**UNIVERSIDADE ESTADUAL PAULISTA – UNESP**
**FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS**
**CÂMPUS DE JABOTICABAL**

# APPLYING MACHINE LEARNING METHODS FOR GENOMIC ANALYSIS OF REPRODUCTIVE TRAITS IN NELLORE CATTLE

**Anderson Antonio Carvalho Alves**
Zootecnista

**2019**

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP**
**FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS**
**CÂMPUS DE JABOTICABAL**

# APPLYING MACHINE LEARNING METHODS FOR GENOMIC ANALYSIS OF REPRODUCTIVE TRAITS IN NELLORE CATTLE

**Anderson Antonio Carvalho Alves**

**Orientadora: Profa. Dra. Lucia Galvão de Albuquerque**

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

**2019**

A474a

Alves, Anderson Antonio Carvalho
Applying machine learning methods for genomic analysis of reproductive traits in Nellore cattle / Anderson Antonio Carvalho Alves. -- Jaboticabal, 2019
114 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientadora: Lucia Galvão de Albuquerque

1. Beef cattle. 2. Genetics. 3. Statistical methods. 4. Neural networks (Computer science). I. Título.

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

Essa ficha não pode ser modificada.

**UNIVERSIDADE ESTADUAL PAULISTA**

**Câmpus de Jaboticabal**

## CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: APPLYING MACHINE LEARNING METHODS FOR GENOMIC ANALYSIS OF REPRODUCTIVE TRAITS IN NELLORE CATTLE

**AUTOR: ANDERSON ANTONIO CARVALHO ALVES**
**ORIENTADORA: LUCIA GALVÃO DE ALBUQUERQUE**

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

Profa. Dra. LUCIA GALVÃO DE ALBUQUERQUE
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

Prof. Dr. RICARDO VIEIRA VENTURA
Departamento de Nutrição e Produção Animal/FMVZ-USP / Pirassununga/SP

Pesquisador Dr. ROBERTO CARVALHEIRO
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Pesquisadora Dra. MARIA EUGÊNIA ZERLOTTI MERCADANTE
Instituto de Zootecnia / Sertãozinho/SP

Prof. Dr. DANISIO PRADO MUNARI
Departamento de Ciências Exatas / FCAV / Unesp - Jaboticabal

Jaboticabal, 28 de novembro de 2019

## DADOS CURRICULARES DO AUTOR

**ANDERSON ANTONIO CARVALHO ALVES** – nascido em 14 de janeiro de 1991, na cidade de Duque de Caxias, Rio de Janeiro, filho de Mirian Carvalho e Inácio de Oliveira Alves. Iniciou a graduação em Zootecnia em agosto de 2009, na Universidade Estadual Vale do Acaraú – UVA, Campus Betânia em Sobral-CE. Durante a graduação, foi monitor das disciplinas Fisiologia da Reprodução Animal e Melhoramento Genético Animal I, estagiou por dois anos na Embrapa Caprinos e Ovinos, atuando no programa de melhoramento de caprinos leiteiros Capragene®. Obteve o título de Bacharel em Zootecnia em janeiro de 2014. Em fevereiro do mesmo ano, ingressou no curso de mestrado do Programa de Pós-Graduação em Zootecnia, com área de concentração em Produção e Melhoramento Animal, na Universidade Federal do Ceará – UFC, Campus do Pici em Fortaleza-CE, sob orientação do Prof. Dr. Raimundo Nonato Braga Lôbo. Obteve o título de Mestre em Zootecnia em dezembro de 2015. Em fevereiro de 2016 iniciou o Doutorado no Programa de Pós-Graduação em Genética e Melhoramento Animal da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista Júlio de Mesquita Filho, Campus de Jaboticabal, sob orientação da Profa. Dra. Lucia Galvão de Albuquerque. Durante o doutorado foi bolsista CAPES e FAPESP. Atualmente é professor EBTT do Instituto Federal do Maranhão – IFMA, atuando nos cursos técnicos em agropecuária, além dos cursos de bacharelado em Agronomia e Zootecnia, em que ministra as disciplinas de Genética, Estatística Básica e Experimentação Estatística.

“All our science, measured against reality, is primitive and childlike - and yet it is the most precious thing we have.”

