained 0.79% and 0.42% of the total genetic variance, respectively. However, for HR and NC53, only 751 and 185 significant variants were selected with a false discovery rate lower than 5% and responded for 0.32% (HR) and 0.08% (NC53) of the genetic variance explained by the significant markers.

The average genomic heritabilities (h^2) between chromosomes (data not shown) estimated by REML were 0.68, 0.38, 0.20, 0.13 and 0.10 for HR, AFC, STAY, HP and NC53. These h^2 were higher than those estimated by the pedigree for the most traits. The results were distinct because we used the EBV as responsible variable and only animals with individual accuracy greater than fifty percent for the GWAS. This fact indicated that the SNPs explained a considerable amount of the variation compared to the pedigree model for HR, AFC, STAY and NC53.

The genomic inflation factors at a whole-genome level were higher than the unit for all reproductive traits, even adjusting the model for population stratification. Account for population stratification in the GWAS model is essential to avoid genomic inflation and false positives (Hinrichs et al., 2009; Price et al., 2006). The values were 1.423, 1.38, 1.338, 1.292, and 1.262 for AFC, HP, NC53, STAY, and HR, respectively. Likely, the method LOCO is the reason for the genomic inflation to be slightly higher than one for all traits in our population. Van den Berg et al. (2019) showed that this approach does not account appropriately for population stratification in a pig population and possibly in other livestock species. The reason is that usually, in the livestock breeding populations, we have many strong family relationships and not unrelated individuals, as, for instance, in humans. A possible solution could be adjusting a **G** matrix using all chromosomes. However, we only had significant SNPs for AFC with this approach (data not shown).

It is known that most of these reproductive traits follow the pattern of complex quantitative traits spread over the entire genome. Our study did not find common SNPs among all reproductive traits. Nevertheless, AFC and HP had 285 significant overlapping SNPs. The region with the highest number of SNPs in common was on chromosome 15 (63 SNPs) and 18 (53 SNPs), confirming that both traits have pleiotropic regions. Mota et al. (2019) reported important genomic regions on chromosomes 2, 3, 5, 14, and 18 associated with AFC. The authors performed a GWAS using a single-step reaction norm model across environments in Nellore cattle. Using a GWAS with the single-step method for HP, Irano et al., (2018) also found SNPs on BTA 18 in Nellore cattle. Concerning the other traits, only a few significant associated SNPs were common, varying between 3 and 11 SNPs. The slight overlap among traits indicates the presence of trait-specific QTL for female reproductive traits in our study. Despite that, a multiple-trait association analysis could help improve the power of GWAS.

Genomic prediction using ssGBLUP

The genomic prediction accuracies with the ssGBLUP method outperformed those from traditional BLUP for all reproductive traits considering all scenarios tested. The LR approach was used to validate the ssGBLUP models with or without inclusion variants from sequence data. Bermann et al. (2021) tested the method with simulation data and a real data set for chicken mortality (binary trait indicating the survival status). They concluded that the LR is useful for estimating the magnitude of (G)EBV quality

metrics for threshold models. In addition, the authors highlighted that the approach applies to any single-step procedure and is easy to run.

The GEBV accuracies ranged from 0.34 for HR to 0.52 for NC53 across all traits. Prediction accuracies from ssGBLUP outperformed those from BLUP for all traits, with an average superiority of 27% for young heifers, 34% for young genotyped heifers and 53% for young genotyped bulls. This confirmed that utilization of genomic information would bring additional benefits to the genetic evaluation of reproductive traits in Nellore cattle in Brazil mainly for young bulls. The higher prediction accuracies with ssGBLUP occurred because genomic information provided more precise relationships among individuals and helped to estimate Mendelian sampling. These improvements could result in greater genetic gains in the Brazilian Nellore population.

The dispersion values (LR coefficients of GEBVw on GEBVp) varied from 0.86 for HP to 1.06 for STAY (Table 3) for young heifers and ranged from 0.81 for NC53 to 1.11 for AFC in young genotyped bulls. The slope (b1) of the regression coefficient should be close to one to ensure that there is no inflation or deflation in (G)EBV for validation animals. Inflation/deflation may be a problem when predictions of genotyped and nongenotyped animals need to be compared for ranking and selection decisions. When the LR coefficient of GEBVw on GEBVp is less than 1, predictions are inflated and extreme GEBV may exist, favoring genotyped individuals (Cardoso et al., 2015).

For all the traits and validation group, the bias values were close to zero (Table 3 and 4). The GEBV had negative biases for AFC, HR and STAY for young heifers. Beyond these traits, NC53 had negative biases for young genotyped bulls, indicating that the (G)EBV means for validation animals from the reduced dataset were lower than those from the complete dataset.

The correlation between GEBVs with the complete dataset and another one with a reduced dataset was used as a measure of consistency between subsequent evaluations. This consistency can be understood as the ability of the reduced dataset to predict the complete dataset. Overall, the use of ssGBLUP models helped increase the correlations between GEBVs from the reduced and complete data for all traits and validation groups.

ssGBLUP including selected sequence variants

Our study is the first using WGS with the ssGBLUP for reproductive traits in Nellore cattle. We tested three scenarios to compare the genomic predictions: using only HD SNPs, HD plus selected sequence variants (HD+SV) and sequence variants filtered

out by LD (4,226,888 SNPs). Including SV in the regular ssGBLUP has not increased prediction accuracies and only marginally increased accuracy in WssGBLUP with nonlinear A method (HD+SV). Using SV combined with SNP panels considerably increased prediction accuracy in simulation studies (Fragomeni et al., 2017) when applying WssGBLUP, compared to the regular ssGBLUP. However, we found only a marginal increase when including SV in weighted ssGBLUP with nonlinearA weights compared to the HD SNP panel. This result is similar to those from Fragomeni et al. (2019) for stature in US Holsteins. The authors compared ssGBLUP with and without weights and did not find any gain by including nearly 17,000 significant variants to a medium- density panel. According to Lourenco et al. (2020), the amount of information used in ssGBLUP overwhelms any a priori assumption on SNP effects, being less sensitive to weighting approaches for large data sets. These authors suggested that an ideal algorithm should assign large variance to causative SNPs while reducing the variance of nearby SNPs. Exploring methods to increase the power of detection of causal variants, as multi-trait meta-analysis or gene expression and genome annotation could help avoid spurious associations. The use of additional biological sources of information (i.e., multi-omics data) could also help to increase the likelihood of identifying mutations that contribute to the variance of complex traits and possibly, increase the prediction accuracies.

References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

van den Berg, S., J. Vandenplas, F. A. van Eeuwijk, M. S. Lopes, and R. F. Veerkamp. 2019. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. J. Anim. Breed. Genet. 136:418–429. doi:10.1111/JBG.12419. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/jbg.12419

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 4:7. doi:10.1186/S13742-015-0047-8/2707533. Available from: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533

Fernandes Júnior, G. A., R. Carvalheiro, H. N. de Oliveira, M. Sargolzaei, R. Costilla, R. V. Ventura, L. F. S. Fonseca, H. H. R. Neves, B. J. Hayes, and L. G. de Albuquerque. 2021. Imputation accuracy to whole-genome sequence in Nellore cattle. Genet. Sel. Evol. 53:1–10. doi:10.1186/S12711-021-00622-5/TABLES/3. Available

from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00622-5

Fragomeni, B. O., D. A. L. Lourenco, A. Legarra, P. M. VanRaden, and I. Misztal. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. J. Dairy Sci. 102:10012–10019. doi:10.3168/JDS.2019-16262.

Fragomeni, B. O., D. A. L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet. Sel. Evol. 49:1–11. doi:10.1186/S12711-017-0335-0/TABLES/2. Available from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-017-0335-0

Geweke, J. 1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In: A. P. D. and A. F. M. S. J. M. Bernardo, J. O. Berger, editor. Bayesian Statistics. Oxford University Press, New York. p. 625–631.

Hinrichs, A. L., E. K. Larkin, and B. K. Suarez. 2009. Population stratification and patterns of linkage disequilibrium. Genet. Epidemiol. 33:S88–S92. doi:10.1002/GEPI.20478. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.20478

Irano, N., G. M. F. De Camargo, R. B. Costa, A. P. N. Terakado, A. F. B. Magalhães, R. M. D. O. Silva, M. M. Dias, A. B. Bignardi, F. Baldi, R. Carvalheiro, H. N. De Oliveira, and L. G. De Albuquerque. 2016. Genome-Wide Association Study for Indicator Traits of Sexual Precocity in Nellore Cattle. PLoS One. 11:e0159502. doi:10.1371/JOURNAL.PONE.0159502. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159502

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50:53. doi:10.1186/s12711-018-0426-6. Available from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-018-0426-6

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman. 2012. Improved linear mixed models for genome-wide association studies. Nat. Methods 2012 96. 9:525–526. doi:10.1038/nmeth.2037. Available from: https://www.nature.com/articles/nmeth.2037

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Aguilar I., Legarra A., and Vitezica Z. 2014. Manual for BLUPF90 family of programs. Available from: http://nce.ads.uga.edu/wiki/doku.php

Mota, L. F. M., F. B. Lopes, G. A. Fernandes Júnior, G. J. M. Rosa, A. F. B. Magalhães, R. Carvalheiro, and L. G. Albuquerque. 2020. Genome-wide scan highlights the role of candidate genes on phenotypic plasticity for age at first calving in Nellore heifers. Sci. Reports 2020 101. 10:1–13. doi:10.1038/s41598-020-63516-4. Available from: https://www.nature.com/articles/s41598-020-63516-4

Pérez-Enciso, M., J. C. Rincón, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 2015 471. 47:1–14. doi:10.1186/S12711-015-0117-5. Available from: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-015-0117-5

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide

association studies. Nat. Genet. 2006 388. 38:904–909. doi:10.1038/ng1847. Available from: https://www.nature.com/articles/ng1847

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

Schmidt, P. I., G. S. Campos, V. M. Roso, F. R. P. Souza, and A. A. Boligon. 2019a. Genetic analysis of female reproductive efficiency, scrotal circumference and growth traits in Nelore cattle. Theriogenology. 128:47–53. doi:10.1016/J.THERIOGENOLOGY.2019.01.032.