(Albert Einstein)

Aos meus pais, Inácio de Oliveira Alves e Mirian Carvalho. Aos meus irmãos, Arlon e Andressa. À minha noiva Rebeka Magalhães. Aos amigos de longa data e a todos os meus professores, por todos os momentos de paciência, convivência e aprendizado.

DEDICO

## AGRADECIMENTOS

Gostaria de expressar inicialmente minha gratidão à vida, por sua beleza, grandeza e complexidade, pelo nitrogênio em nosso DNA, o cálcio nos nossos dentes, o carbono em nossas células, o ferro em nosso sangue, por cada átomo que compõe seres vivos, planetas e galáxias.

Sou eternamente grato à minha família, em especial aos meus pais Inácio de Oliveira Alves e Mirian Carvalho Oliveira e meus irmãos Arlon e Andressa pelo apoio incondicional em todas as minhas decisões, pelos ensinamentos e amor durante todos esses anos, apesar de toda a distância.

Minha noiva Rebeka Magalhães é parte importante dessa conquista, me sinto muito afortunado de poder contar com sua cumplicidade e apoio em todos os meus projetos. Obrigado por estar sempre ao meu lado nos momentos bons e ruins e por me permitir conhecer o amor em sua forma mais pura e simples. Também dedico estes agradecimentos ao meu pequeno amigo peludo, Snoopy, por me esperar pacientemente voltar do trabalho e me ensinar todos os dias que precisamos de pouco para sermos felizes. Vocês são a razão da minha felicidade, amo vocês.

Agradeço aos meus cunhados Alex, Ariel, Demétrios e Philipe, aos meus sogros José Maria e Maria de Fátima e à Sheila, Lara e Larissa por serem uma segunda família para mim em Fortaleza.

Não poderia deixar de agradecer aos orientadores que passaram por minha carreira acadêmica nas etapas anteriores, Dra. Ana Bezerra, Dr. Olivardo Facó e Dr. Raimundo Lôbo, pela grande contribuição no meu aprendizado, meu muito obrigado.

Gostaria de agradecer imensamente à minha orientadora de doutorado, Dra. Lucia Galvão de Albuquerque, pela oportunidade concedida, paciência, confiança e pelas valiosas contribuições para minha vida pessoal e profissional ao longo desses anos.

Aos amigos de longa data: Neto, Orlando, André, Rodrigo e Pedro pela convivência e irmandade e também aos amigos da banda Turgal (Fagner, Ítalo, João Guilherme, Neto, Paulo Roberto e Tiago) pelos momentos de diversão, companheirismo e aprendizado musical.

Aos docentes que contribuem para o programa de Pós-graduação em Genética e Melhoramento animal, por compartilharem seu conhecimento, em especial aos professores Danísio Munari, Roberto Carvalheiro, Henrique Nunes, Ricardo da Fonseca e Guilherme Rosa.

Aos colegas e amigos dos departamentos de Zootecnia e Ciências Exatas, Ana Cristina, Ana Fabricia, André Mauric, André Vieira, Ândrea, Andres Chaparro, Bruna, Dani Beraldo, Diogo Osmar, Gabriela, Larissa, Lucas Sales, Lucas Lopes, Lucio, Malane, Natália, Patrícia, Rafael Espigolan, Rafael Watanabe, Samuel, Samla, Tiago Bresolin, Thaise, William e aos demais, pelos momentos de café, compartilhamento de experiência e confraternizações.

Aos companheiros do Instituto Federal do Maranhão (IFMA) do campus São Raimundo das Mangabeiras, em especial aos amigos Luís Paulo, Marcones, Clemeson Vale, Antonia Lima, Marlon Cardozo, Sigfran, Guilherme Ramon, Well Max, Marlon da Costa, Adeneide, Carine, Júlio e Felipe Saraiva.

À Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), pela oportunidade concedida. Tenho profunda admiração e respeito por esta instituição.

Aos membros das bancas examinadoras de qualificação e de defesa pelas valiosas sugestões propostas para o aprimoramento deste trabalho.