Schmidt, P. I., I. A. Ferreira, D. D. Silveira, G. S. Campos, F. R. P. Souza, R. Carvalheiro, and A. A. Boligon. 2019b. Reproductive performance of cows and genetic correlation with weight gains and principal components of traits used in selection of Nelore cattle. Livest. Sci. 229:77–84. doi:10.1016/J.LIVSCI.2019.09.011.

Shiotsuki, L., J. A. de Vasconcelos Silva, and L. G. de Albuquerque. 2009. Associação genética da prenhez aos 16 meses com o peso à desmama e o ganho de peso em animais da raça Nelore. Rev. Bras. Zootec. 38:1211–1217. doi:10.1590/S1516-35982009000700008. Available from: http://www.scielo.br/j/rbz/a/XWnWv8cgF3sL4GpB9LTQwbC/abstract/?lang=pt

Smith, B. J., Smith, and B. J. 2007. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. J. Stat. Softw. 021.

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

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol. 2017 491. 49:1–12. doi:10.1186/S12711-017-0307-4. Available from: https://link.springer.com/articles/10.1186/s12711-017-0307-4

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436. doi:10.3168/JDS.2008-1758.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and & Alkes. 2014. Advantages and pitfalls in the application of mixed-model association methods. doi:10.1038/ng.2876.

Table1. Data structure and descriptive statistics for the Nellore cattle population.

Trait[1]	Number of records	Mean ± SD	Minimum	Maximum	Number of contemporary groups	Number of sires	Number of dams
AFC	224,219	1010.07 ± 133.42	540	1220	11,582	4,842	172,028
HR	178,152	0.59 ± 0.49	0	1	21,647	3,957	138,527
HP	166,842	0.21 ± 0.40	0	1	4,701	3,474	126,020
NC53	356,472	0.84 ± 0.88	0	3	10,655	5,150	247,176
STAY	137,620	0.47 ± 0.50	0	1	16,566	2,950	107,619

[1]Trait: Age at first calving (AFC), heifer pregnancy at 18 months at age (HP), heifer rebreeding (HR), number of calves at 53 months of age (NC53) and stayability (STAY).

Table 2. Variance components and heritabilities for reproductive traits in Nellore cattle.

Trait[1]	σ_a^2	σ_e^2	h_a^2
AFC	393.74	6279.60	0.06±0.01
HR	0.108	1.034	0.09±0.02
HP	0.197	1.013	0.16±0.01
NC53	0.061	0.695	0.08±0.01
STAY	0.127	1.032	0.11±0.02

[1]Trait: Age at first calving (AFC), heifer pregnancy at 18 months at age (HP), heifer rebreeding (HR), number of calves at 53 months of age (NC53) and stayability (STAY).

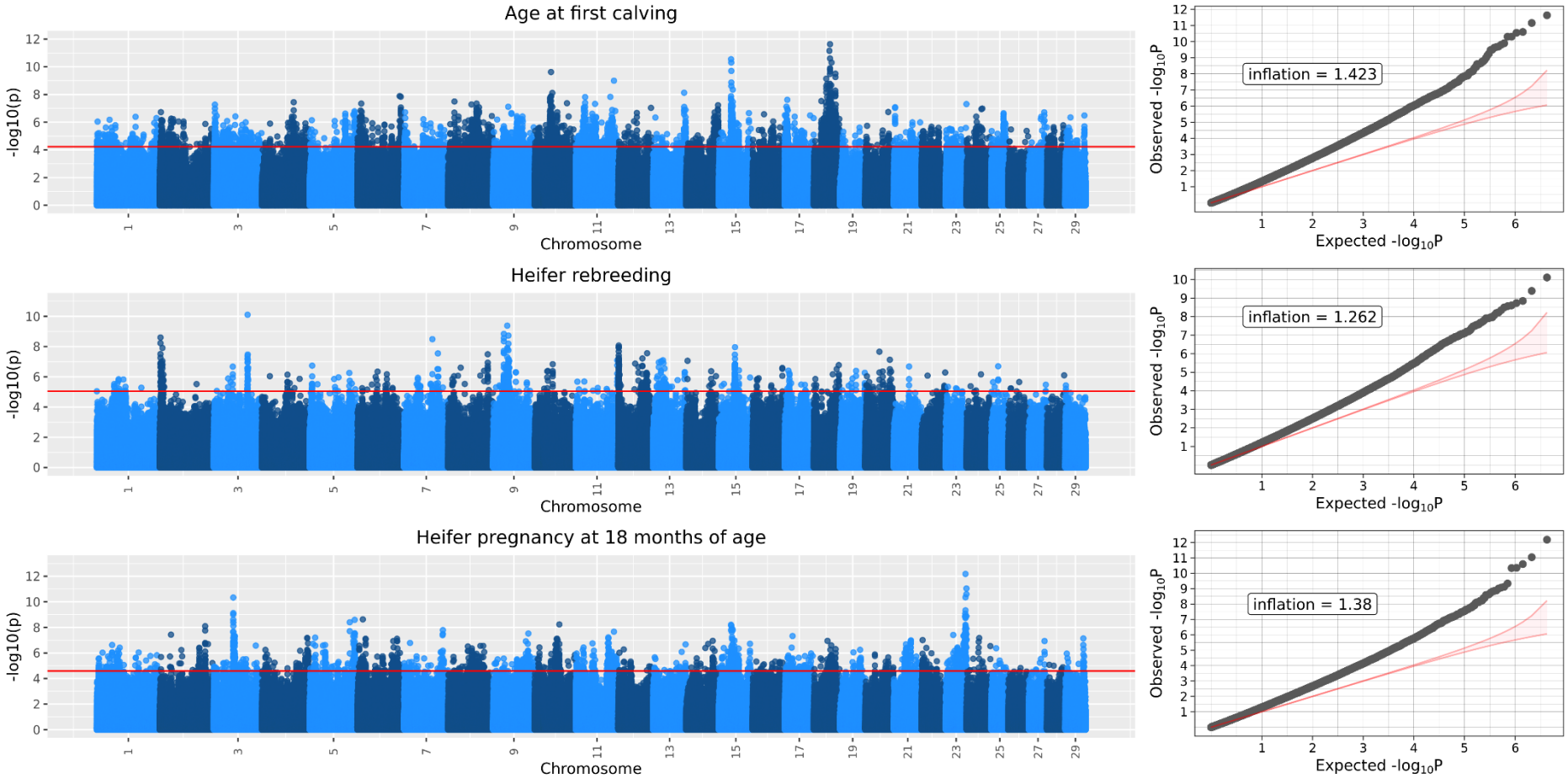
Age at first calving
-log10(p)
Chromosome
Observed -log10P
Expected -log10P
inflation = 1.423
Heifer rebreeding
-log10(p)
Chromosome
Observed -log10P
Expected -log10P
inflation = 1.262
Heifer pregnancy at 18 months of age
-log10(p)
Chromosome
Observed -log10P
Expected -log10P
inflation = 1.38

Figure 1. Manhattan plot charts and Q-Q plots showing the results for the whole genome association study for reproductive traits in Nellore cattle. The red line is the significance threshold for the false discovery rate of 5%. Q-Q plots are displayed as scatter plots of observed and expected -log10 (p-values). Inflation values are reported within the Q-Q plots.

Table 3. Linear regression statistics of (G)EBV for all and young genotyped heifers when including or excluding selected variants (SV) and variants filtered out by linkage disequilibrium in single-step GBLUP (ssGBLUP) for reproductive traits.