Finalmente, é fundamental ressaltar que a execução deste estudo só foi possível graças ao suporte financeiro concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e pela Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (processo 16/24227-2).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

**SUMÁRIO**

Página

# APLICAÇÃO DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA PARA ANÁLISE GENÔMICA DE CARACTERÍSTICAS REPRODUTIVAS EM BOVINOS NELORE

**RESUMO –** A seleção de animais geneticamente superiores com base na informação genômica tem sido uma tendência crescente e promissora em programas de melhoramento. No entanto, os principais métodos de predição genômica envolvem modelos paramétricos, que em sua maioria, assumem somente variância aditiva para o efeito dos marcadores, ignorando-se possíveis relações não-lineares. A consideração de tais efeitos pode ser importante para melhorar a habilidade de predição em características com arquitetura genética complexa. Recentemente, tem crescido o interesse em métodos de predição semi e não paramétricos. Dentro desse contexto, os métodos de aprendizagem de máquina tais como Redes Neurais Artificiais (ANN), "Random Forest" (RF) e "Support Vector Machines" (SVM) são alternativas interessantes. Os objetivos do presente estudo foram: i) Comparar o desempenho preditivo do modelo "Genomic Best Linear Unbiased Predictor" (GBLUP) e de métodos de aprendizagem de máquina em populações simuladas de bovinos de corte, apresentando diferentes níveis para efeitos de dominância; ii) Investigar a habilidade de predição de diferentes métodos de aprendizagem de máquina para predição genômica de características reprodutivas em bovinos da raça Nelore; iii) Desenvolver um estudo de associação genômica ampla (GWAS) utilizando a metodologia "Random Forest", a fim de buscar genes candidatos para idade ao primeiro parto em novilhas da raça Nelore. No primeiro estudo, o genoma simulado compreendeu um painel de SNPs ("Single Nucleotide Polymorphisms") com densidade de 50k e 300 QTLs ("Quantitative Trait Loci"), espalhados aleatoriamente ao longo de 29 cromossomos. Foram simuladas ao todo seis características, considerando-se diferentes valores de herdabilidade no sentido restrito e amplo. No cenário puramente aditivo e com baixa herdabilidade ($h^2$ = 0,10), a habilidade de predição utilizando o método GBLUP foi levemente superior em relação aos outros métodos (aproximadamente de 0,8% a 5,0%), ao passo que as ANN obtiveram melhor acurácia nos cenários com moderada herdabilidade ($h^2$ = 0,30). As acurácias para os efeitos de dominância variaram entre 0,180 e 0,350 no modelo GBLUP considerando a matriz de relacionamento de dominância (GBLUP-D), entre 0,062 e 0,185 para o RF e foram nulas utilizando-se os métodos ANN e SVM. Entre os métodos de aprendizagem de máquina, apenas o RF foi capaz de capturar implicitamente os efeitos de dominância, resultando em maiores acurácias de predição para os valores genéticos totais e fenotípicos quando a variância devido ao efeito de dominância aumentou. No segundo estudo, dados referentes a bovinos da raça Nelore nascidos entre 1984 e 2015 foram utilizados. As características estudadas foram Idade ao Primeiro Parto (AFC), Circunferência Escrotal (SC), Prenhez Precoce (EP) e Habilidade de Permanência (STAY). Após o controle de qualidade, o número de animais com genótipos e de marcadores SNP disponíveis foram respectivamente, 2.342 e 321.419 (AFC), 4.671 e 309.486 (SC), 3.356 e 319.108 (EP) e 2.681 e 319.619 (STAY). A habilidade preditiva de diferentes métodos de aprendizagem de máquina tais como "Support Vector Regression" (SVR), "Bayesian Regularized Artificial Neural Network" (BRANN) e RF foi avaliada. Os resultados foram comparados aos obtidos pelos modelos paramétricos GBLUP e BLASSO ("Bayesian Least Absolute Shrinkage and Selection Operator"). Para o modelo SVR, investigou-se a influência de diferentes