Trait[1]	Method[2]	Scenario[3]	All young heifers				Young genotyped heifers			
			Accuracy	b_1	Bias	Correlation	Accuracy	b_1	Bias	Correlation
AFC	BLUP	Pedigree	0.35	0.87	-0.494	0.83	0.44	0.85	-0.644	0.86
	ssGBLUP	HD	0.42	0.88	-0.426	0.85	0.62	0.88	-0.605	0.91
	WssGBLUP	HD	0.42	0.88	-0.435	0.85	0.63	0.89	-0.603	0.90
	ssGBLUP	HD+SV	0.42	0.88	-0.429	0.86	0.63	0.89	-0.586	0.91
	WssGBLUP	HD+SV	0.43	0.88	-0.447	0.86	0.64	0.89	-0.602	0.91
	ssGBLUP	LD	0.42	0.87	-0.432	0.84	0.60	0.89	-0.646	0.89
HR	BLUP	Pedigree	0.29	0.90	-0.018	0.76	0.33	0.94	-0.014	0.76
	ssGBLUP	HD	0.34	0.90	-0.014	0.80	0.39	0.93	-0.004	0.80
	WssGBLUP	HD	0.34	0.89	-0.014	0.79	0.40	0.92	-0.004	0.79
	ssGBLUP	HD+SV	0.35	0.90	-0.015	0.79	0.39	0.86	-0.002	0.77
	WssGBLUP	HD+SV	0.35	0.90	-0.014	0.80	0.41	0.88	-0.001	0.77
	ssGBLUP	LD	0.33	0.88	-0.016	0.80	0.39	0.91	0.002	0.78
HP	BLUP	Pedigree	0.29	0.88	0.016	0.73	0.30	0.91	0.017	0.76
	ssGBLUP	HD	0.37	0.86	0.024	0.80	0.47	0.90	0.033	0.87
	WssGBLUP	HD	0.37	0.86	0.023	0.80	0.47	0.90	0.032	0.86
	ssGBLUP	HD+SV	0.37	0.87	0.026	0.80	0.47	0.91	0.037	0.87
	WssGBLUP	HD+SV	0.38	0.87	0.027	0.80	0.48	0.93	0.033	0.86
	ssGBLUP	LD	0.36	0.79	0.021	0.78	0.46	0.88	0.044	0.85
NC53	BLUP	Pedigree	0.42	0.94	-0.006	0.83	0.46	0.97	-0.031	0.85
	ssGBLUP	HD	0.52	0.93	0.001	0.89	0.49	0.93	-0.018	0.87
	WssGBLUP	HD	0.52	0.93	0.001	0.88	0.50	0.94	-0.020	0.87
	ssGBLUP	HD+SV	0.52	0.94	0.002	0.90	0.49	0.98	-0.013	0.87
	WssGBLUP	HD+SV	0.52	0.93	0.002	0.89	0.50	0.98	-0.015	0.88

	ssGBLUP	LD	0.51	0.93	0.001	0.88	0.49	1.05	-0.019	0.87
STAY	BLUP	Pedigree	0.26	1.07	-0.011	0.73	0.27	0.79	-0.018	0.66
	ssGBLUP	HD	0.38	1.03	-0.007	0.83	0.39	0.89	-0.007	0.76
	WssGBLUP	HD	0.39	1.03	-0.006	0.83	0.39	0.89	-0.008	0.76
	ssGBLUP	HD+SV	0.38	1.02	-0.006	0.83	0.40	0.94	-0.005	0.77
	WssGBLUP	HD+SV	0.39	1.02	-0.005	0.83	0.41	0.93	-0.005	0.77
	ssGBLUP	LD	0.38	1.05	-0.009	0.81	0.39	1.11	-0.015	0.75

[1]Trait: Age at first calving (AFC), heifer pregnancy at 18 months at age (HP), heifer rebreeding (HR), number of calves at 53 months of age (NC53) and stayability (STAY).

[2]BLUP: regular pedigree BLUP; ssGBLUP: regular single-step genomic BLUP; WssGBLUP: weighted single-step GBLUP with nonlinear A method in the fifth iteration.

[3]HD: high density SNP panel; SV: selected sequence variants; LD: sequence data filtered out by linkage disequilibrium.

Table 4. Linear regression statistics of (G)EBV for young genotyped bulls when including or excluding selected sequence variants (SV) and variants filtered out by linkage disequilibrium in single-step GBLUP (ssGBLUP) for reproductive traits.

Trait[1]	Method[2]	Scenario[3]	Young genotyped bulls			
			Accuracy	b_1	Bias	Correlation
AFC	BLUP	Pedigree	0.36	0.86	-0.446	0.89
	ssGBLUP	HD	0.55	0.90	-0.458	0.93
	WssGBLUP	HD	0.56	0.89	-0.443	0.93
	ssGBLUP	HD+SV	0.55	0.90	-0.430	0.93
	WssGBLUP	HD+SV	0.57	0.91	-0.422	0.93
	ssGBLUP	LD	0.54	0.89	-0.601	0.93
HR	BLUP	Pedigree	0.27	1.09	-0.049	0.72
	ssGBLUP	HD	0.36	0.86	-0.056	0.76
	WssGBLUP	HD	0.37	0.89	-0.058	0.75
	ssGBLUP	HD+SV	0.36	0.84	-0.059	0.76
	WssGBLUP	HD+SV	0.37	0.86	-0.061	0.77
	ssGBLUP	LD	0.36	1.11	-0.059	0.79
HP	BLUP	Pedigree	0.26	0.96	0.015	0.86
	ssGBLUP	HD	0.43	0.93	0.032	0.90
	WssGBLUP	HD	0.44	0.93	0.031	0.90
	ssGBLUP	HD+SV	0.43	0.93	0.034	0.90
	WssGBLUP	HD+SV	0.44	0.98	0.031	0.90
	ssGBLUP	LD	0.43	0.89	0.039	0.89
NC53	BLUP	Pedigree	0.37	0.93	-0.035	0.88
	ssGBLUP	HD	0.49	0.81	-0.033	0.84
	WssGBLUP	HD	0.49	0.94	-0.027	0.89
	ssGBLUP	HD+SV	0.51	0.83	-0.035	0.86
	WssGBLUP	HD+SV	0.53	0.96	-0.029	0.89
	ssGBLUP	LD	0.47	1.09	-0.042	0.87
STAY	BLUP	Pedigree	0.25	1.05	-0.026	0.85
	ssGBLUP	HD	0.47	1.04	0.021	0.91
	WssGBLUP	HD	0.44	0.83	-0.026	0.87
	ssGBLUP	HD+SV	0.44	0.89	-0.028	0.87
	WssGBLUP	HD+SV	0.45	0.89	-0.026	0.87
	ssGBLUP	LD	0.43	0.86	-0.061	0.86

[1]Trait: Age at first calving (AFC), heifer pregnancy at 18 months at age (HP), heifer rebreeding (HR), number of calves at 53 months of age (NC53) and stayability (STAY).
[2]BLUP: regular pedigree BLUP; ssGBLUP: regular single-step genomic BLUP; WssGBLUP: weighted single-step GBLUP with nonlinear A method in the fifth iteration.
[3]HD: high density SNP panel; SV: selected sequence variants; LD: sequence data filtered out by linkage disequilibrium.

4. Colaboração em outros trabalhos paralelamente ao projeto de pesquisa

4.1 Unraveling genetic architecture of depigmentation in nellore cattle by genome wide association studies of imputed whole-genome sequence.

PPG: Genética e Melhoramento Animal, FCAV UNESP

Pesquisadora: Giovana Vargas

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.

Unraveling genetic architecture of depigmentation in nellore cattle by genome wide association studies of imputed whole-genome sequence

ABSTRACT

The introduction of the Special Certificate of Identification and Production (CEIP) to beef cattle breeding programs has contributed to successfully promote the genetic progress of economic important traits in Brazil. The top 20-30% animal candidates to receive the certificate are evaluated for problems related to functional traits that can negatively impact the animal's performance. Depigmentation (DP) is one of the main causes of disqualification since it can contribute to the appearance of skin lesions and consequent infections impairing the animal's performance. The objectives of the present study were (1) to perform a genome-wide association study (GWAS) of imputed whole-genome sequence data for DP in Nellore cattle; (2) to detect potential candidate genes that could be used to assist Nellore cattle breeding programs in the selection process; and (3) to detect structural variation (SVs) associated with DP by performing a GWAS using whole-genome sequence data on influential Nellore sires. Phenotypes and pedigree data were available for 340,220 Nellore animals. The genotype was imputed to the sequence using as reference population a set of 151 Nellore sires already sequenced. After quality control analyses, 4,260,018 SNPs were retained for subsequent analysis. Functional candidate gene analysis revealed 31 candidate genes for DP of which KIT and MITF were previously found to play roles in skin pigmentation. Significant enrichment results were identified for eye area pigmentation and skin pigmentation QTLs within the candidate positions. Gene ontology enrichment analysis including biological process, molecular functions, and cellular components revealed significant GO terms associated with DP.

The identification of these regions contributes to a better understanding of DP and provided knowledge of positional candidate genes that could be used for future investigation of causal mutations and their implications.
Keywords: depigmentation, genome sequencing, genome-wide association, Nellore

INTRODUCTION

Beef cattle industry from Brazil has successfully undergone major transformations thanks to improvements involving management, environmental progress, and genetic selection performed by breeding programs that have been focusing on selecting traits with clear economic value. The genetic progress achieved by these breeding programs can be assigned due the implementation of CEIP (Special Certificate of Identification and Production), a certificate of genetic superiority launched by the Ministry of Agriculture, Livestock and Food Supply (MAPA, 1995). The certificate is issued for the top 20-30% animals (young bulls and heifers), evaluated at weaning and yearling, within each breeding program. Besides being genetically superior, the animals potentially eligible to receive the CEIP must not present serious morphological and functional defects or disorders such as feet and legs problems, nasal deviation, testicular hypoplasia, and depigmentation (DP).

From a biological standpoint, the external morphology of the animal represents the first layer in contact with environmental effects, such as solar radiation and other climatic factors, which makes the skin color an important faculty of heat tolerance in livestock. The impact of heat stress on the animals' performance is highlighted when the external temperature increases combined with solar radiation, wind speed and humidity, leading to reduced feed intake consumption in an attempting to decrease metabolic heat. The dark pigmentation typical of Nellore cattle is a powerful characteristic of this breed that allowed its adaptation to tropical climates. The composition of dark skin and white hair is the most efficient for animals' thermal balance, since depigmentation can contribute to the appearance of skin lesions and consequent infections impairing the performance of the animals. In this sense, the selection focusing on well pigmented animals (dark skin) against depigmented animals (pink skin areas) is crucial for a successful productive performance.

Since the Bovine Genome Sequencing (Elsik et al., 2009), an extensive use of genotyped animals has allowed the identification of genomic regions containing

quantitative trait loci (QTL) that affect traits of interest, such as DP. By combining genotyping information for genomic selection with reference data from whole-genome sequence, it becomes possible to carry out genome-wide association studies (GWAS) for identification of causative variants associated with complex traits across the cattle genome. The imputation of SNPs from genotyped lower-density panels to the whole-genome sequence increases the possibility of identifying potential genetic variants due its high DNA marker panel. Currently, GWAS are applied to identify candidate genes for many traits in Nellore cattle (Magalhães et al., 2016; Melo et al., 2017; Neves et al., 2019; Vargas et al., 2018, 2020); however, a few studies have explored sequence information of indicine cattle breeds.

In addition to use genome sequence data to perform associations studies based on molecular markers, a different source of information that can be explored in an attempting to identify genomic regions of interest are genomic structural variations (SVs) including deletions, insertions, inversions, and duplications. SVs are the main source of variation in the genome of complex organisms, contributing to a greater proportion of the differences between individuals when compared to SNPs, and therefore representing an important source of genetic diversity (Baker, 2012). It is well known that SVs are responsible for variation in many complex traits in cattle, including milk production, fertility, and other traits. Therefore, the objectives of the present study were (1) to perform a genome-wide association study (GWAS) of imputed whole-genome sequence data for DP in Nellore cattle; (2) to detect potential candidate genes that could be used to assist Nellore cattle breeding programs in the selection process; and (3) to detect SVs associated with DP by performing a GWAS using whole-genome sequence data on influential Nellore bulls.

MATERIAL AND METHODS

Phenotype and pedigree data

Phenotype and pedigree information of 340,220 Nellore cattle from the Alliance Nellore database (www.gensys.com.br) collected between 2000 and 2018 was considered for this study. The animals belong to herds that systematically perform evaluations for morphological and functional defects or disorders. Depigmentation (DP) is evaluated as a binary trait at weaning (about 205 days of age, DP_W) and at yearling (about 550 days of age, DP_Y).

Contemporary groups (CG) at weaning was defined by concatenating sex, season of birth, farm at birth, management group at birth, farm at weaning, management group at weaning, measurement date. The CG at yearling were defined by concatenating CG at weaning, farm at yearling, management group at yearling and measurement date. Data from CG with fewer than 10 records and/or without variability in the trait was removed from further analyses. Connectedness among CG was determined by using the AMC software (Roso & Schenkel, 2006) and only CG with at least 10 genetic links was kept for the analyses.

Sequencing data and genotype imputation

A total of 17.558 animals from the Alliance Nellore database (www.gensys.com.br), genotyped using the Illumina BovineHD Genotyping BeadChip (HD; Illumina, Inc., San Diego, CA, USA) or with a lower density panel (from 28K to 74K SNPs), had their genotypes imputed to the sequence level. The genotype quality control considered as inclusion criteria: SNPs located in autosomes; GenCall score greater than 0.70; minor allele frequency greater or equal to 0.02; call rate per SNP greater than 0.05; call rate per animal greater than 0.90. The imputation was carried out in two steps. First, the animals genotyped with the lower density panels had their genotype imputed to the HD panel. Subsequently, the HD genotypes were imputed to the sequence. The whole-genome sequencing information from 151 influential Nellore sires were chosen as reference population for imputation of the genotypes aiming to optimize the imputation accuracy of our genotype database. The 151 Nellore sires were sequenced using the Illumina HiSeq X™ Ten and Illumina NovaSeq™ platforms (Illumina Inc., San Diego, USA). Sequencing of all samples was performed using the Illumina sequencing-by-synthesis technology at an overall average sequence coverage of 14.5×, ranging from 7.8 to 26.3×. The SAMtools program (Li et al., 2009) was used to index the sequences, and the alignment to the ARS-UCD1.2 reference genome was performed using HiSat (Kim et al., 2015). After this step, non-biallelic markers were excluded and also those with a minor allele frequency lower than 0.01. Marker genotypes with a phred-scaled confidence (a genotype quality score) less than 15 were treated as missing and those SNPs with missing values for more than 40 individuals (26.5% of the total population) were removed from the analyses. A total of 30,394,484 sequence SNPs in autosomal regions remained after quality control. All genotyped individuals were imputed to WGS using those 151 sires as a reference population. After imputation with the software FImpute3 (Sargolzaei et al., 2014), we filtered out variants with minor allele frequencies lower than

0.001 and linkage disequilibrium values (r2) greater than 0.9 using PLINK 2.0 (Chang et al., 2015). After quality control analyses, 4,260,018 SNPs were retained for subsequent analysis.

Variant calling

Single sample SV discovery (including deletions, insertions, inversions, and duplications) was performed using Parliament2 (Zarate et al., 2020) considering the following combination of callers CNVnator (Abyzov et al., 2011), Lumpy (Layer et al., 2014), Delly (Rausch et al., 2012) and BreakDancer (Fan et al., 2014). The BAM files and associated reference genome file (ARS-UCD1.2) were used to run Parliament2 on each sample via a single Docker image using the DNAnexus platform. A merged VCF file was prepared for each variant calling by merging individual VCF files of the 81 samples using the BCFtools merge function. The output file containing SVs was then transformed to 012 genotypes format using VCFtools (Danecek et al., 2011), where 0 represented the homozygous genotype for the reference allele (no SV), 1 was the heterozygous genotype (heterozygous SV) and 2 was the homozygous genotype for the alternative allele (homozygous SV).

Genome-wide association analyses

The GWAS analysis was undertaken independently on 81 sequenced animals using SV information as genotype and on imputed sequence data. The MLMA (mixed linear model association) module implemented in the GCTA software was used for analyses (Yang et al., 2011), fitting the following linear mixed model:

$y=\mu+Xb+Zu+e,$

where y is the vector of phenotypes, μ is the overall mean, b is the allele substitution effect, X is the incidence matrix of the allele copies (coded as 0, 1, 2), Z is the incidence matrix connecting the phenotypes to animals, u is the vector of additive genetic effect, and e is the vector of random residuals. The genomic relationship matrix (GRM) was derived as defined in Yang et al. (2010, 2011). It was assumed that $u \sim N(0, G\sigma_a^2)$, where $\sigma_a^2$is the additive genetic variance and G is the realized genomic relationship matrix calculated from all remaining SNPs, and $e \sim N(0, I\sigma_e^2)$. The percentage of genetic variances explained by each significant SNP was calculated according to the following formula:

$$\%V_g = \frac{2pq\alpha^2}{\sigma_g^2} * 100$$

Where p_i and q_i are the allele frequencies for the i_{th} SNP, α_i is the estimated additive genetic effect of the i_{th} SNP on depigmentation, σ_g^2 is the estimated genetic variance.

Quantile-quantile plot and manhattan plots were generated using the R package qqman (Turner, 2014). For further downstream analysis, SNP markers with a − log10(P-value) which exceeded the Bonferroni threshold (i.e., 0.05/number of independent tests) were considered to be statistically significant and were further annotated into coding regions (genes) of the cattle genome. The genomic inflation factor (λ = the observed P value/the expected P values) was calculated to evaluate the extent of false positive signals.

Enrichment analysis

Significant SNPs were investigated for gene and quantitative trait loci (QTL) annotation using the Genomic Annotation in Livestock for positional candidate Loci (GALLO; https://github.com/pablobio/GALLO) package in R (Version 4.0.0.; R Core Team, 2020). The .gtf annotation file corresponding to the bovine gene annotation and the .gff file with the QTL information from Animal QTL Database (Hu et al., 2013) were used for genes and QTL annotation using the ARS-UCD1.2 assembly coordinates, respectively. A QTL enrichment analysis was conducted for all the QTL information annotated using a chromosome-based enrichment analysis to identify the potential effects of polymorphisms associated to any of the six trait classes, e.g. ′Reproduction′, ′Milk′, ′Production′, 'Exterior', 'Meat and Carcass', and 'Health'. A P-value for the QTL enrichment status of each annotated trait was calculated. In addition, the P-values were corrected for multiple testing using FDR (5%). The list of candidate genes was used to identify Gene Ontology (GO) terms (e.g. biological process, molecular function, and cellular component) and to search metabolic pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG). In this study, a protein-protein interaction network analysis was performed for the candidate genes using the NetworkAnalyst 3.0 (Xia et al., 2015; https://www.networkanalyst.ca) in order to identify interactions between the positional candidate genes and other genes in the genome.