valores para o parâmetro de largura de banda do kernel na habilidade de predição do modelo. Para o modelo BRANN, diferentes números de neurônios na camada oculta (de 1 a 4 neurônios) foram examinados para se identificar a melhor arquitetura de rede. Além disso, duas estruturas de informação genômica foram testadas como informação de entrada no modelo BRANN, a matriz de relacionamento genômica (G) e a matriz de componentes principais (PC). A habilidade de predição dos modelos foi avaliada por meio de um esquema de validação cruzada em 5 “folds”. As acurácias obtidas foram de baixas a moderadas de acordo com a característica e modelos considerados, variando entre 0,555 e 0,625 (AFC), 0,268 e 0,359 (SC), 0,573 e 0,666 (EP) e entre 0,517 e 0,618 (STAY). O modelo SVR obteve desempenho ligeiramente superior em relação aos métodos paramétricos (GBLUP e BLASSO) para todas as características avaliadas, aumentando a acurácia de predição da AFC em aproximadamente 5,1% e 3,7%, quando comparados aos modelos GBLUP e BLASSO, respectivamente, e em 7,2% para SC, 3,4% para EP e 5% para STAY quando comparado aos resultados obtidos por ambos GBLUP e BLASSO. Por outro lado, os modelos RF, BRANN_G e BRANN_PC não apresentaram habilidade de predição competitiva com os métodos tradicionais, apresentando menor acurácia de predição e maiores erros de predição para todas as características. Os resultados indicam que o SVR é um método adequado para a predição de valores genéticos genômicos para características reprodutivas em bovinos da raça Nelore, apresentando melhor habilidade de predição e eficiência no tempo de computação em relação as metodologias paramétricas estudadas. Além disso, o valor mais adequado para o parâmetro de largura de banda do kernel no método SVR dependeu da característica avaliada, desse modo, a correta predefinição desse parâmetro na fase de treinamento do modelo é aconselhável. Por último, um estudo de associação genômica ampla foi realizado utilizando a abordagem RF, a fim de se identificar genes candidatos para a idade ao primeiro parto em bovinos da raça Nelore. Os valores examinados para o parâmetro $M_{try}$ (ou seja, o número de SNPs testados em cada nó das árvores) foram 1, $\sqrt{p}$, $0.01p$ e $0.1p$, em que $p$ representa o número total de SNPs. Os parâmetros que produziram o menor erro quadrático nos dados *out-of-bag* ($MSE_{OOB}$) foram mantidos para análises posteriores. Foram realizadas 5 análises independentes com diferentes sementes de inicialização do algoritmo e os escores de importância dos SNPs foram computados como a média das 5 análises. Foram identificados 118 SNPs associados com AFC, localizados em oito cromossomos autossômicos (BTA 3, 5, 10, 11, 18, 21, 25 e 27). No total, 23 regiões não sobrepostas cobriram 172 genes candidatos para AFC. Regiões genômicas previamente associadas com características de fertilidade e crescimento em bovinos Nelore foram reportadas neste estudo, o que reforça a efetividade do RF como um método para a varredura inicial de regiões candidatas associadas com características complexas. O estudo de associação baseado no método RF e a análise funcional apontaram genes candidatos com funções chave na regulação da fertilidade, incluindo a pré-implantação de embriões e seu desenvolvimento, viabilidade embrionária, maturação de células germinais masculinas e reconhecimento de feromônios.

**Palavras-chave**: bovinos de corte, fertilidade, genes candidatos, métodos não-paramétricos, precocidade, predição genômica