RESULTS

Characterization of structural variants

A total of 146,003 deletions, 52,208 insertions, 84,306 duplications, and 226,036 inversions were identified from the sequence data of the Nellore population, which jointly affect ~ 2.38 Mb of sequences (Table 1). The distribution of the number of structural

variants on the 29 autosomes is shown in Figure 1. In general, the number of deletions and inversions detected in each chromosome was greater than the number of duplications and insertions. In addition, the distribution of the SVs detected on each chromosome is quite uniform. Chromosome 1 had the highest density of structural variants, while chromosome 25 had the lowest. The annotation of detected variants was performed to further explore the distribution of SVs in genic regions. Within genic regions there were 46,850 (32.08%), 15,690 (30.05%), 7,937 (3.51%), and 30,446 (36.11%) identified deletions, insertions, inversions, and duplications, respectively. In addition, the number of total genes for deletions, insertions, inversions, and duplications contained in the Ensembl gene database of bovine were 6,483 (37.00%), 5,413 (30.89%), 2,104 (12.01%), and 1,966 (11.22%). The range of the number of deletions, relative to genes distributed on different chromosomes was from 24.23% to 52.87%; for insertions, from 18.30% to 45.97%; for inversions, from 5.08% to 19.4%; and for duplications, from 4.59% to 89.09%. Chromosomes 6, 24, and 28 had the highest number of genes with deletions, insertions, inversions, and duplications.

Table 1. Summary of structural variants.

	Deletions	Insertions	Duplications	Inversions	Total
No.	146,003	52,208	84,306	226,036	508,553
Average size (bp)	686,225	1	1,013,137	683,547	2,382,910

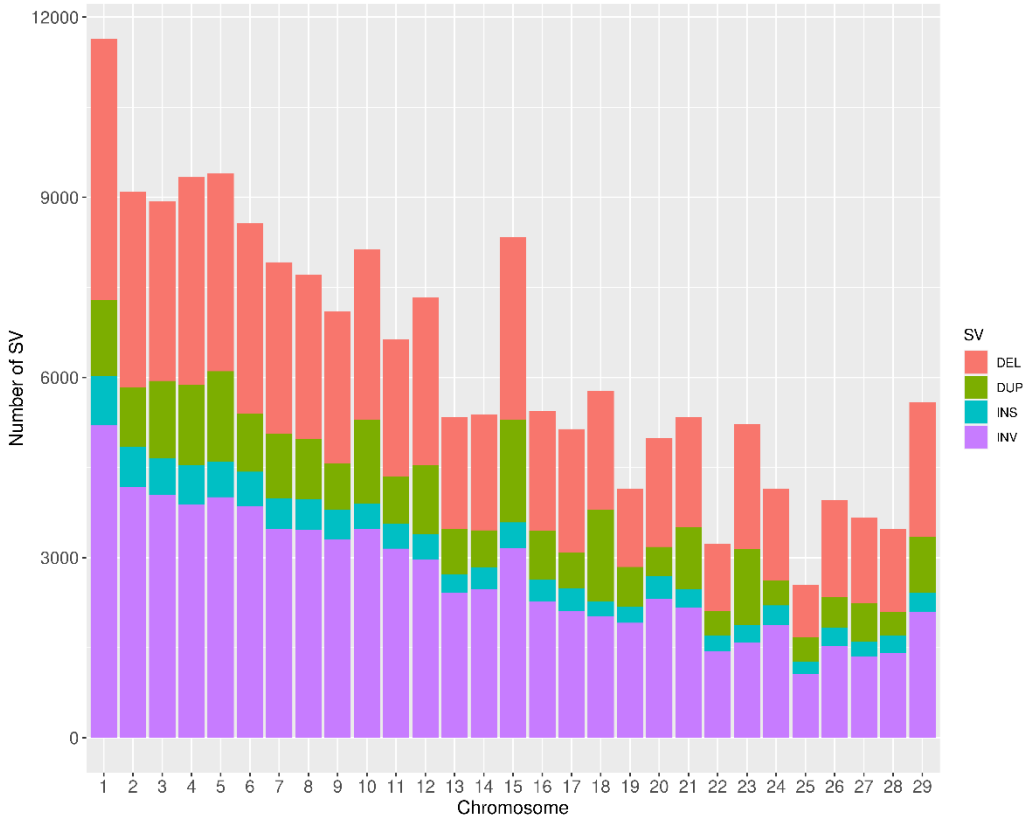

Figure 1. Distribution of the number of deletions (DEL), duplications (DUP), insertions (INS), and inversions (INV) across chromosomes.

GWAS for sequencing data

In this work, we took advantage of whole-genome sequence data to identify potential SVs that are highly associated with DP for 81 sequenced Nellore bulls. Single-trait GWAS revealed 2 significant deletions and 9 significant insertions located on BTA4, BTA7, BTA8, BTA13, BTA15, BTA16, BTA17, BTA20 and BTA28 (Table 2) that together explained 2.61% of the total additive genetic variance for DP. Three potential candidate genes (MAML2, UMAD1 and SAMD8) were annotated on three associated chromosomes (BTA4, BTA15, and BTA28). KEGG pathway analysis indicated that the genes with insertions were involved in 3 KEGG pathways (bta04330: Notch signalling pathway; bta04658: Th1 and Th2 cell differentiation; bta05165: Human papillomavirus infection).

Table 2. Significant SVs detected from genome-wide association study for depigmentation in Nellore cattle.

SV Type[1]	Chr[2]	Start (bp)	End (bp)	Var[3]	*p*-Value	Candidate gene

DEL	13	80111745	80111794	0.5523	1.19E-08	-
DEL	16	16787955	16788618	0.3831	8.57E-08	-
INS	4	15831101	15831102	0.0999	1.18E-06	*UMAD1*
INS	4	82871449	82871450	0.1296	4.29E-06	-
INS	7	48306627	48306628	0.1488	9.71E-07	-
INS	8	57768936	57768937	0.1450	1.36E-06	-
INS	15	14161805	14161806	0.1166	2.35E-07	*MAML2*
INS	16	9833208	9833209	0.3869	8.31E-08	-
INS	17	25309994	25309995	0.1179	8.03E-07	-
INS	20	69296489	69296490	0.4099	2.43E-06	-
INS	28	30910133	30910134	0.1177	6.70E-06	*SAMD8*

[1] SV Type: deletion (DEL); insertion (INS)
[2] Chr: chromosome
[3] Var: additive genetic variance

GWAS for imputed sequence data

The quantile-quantile (QQ) plot for DP is shown in Figure 2. The genomic inflation factor (lambda value) obtained from the QQ plot was 0.97 indicating a certain degree of deviation between the predicted and actual data of DP. Manhattan plot illustrating the results from the GWAS for DP is shown in Figure 3. The GWAS distinguished genomic regions across 9 chromosomes (BTA2, BTA6, BTA9, BTA13, BTA17, BTA22, BTA23, BTA24 and BTA29) comprising of 312 significant SNPs that passed the alternative significance ($p < 1x10^{(-6)}$) p-value threshold and that together explain 2,08% of the total additive genetic variance. The majority of associations was found on BTA6 (n=66) and BTA22 (n=230). The additive genetic variance explained by the significant SNPs located on BTA6 and BTA22 varied from 0.39% to 1.61%, respectively. The significant SNPs

on BTA6 were located between 60,602,592-70,645,707 bp and the most significant (p = 7.53×10−11) SNP explained 0.009% of the genetic variance. The most significant SNP on BTA22 was located at 31,657,074 bp with the p-value of 2.22×10−38, and 0.026% of the genetic variance was explained by this SNP.

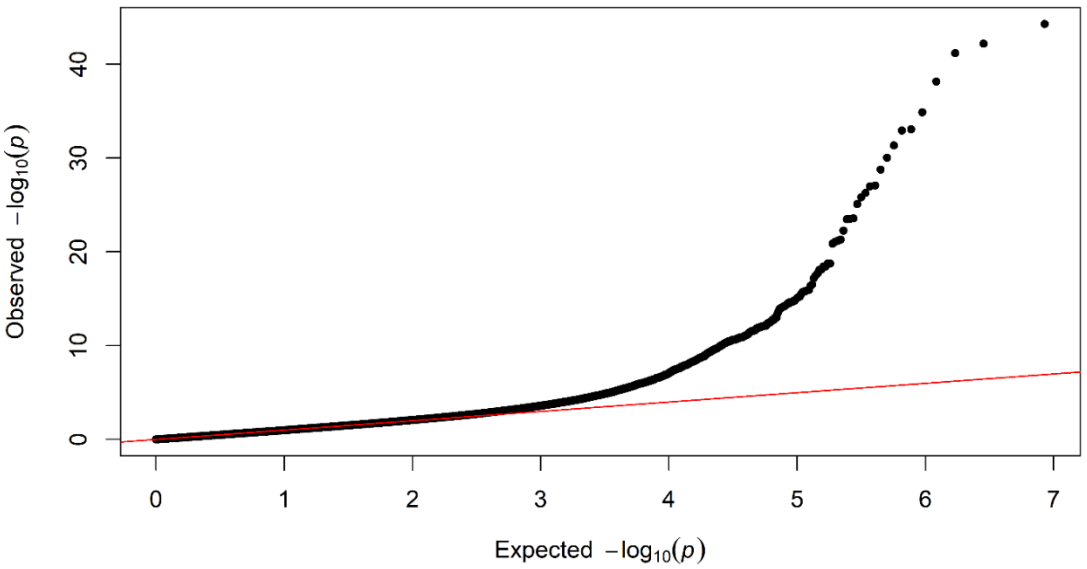

Figure 2. Quantile-quantile plot showing the observed vs. expected p-values outputted from the GWAS analyses for DP. The red line represents the 95% concentration band under the null hypothesis of no association among traits and SNPs. The black dots represent the P-values of the entire study.

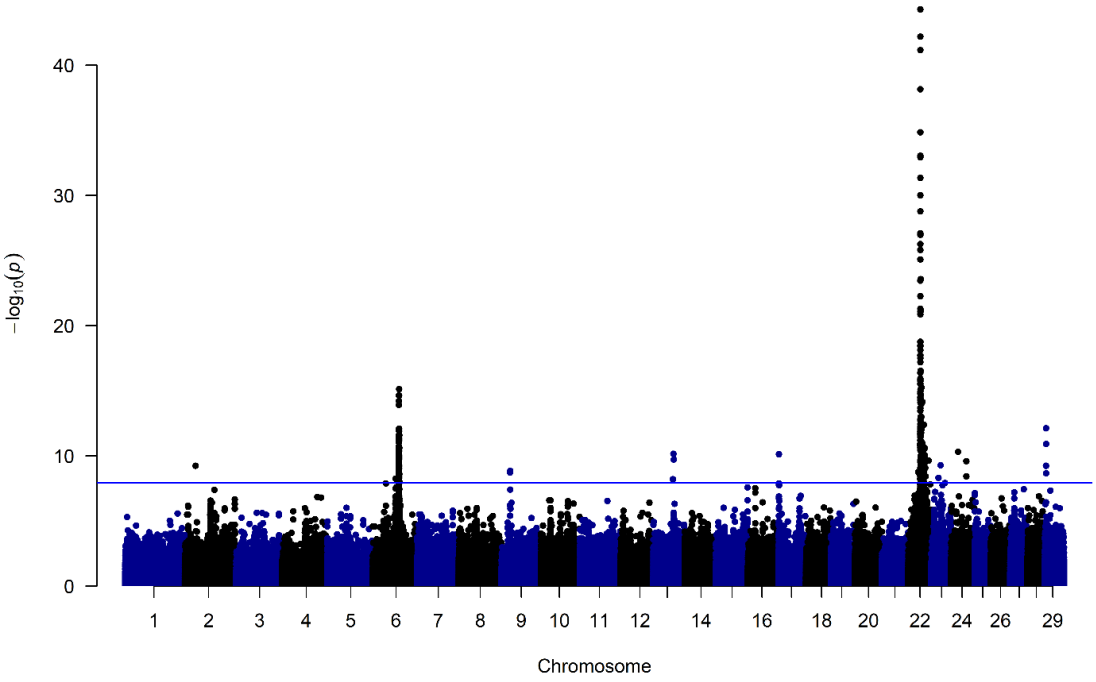

Figure 3. Manhattan plot from the genome-wide association analysis for depigmentation. The blue line indicates the Bonferroni statistical threshold.

The identified potential candidate genes that significantly influence direct genetic effects on DP are in Table 3. Twenty-nine positional candidate genes were identified for DP located on BTA6, BTA9, BTA13, BTA22, BTA23, and BTA24. The QTL enrichment analysis revealed four significant QTLs (FRD-corrected P-value ≤ 0.05) annotated for the positions of significant SNPs (Table 4). These QTLs were associated with 'Exterior' (50%) and 'Milk' (50%) phenotypic classes. The exterior conformation traits previous reported as associated to annotated QTLs include eye area pigmentation and skin pigmentation traits (Pausch et al., 2012).

Table 3. Potential candidate genes related to the identified SNPs significantly associated with depigmentation trait in Nellore cattle.

BTA	Gene Start (bp)	Gene End (bp)	Gene Name*
6	70166692	70254044	*KIT*
6	69723655	69771549	*PDGFRA*

6	70564927	70612683	*KDR*
9	22791492	22958605	*UBE3D*
9	23131908	23352169	*ME1*
13	54606948	54658859	*NTSR1*
22	31285296	31402063	*MDFIC2*
22	32022449	32381786	*FRMD4B*
22	32457271	32491697	*UBA3*
22	31617620	31843533	*MITF*
22	32494168	32525378	*TMF1*
22	32582496	32789965	*TAFA4*
22	32966389	33471943	*TAFA1*
22	33859633	34134874	*SUCLG2*
22	34953094	35066971	*LRIG1*
22	35068281	35215061	*SLC25A26*
22	35758131	36020214	*MAGI1*
22	36760610	36922968	*ADAMTS9*
22	37061676	37323258	*PRICKLE2*
22	37470887	37559158	*ATXN7*
22	37601162	37619598	*THOC7*
22	38430995	38900735	*CADPS*
22	39028434	39358064	*PTPRG*
22	40819274	41605063	*FHIT*
22	43322856	43471741	*SLMAP*

22	44539154	45518463	*ERC2*
23	19185150	19351221	*CLIC5*
23	24893081	24957939	*TRAM2*
24	40329648	40995218	*PTPRM*

Table 4. Significant enriched QTL identified annotated for depigmentation in Nellore cattle.

Trait	**QTL Trait**	**N_QTLs**	***P*-value**
Exterior	Eye area pigmentation Skin pigmentation	2	0.0014
Milk	Milk fat yield	2	0.1024

Gene ontology enrichment analyses was performed for all candidate genes for DP to identify genes involved in known metabolic pathways. Our analyses revealed that the identified potential candidate genes were involved in 100 biological process, 19 cellular component, and 20 molecular function. Pathway analyses was performed for the 312 significant SNPs identified in the GWAS. Pathway analyses showed great diversity in the biological pathways liked to DP, including melanogenesis (bta04916) and melanoma (bta05218).

DISCUSSION

This is the first study investigating genome-wide associations and candidate genes for depigmentation in Nellore cattle based on imputed whole-genome sequence data. Based on low-density SNP chips, association analyses achieved low power and accuracy for detecting loci associated with complex traits. Using whole-genome sequence data, GWAS is optimized to identify genetic variants for complex traits. In this study, a reference population of 151 sequenced Nellore bulls was considered for imputation, resulting in an accuracy of imputation equal to 94% (Fernandes Júnior, 2021). Achieving

an imputation accuracy above 94%, we expected a high power in our GWAS to identify genomic loci associated with complex functional trait in Nellore cattle.

Using 81 sequenced Nellore bulls, this study conducted single-marker association study for DP. Single-trait GWAS revealed 11 significant SVs located on 9 chromosomes. A total of three candidate genes (MAML2, UMAD1 and SAMD8) were annotated and harboring three associated chromosomes. The gene protein MAML2 belongs to the three-member MAML (mastermind-like) family that co-activates Notch receptor-induced transcription. Notch activation regulates skin homeostasis by balancing primarily growth arrest and progressive differentiation processes of keratinocytes, which are associated to skin pigmentation process by accumulating pigmented organelles including melanin (Benito-Martínez et al., 2021).

GWAS with imputed whole-genome sequence data may be effective to detect putative candidate genes. The QQ plot for DP illustrated that some potential bias may still exist and therefore the more stringent Bonferroni threshold was used to identify significant peaks. We identified significant SNPs on BTA2, BTA6, BTA9, BTA13, BTA17, BTA22, BTA23, BTA24 and BTA29 with the largest number of associations on BTA 6 and BTA 22. The significant peaks on BTA6 and BTA22 allocate the genes KIT and MITF. The KIT gene on BTA 6 is one of the main candidate genes for DP and is located between 70166692-70254044 bp. The KIT gene encodes the mast/stem cell growth factor receptor, a large protein with an extracellular domain consisting of five Ig-like domains, a transmembrane region and a tyrosine kinase domain (Ray et al., 2008). KIT plays key roles in melanogenesis and its function is due to its involvement in driving the melanocyte migration from the neural crest along the dorsolateral pathway to colonize the final destination in the skin (Besmer et al. 1993; Yoshida et al. 2001). A large number of KIT mutations have been reported to cause pigmentation defects in human and mouse (Chabot et al. 1988; Geissler et al. 1988; Fleischman et al. 1991; Spritz et al. 1992). Different pigmentation patterns and colours in horses and pigs derive from mutations in the KIT gene (Marklund et al. 1998, 1999; Giuffra et al. 1999, 2002; Pielberg et al. 2002; Brooks & Bailey 2005; Haase et al. 2007). In feline, white spotting locus has been mapped near the KIT gene (Cooper et al. 2006). In cattle, the KIT gene have been previously associated with circumocular pigmentation (Pausch et al., 2012). In addition, the KIT gene has also been described in the literature as associated with coat color in Nelore x Angus crossbreeds (Hulsman Hanna et al., 2014).

MITF is a strong candidate gene for DP and located between 31617620-31843533 bp on BTA22. The MITF gene has been identified as a master transcriptional regulator that orchestrates key developmental and differentiation programs in the melanocyte lineage. MITF together with other three transcription factors (TFE3, TFEB, and TFEC) form the MiT transcription factor family, which dictates tissue-specific gene expression of critical pigmentation enzymes (Hemesath et al., 1994). In humans, mutation in MITF gene leads to Waardenburg syndrome type II and Tietz syndrome, characterized by sensorineural deafness and pigmentation abnormalities of hair and eyes (Hughes, Newton, Liu, & Read, 1994). In mice, a large number of MITF mutations have been reported, causing mainly coat color phenotypes with white spotting or coat color dilution and in some cases microphthalmia and/or osteopetrosis. In cattle, a missense mutation in exon 7 was described that the authors suggest caused the German white Fleckvieh syndrome, an autosomal dominant disorder with complete penetrance characterized by pure white coat color, pigmentless skin and bilateral deafness (Philipp et al. 2011). Additionally, SNPs in the region of MITF have been described to be associated with piebaldism in Holstein and Simmental cattle (Fontanesi et al. 2012).