**APPLYING MACHINE LEARNING METHODS FOR GENOMIC ANALYSIS OF REPRODUCTIVE TRAITS IN NELLORE CATTLE**

**ABSTRACT –** The selection of genetically superior animals based on genomic information has been an increasing and promising trend in breeding programs. However, the main methods used for genome-enabled prediction involve parametric models that mostly assume only additive variance for markers effects, ignoring possible nonlinear relationships. Accounting for such effects may be important to improve the predictive ability for traits with complex genetic architecture. The interest in semi and non-parametric prediction methods has recently increased. Within this context, machine learning methods such as Artificial Neural Networks (ANN), Random Forest (RF) and Support Vector Machines (SVM) are an interesting alternative. The aims of the present study were: i) To compare the predictive performance of Genomic Best Linear Unbiased Predictor (GBLUP) and machine learning methods in simulated beef cattle populations presenting different degrees of dominance; ii) To investigate the predictive ability of different machine learning for genome-enabled prediction of reproductive traits in Nellore cattle and compare their performance with parametric approaches (GBLUP and BLASSO); iii) To perform a genome-wide association study (GWAS) using the Random Forest approach for scanning candidate genes for age at first calving in Nellore heifers. In the first study, the simulated genome comprised 50k single nucleotide polymorphisms (SNPs) and 300 QTL (Quantitative Trait Loci), both biallelic and randomly distributed across 29 chromosomes. A total of six traits were simulated considering different values for the narrow and broad-sense heritability. In the purely additive scenario with low heritability ($h^2$ = 0.10), the predictive ability obtained using GBLUP was slightly higher than the other methods (approximately 0,8% to 5,0%) whereas ANN provided the highest accuracies for scenarios with moderate heritability ($h^2$ = 0.30). The accuracies of dominance deviations varied from 0.180 to 0.350 in the GBLUP model considering the dominance genomic relationship matrix (GBLUP-D), from 0.062 to 0.185 in the RF and were null using ANN and SVM methods. Among machine learning methods, only the RF was capable to cover implicitly dominance effects without increasing the number of covariates in the model, resulting in higher accuracies for the total genetic and phenotypic values as the dominance ratio increased. In the second study, data of Nellore cattle from commercial herds born between 1984 and 2015 were used. The studied traits were Age at First Calving (AFC), Scrotal Circumference (SC), Early Pregnancy (EP) and Stayability (STAY). After quality control, the number of genotyped animals and SNP markers available were respectively, 2,342 and 321,419 (AFC), 4,671 and 309,486 (SC), 3,356 and 319,108 (EP) and 2,681 and 319,619 (STAY). The predictive ability from different machine learning models such as Support Vector Regression (SVR), Bayesian Regularized Artificial Neural Network (BRANN) and RF, was assessed. Results were compared with that obtained using GBLUP and BLASSO (Bayesian Least Absolute

Shrinkage and Selection Operator) parametric models. For the SVR, the influence of different kernel bandwidth parameter values on the model predictive ability was assessed. In the BRANN models, different numbers of neurons in the hidden layer (1 to 4 neurons) were examined to assess the best ANN architecture. Further, two genomic structures were assessed as input information in the BRANN model, the marker-based genomic relationship matrix (G) and the principal components scores matrix (PC). The predictive ability of the studied models was evaluated by a 5-fold cross-validation scheme. The average accuracies were from low to moderate according to the trait and model considered, ranging between 0.555 and 0.625 (AFC), 0.268 and 0.359 (SC), 0.573 and 0.666 (EP) and 0.517 and 0.618 (STAY). The SVR provided slightly better performance than the parametric models for all traits, increasing the prediction accuracy for AFC around 5.1% and 3.7% compared to GBLUP and BLASSO models, respectively, and around 7.2% for SC, 3.4% for EP and 5% for STAY, comparing to both GBLUP and BLASSO. In contrast, the RF, BRANN_G and BRANN_PC models did not present competitive predictive ability compared to the benchmark approaches, presenting lower prediction accuracies and higher MSE for all traits. Our results indicate that the SVR is a suitable method for genomic breeding values prediction for reproductive traits in Nellore Cattle, presenting better predictive ability and computational time efficiency than the studied parametric approaches. Further, the optimal kernel bandwidth parameter in the SVR model was trait-dependent, thus, the correct pre-definition of this parameter in the training phase is advisable. Lastly, a genome-wide association study (GWAS) was performed using the RF approach for scanning candidate genes for AFC in Nellore cattle. The assessed values for the $M_{try}$ parameter (*i.e*. the number of SNPs to search at each node) were 1, $\sqrt{p}$, $0.01p$ and $0.1p$, in which $p$ represents the total number of SNPs. The RF parametrization which produced the lowest mean squared error in the out-of-bag data ($MSE_{OOB}$) was maintained for further analysis. We run five independent analyses with different initialization seeds for the algorithm and the SNPs importance scores were averaged. There were identified 118 SNPs associated with AFC, located over eight autosomes (BTA 3, 5, 10, 11, 18, 21, 25 and 27). In total, 23 non-overlapping genomic regions embedded 172 candidate genes for AFC. Genomic regions previously associated with fertility and growth traits in Nellore cattle were reported in the present study, which reinforces RF effectiveness for pre-screening candidate regions associated with complex traits. The RF-based genome-wide scan and functional analysis highlighted candidate genes with key roles in fertility, including embryo pre-implantation and development, embryonic viability, male germinal cell maturation and pheromone recognition.