The analysis of enriched pathways reinforced what has been mentioned for the single genes, namely that in our study the mechanisms regulating DP were mainly those linked to skin pigmentation. Among the GO terms enriched, there were: epithelial cell proliferation (GO:0050673) and epithelial cell apoptotic process (GO:1904019). Melanosomes, “pigmentary units” produced from melanocyte, are transported and accommodate into the epithelial cells which are essential in terms of general regulation of pigmentary tasks in many species.

In this study, two chromosomes were found as being overlapped between the two GWAS performed, namely BTA13 and BTA17, indicating that they might play important and consistent roles associated with the trait of interest. The candidate gene NTSR1 identified from the GWAS using the whole-genome sequence data and located between 54606948-54658859 bp on BTA13 is commonly high expressed in melanoma cells. NTS is a single polypeptide chain and its receptors were reported to contribute to melanocyte proliferation and melanogenesis (Kim et al., 2019).

CONCLUSIONS

Based on the results of the genome-wide association studies presented here, depigmentation is a largely polygenic trait influenced by numerous DNA variants

throughout the genome. Structural variations from 81 Nellore sires, including deletions, insertions, inversions, and duplications were identified. The annotation of SVs in different gene families suggests their relevance on genetic variation for the breed's evolution. Highlighted SNPs significantly associated with DP were found associated with QTL for external classification traits, such as eye area pigmentation and skin pigmentation. Several putatively associated genes were identified as potentially associated with DP. These genes are involved in the regulations of melanocyte differentiation, melanogenesis. The variety of GO terms and functional elements related to depigmentation was significant. We could detect key roles of pathways and biological processes related to epithelial cell proliferation and melanogenesis, circumocular pigmentation, and pigmentation abnormalities. The set of significant genetic variants reported here could be useful for further studies to investigate their incorporation into genomic prediction in order to enhance selection and mating decisions aimed to reduce the incidence of skin depigmentation in Nellore.

ACKNOWLEDGMENTS

To CNPq (#422863/2018-1), FAPESP (#18/20026-8; # 17/10630-2), and Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), in the scope of the Program CAPES-PrInt (#88887.571090/2020-00) for financial support.

REFERENCES

Baker M. (2012). Structural variation: the genome's hidden architecture. Nat Methods, 9:133–137.

Besmer P., Manova K., Duttlinger R., Huang E.J., Packer A., Gyssler C., Bachvarova R.F. (1993) The kit-ligand (steel factor) and its receptor c-kit/W: pleiotropic roles in gametogenesis and melanogenesis. Development Supplement, 125–37.

Brooks S.A. & Bailey E. (2005) Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. Mammalian Genome, 16:893–902.

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 128(14).

Cooper M.P., Fretwell N., Bailey S.J., Lyons L.A. (2006) White spotting in the domestic cat (Felis catus) maps near KIT on feline chromosome B1. Animal Genetics, 37:163–5.

Elsik, CG, Tellam RL, Kim C., Worley Kim C. (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science, 324(5926):522-528. doi:10.1126/science.1169588

Fan B, Onteru SK, Du Z-Q, Garrick DJ, Stalder KJ, Rothschild MF. (2011) Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. PLoS One, 6:e14726. doi:10.1371/journal.pone.0014726.

Fleischman R.A., Saltman D.L., Stastny V., Zneimer S. (1991) Deletion of the c-kit protooncogene in the human developmental defect piebald trait. Proceedings of the National Academy of Sciences of the United States of America, 88:10885–9.

Fontanesi L., Scotti E., Russo V. (2012) Haplotype variability in the bovine MITF gene and association with piebaldism in Holstein and Simmental cattle breeds. Animal Genetics, 43:250–6.

Haase B., Brooks S.A., Schlumbaum A. et al. (2007) Allelic heterogeneity at the equine KIT locus in dominant white (W) horses. PLoS Genetics, 3: e195.

Hemesath TJ, Steingrímsson E, McGill G, Hansen MJ, Vaught J, Hodgkinson CA, Fisher DE (1994). microphthalmia, a critical factor in melanocyte development, defines a discrete transcription factor family. Genes & Development, 8(22):2770–2780. 10.1101/gad.8.22.2770

Hulsman H., Sanders L.L., Riley J.O., Abbey D.G., Gill C.A. (2014). Identification of a major locus interacting with MC1R and modifying black coat color in an F2 Nellore-Angus population. Genetics, Selection, Evolution, 46(1).

Hughes AE, Newton VE, Liu XZ, Read AP (1994). A gene for Waardenburg Syndrome type 2 maps close to the human homologue of the microphthalmia gene at chromosome. Nature Genetics, 7:509.

Kim JS, Park MJ, Kang HY, Hong SP, Park BC, Kim MH. (2019) Neuropeptides Profile and Increased Innervation in Becker's Nevus. Ann Dermatol. 31(2):154-163. doi:10.5021/ad.2019.31.2.154

Magalhães AFB, de Camargo GMF, Fernandes GA Junior, Gordo DGM, Tonussi RL, Costa RB, et al. Genome-Wide Association Study of Meat Quality Traits in Nellore Cattle. PLoS ONE, 11(6): e0157845, 2016. https://doi.org/10.1371/journal.pone.0157845

Ministério da Agricultura, Pecuária e Abastecimento, MAPA. 1995. Instituir o Certificado Especial de Identificação e Produção –CEIP, para bovinos resultantes de cruzamentoplanificado e para animais de raças puras com aptidão para corte. Portaria

Ministerial nº 267, de 04 de Maio de 1995. Lex: Publicado no Diário Oficial da União de 05/05/1995, Seção 1,Página 6409.

Melo T.P., de Camargo G.M.F., de Albuquerque L.G., Carvalheiro R. (2017) Genome-wide association study provides strong evidence of genes affecting the reproductive performance of Nellore beef cows. PLoS ONE, 12(5):e0178551. https://doi.org/10.1371/journal.pone.0178551

Neves H.H.R., Vargas G., Brito L.F., Schenkel F.S., Albuquerque L.G., Carvalheiro R. (2019)Genetic and genomic analyses of testicular hypoplasia in Nellore cattle. PLoS ONE, 14(1):e0211159. https://doi.org/10.1371/journal.pone.0211159

Nicolazzi E.L., Caprera A., Nazzicari N., Cozzi P., Strozzi F. et al. (2015) SNPchiMp v.3: integrating and stan dardizing single nucleotide polymorphism data for livestock species. BMC Genomics, 16:283.

Pausch H., Wang X., Jung S., Krogmeier D., Edel C., Emmerling R., Götz K-U., Fries R. (2012). Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. PLoS ONE, 7(5):e36346.

Philipp U., Lupp B., Mömke S., Stein V., Tipold A., Eule J.C., Rehage J.,& Distl O. (2011) A MITF mutation associated with a dominant white phenotype and bilateral deafness in German Fleckvieh cattle. PLoS ONE, 6:e28857.

Pielberg G., Olsson C., Syvänen A.C. & Andersson L. (2002) Unexpectedly high allelic diversity at the KIT locus causing dominant white color in the domestic pig. Genetics, 160:305–11.

Ray P., Krishnamoorthy N. & Ray A. (2008) Emerging functions of c-kit and its ligand stem cell factor in dendritic cells: regulators of T cell differentiation. Cell Cycle, 7:2826–32.

Roso V.M. & Schenkel F.S. (2006) A computer program to assess the degree of connectedness among contemporary groups. Proceeding of the 8th World Congress on Genetics Applied to Livestock Production. August 13–18; Belo Horizonte, Brazil.

Saatchi M., Schnabel R.D., Taylor J.F., Garrick D.J. (2014) Large effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics, 15(442).

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

Spritz R.A., Giebel L.B. & Holmes S.A. (1992) Dominant negative and loss of function mutations of the c-kit (mast/stem cell growth factor receptor) proto-oncogene in human piebaldism. American Journal of Human Genetics, 50:261–9.

Turner, S. D. (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv10.1101/005165

Vargas G., Neves H.H.R., Camargo G.M.F., Cardoso V., Munari D.P., Carvalheiro R. (2018) Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle. Journal of Animal Science, 96(5):1617-1627. doi: 10.1093/jas/sky079

Vargas G., Schenkel F.S., Brito L.F., Neves H.H.R., Munari D.P., Albuquerque L.G., Carvalheiro R. (2020) Genomic regions associated with principal components for growth, visual score and reproductive traits in Nellore cattle. Livestock Science, 233:103936. doi:10.1016/j.livsci.2020.103936

Whitacre L.K., Hoff J.L., Schnabel R.D., Albarella S., Ciotola F., et al. (2017) Elucidating the genetic basis of an oligogenic birth defect using whole genome sequence data in a non model organism, Bubalus bubalis . Sci. Rep., 7:39719. doi: srep39719

Yang J., Lee S.H., Goddard M.E., Visscher P.M. (2011) GCTA: a tool for Genome wide Complex Trait Analys is. Am J Hum Genet. 88(1):76 82.

Yoshida H., Kunisada T., Grimm T., Nishimura E.K., Nishioka E., Nishikawa S.I. (2001) Review: melanocyte migration and survival controlled by SCF/c-kit expression. Journal of Investigative Dermatolology Symposium Proceedings 6, 1–5.

4.2 Interação genótipo-ambiente em bovinos Nelore para características de crescimento e escores visuais com utilização de dados de sequência.
Projeto de doutorado; PPG: Genética e Melhoramento Animal, FCAV UNESP

Aluno: Ivan Carvalho Filho

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.

Seleção genômica considerando a interação genótipo ambiente para características de importância econômica em bovinos Nelore

A seleção genética pode se tornar mais assertiva quando diferentes sistemas de produção são considerados e quando a informação genômica é incluída. Assim, o objetivo deste estudo foi avaliar o ganho em acurácia da predição em três cenários e compará-los, dois dele usando BLUP genômico *single-step*, um considerando os marcadores presentes no SNPChip HD (ssGBLUP_HD), outro a sequência completa do genoma (ssGBLUP_SEQ) e um cenário com apenas pedigree (BLUP), na presença de interação genótipo-ambiente (GxE). Os dados fenotípicos para ganho de peso pós-desmame (PWG), idade ao primeiro parto (AFC), peso ao sobreano (YW) e perímetro escrotal (SC) foram obtidos do conjunto de dados Nelore Alliance. Havia 1.578.591 animais no pedigree, dos quais 51.485 foram genotipados para 460.578 SNPs após o controle de qualidade para o ssGBLUP_HD. Para o ssGBLUP_SEQ após controle de qualidade, permaneceram 2.437.948 SNPs. Para cada característica foi realizada consistência de dados separadamente, sendo em comum, a presença de ao menos 20 animais por grupo de contemporâneos (CG). Efeitos de CG previamente estimados usando um modelo animal misto (sem modelagem GxE) foram usados no modelo de norma de reação (RN) para definir os gradientes ambientais (EG). A sensibilidade genética à variação ambiental foi avaliada por meio do ajuste de três modelos RN lineares diferentes, um considerando apenas o pedigree (BLUP) e outros dois considerando também a informação genômica (ssGBLUP_HD e ssGBLUP_SEQ). A validação da predição foi realizada para touros jovens genotipados, sendo removido os registros de progênie nos dados reduzidos, mas com pelo menos um registro de progênie nos dados completos. Os modelos foram comparados usando a acurácia realizada com base no método de regressão linear. Foram observadas reclassificações substanciais entre os animais e heterogeneidade da variância genética nos diferentes EG, o que sugere a presença de GxE. Os resultados para os coeficientes de regressão do modelo RN

mostraram que a inclusão de informações genômicas com auxílio dos marcadores presentes, tanto no ssGBLUP_HD e ssGBLUP_SEQ, aumentou a acurácia de predição para todas as características nos coeficientes de regressão dos modelos (intercepto e inclinação) em relação ao BLUP tradicional, sem inclusão de marcadores. Além disso, a acurácia de predição em diferentes EG aumentou à medida que o EG melhorou (Figura 1), quando os valores genéticos estimados para touros jovens genotipados foram considerados. Em conclusão, o ssGBLUP_SEQ apresentou maior acurácia de predição do que o BLUP em todos os EG, indicando que a implementação da seleção genômica contabilizando GxE seria benéfica nesta população Nelore.

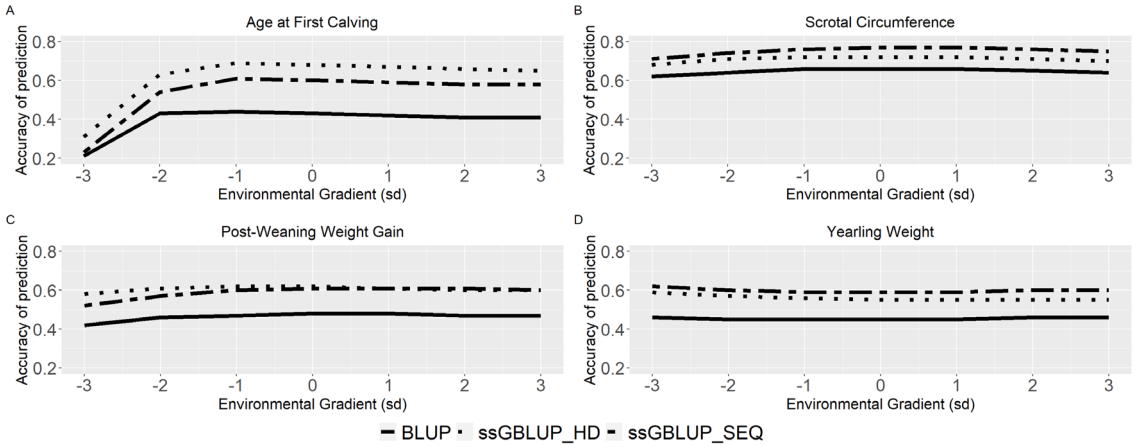

Figura 1. Acurácia de predição ao longo dos gradientes ambientais para as características idade ao primeiro parto (AFC), perímetro escrotal (SC), ganho de peso pós desmame (PWG) e peso ao sobreano (YW), considerando o modelo sem inclusão de informação genômica (BLUP, linha contínua), com inclusão da informação genômica proveniente do SNPchip HD (ssGBLUP_HD, linha pontilhada) e com inclusão da informação genômica proveniente da sequência (ssGBLUP_SEQ, linha tracejada).

Estudos de associação genômica ampla considerando a interação genótipo ambiente para características de importância econômica em bovinos Nelore

A inclusão de marcadores moleculares na estimativa de normas de reação permite a identificação de genes candidatos ao longo dos gradientes ambientais por meio de estudos de associação do genoma. O objetivo deste trabalho é realizar estudos de associação genômica ampla (GWAS) usando dados de sequenciamento para mapear genes candidatos a características em diferentes gradientes ambientais. Os dados fenotípicos para ganho de peso pós-desmame (PWG), idade ao primeiro parto (AFC), peso ao sobreano (YW) e perímetro escrotal (SC) foram obtidos do conjunto de dados Nelore Alliance. Havia 1.578.591 animais no pedigree, dos quais 151 touros foram sequenciados, e após imputação, 51.485 animais passaram a ter a informação da sequência de todo o genoma. Devido as limitações de software e hardware, foi estabelecido que no máximo 20.000 animais genotipados teriam seus genótipos incorporados as análises. Após os

critérios de controle de qualidade estipulados, o número de SNPs variou de 2 milhões de marcadores a 7 milhões, dependendo da característica e do número de animais genotipados utilizados. Foi realizado GWAS considerando dois modelos de norma de reação heterocedásticos distintos, um linear e um spline linear-linear, com nó no gradiente zero. Os estudos de associação do genoma estão sendo realizados usando a abordagem GBLUP de passo único. Para AFC, PWG e SC, resultados prévios mostram que houve pequena diferença entre os coeficientes da regressão e entre os gradientes estudados quanto a significância dos SNPs. Para YW as análises estão em andamento. Para todas as características, uma rede de genes agrupados funcionalmente em termos biológicos será construída para investigar evidências da associação de processos biológicos relacionados à sensibilidade à variação ambiental.

Projeto de doutorado: estudo de associação genômica ampla e análise funcional com dados de sequenciamento completo para hipoplasia testicular e conformação de pernas e pés em bovinos da raça Nelore.

PPG: Genética e Melhoramento Animal, FCAV UNESP

Aluno: Thales de Lima Silva

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.

4.3 Estudo quantitativo de regiões codificantes e regulatórias para características de crescimento em bovinos Nelore.

PPG: Genética e Melhoramento Animal, FCAV UNESP

Aluno: Miller de Jesus Teodoro

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.

Características de crescimento têm sido usadas na seleção de bovinos de corte no Brasil, estando diretamente associadas com a produtividade e a lucratividade dos rebanhos. O acompanhamento do peso corporal do animal, em diferentes idades, permite avaliar a eficiência da atividade e auxilia na tomada de decisões futuras. Com os avanços tecnológicos da genômica, a genotipagem e o sequenciamento dos animais possibilitaram uma avaliação mais acurada dos indivíduos. O desenvolvimento de chips de alta densidade de marcadores do tipo polimorfismo de nucleotídeos únicos (SNPs) possibilitou a utilização de estudos de associação genômica ampla (GWAS), em que é possível identificar variações genéticas associadas a fenótipos de interesse. Com isso, o objetivo deste estudo será identificar regiões do cromossomo 14 associadas com

características de crescimento em bovinos da raça Nelore, além de verificar quais classes funcionais de regiões codificadoras e regulatórias do genoma explicam maior proporção da variância. Para atingir os objetivos propostos, serão usados dados de sequenciamento de 150 touros da raça Nelore e genótipos de 3.800 animais, pertencentes a diferentes programas de melhoramento genético, e registros fenotípicos de 355.524 animais para peso corporal aos 205, 365, 550 e 720 dias de idade. Ao todo, serão consideradas as anotações de 20 classes funcionais. As análises de componentes de variância serão conduzidas por duas metodologias, sendo a primeira um teste de permutação aleatória das variantes e, a segunda, a estimação da variância explicada pelas classes funcionais pelo método da máxima verossimilhança restrita.

4.4 Estudo de associação genômica ampla e análise funcional com dados de sequenciamento completo para hipoplasia testicular e conformação de pernas e pés em bovinos da raça Nelore.
PPG: Genética e Melhoramento Animal, FCAV UNESP

Aluno: Thales de Lima Silva

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.

4.5 Identificação de polimorfismos pleiotrópicos para características reprodutivas em bovinos da raça Nelore.
PPG: Genética e Melhoramento Animal, FCAV UNESP

Aluno: Delvan Alves Silva

Atividades:

- Colaboração na realização das análises e no manuscrito que está em processo de preparação para submissão.