**Keywords**: beef cattle, fertility, candidate genes, nonparametric methods, precocity, genomic prediction

# CHAPTER 1 – General considerations

## 1 Introduction

The use of molecular information has provided a paradigm shift in animal breeding, allowing us to investigate the genetic basis of complex traits at the genomic level. Due to the constant advance and cost reduction on genotyping technologies, currently, it is possible to identify and sequence a large number of single nucleotide polymorphisms (SNPs) throughout the whole genome. In this regard, the selection of genetically superior animals based on genomic information has been an increasing and promising trend in breeding programs.

Such an approach, termed genomic selection (GS), relies on the assumption that genetic markers are in linkage disequilibrium with quantitative trait loci (QTL). Hence, genomic predictions of breeding values are derived considering the estimated effects of thousands of SNP genotypes spread throughout the genome of a selection candidate sire. This methodology has allowed predicting breeding values of young animals, with not yet observed phenotypes, with higher accuracy than the traditional pedigree-based approach (Meuwissen et al., 2001; De Los Campos et al., 2013). Analogously, it is possible to use the genotype information to identify markers directly associated with the expression of an interest trait, an approach known as genome-wide association study (GWAS).

However, with the genomic information advent, new computational and statistical challenges have emerged. The simultaneous association between thousands of markers (p) and the observed variability in the individuals (phenotypes), due to a limited number of available samples (n) may generate restrictions on data estimation and inference. In order to deal with the “large p, small n” problem, most statistical methods adopt some type of variable selection or shrinkage estimation (De Los Campos et al., 2013). Arguably, parametric approaches have been shown to be effective for whole genome-enabled prediction and GWAS purposes. Notwithstanding, such methodologies tend to assume strong assumptions about the genetic architecture of the trait that not always hold in practice. For instance, ridge regression best linear unbiased prediction assumes that all marker effects share a common variance in their

distribution (Okut et al., 2013). On the other hand, in the linear Bayesian regression models, different prior distributions are assigned to the marker effects in order to perform differential shrinkage. Nonetheless, a posteriori inference depends heavily on the prior assumptions used in the model formulation (Gianola et al., 2009). Additionally, the aforementioned approaches mostly assume only additive inheritance and do not account for complex interactions among genes and other non-linear effects that may exist (Gianola et al., 2011).

In recent years there has been a growing interest in using semi and non-parametric methods for genome-enabled prediction (Gianola and De Los Campos, 2008; De Los Campos et al., 2009; De Los Campos et al., 2010; Long et al., 2010), mainly due to the theoretical flexibility offered by such models into cover complex relationships between markers and phenotype, which can potentially enhance prediction ability. In this context, interesting alternatives are machine learning (ML) methods, such as artificial neural networks (ANN), random forest (RF) and support vector machines (SVM). Latter methods provide an ensemble of attributes that make them suited for dealing with complex data and have been successfully applied in a wide range of studies on genomics and other bioinformatics branches (Yang et al., 2010; Upstill-Goddard et al., 2012; Libbrecht and Noble, 2015).

For genome-enabled prediction, a particular advantage is that ML methods are model-free, in other words, there is no necessity to impose a specific genetic structure, so that, no assumptions are required about the genetic architecture of the target traits. In practice, previous studies, especially using simulated data, indicate that such approaches provide similar or superior prediction abilities when compared to the parametric models (González-Recio and Forni, 2011; Ogutu et al., 2011; Ehret et al., 2015; Ghafouri-Kesbi *et al*., 2016). However, results considering empirical applications of machine learning algorithms on real data still have been few explored, particularly for Zebu cattle.

In Brazil, some efforts have been taking into account for the implantation of genomic selection on commercial livestock, especially for the Nellore cattle (Carvalheiro, 2014), a commonly explored breed under commercial conditions and rather widespread on that country. Nevertheless, despite its economic importance, reproductive traits remain few explored as selection criteria, mostly due to the fact that

such traits in general exhibit low heritability estimates, are sex-limited and are expressed belatedly (Dias et al., 2004; Boligon et al., 2010). Even so, one must consider the inclusion of reproductive traits in beef cattle breeding programs, since they are closely related to the productive and reproductive efficiency of the herds, presenting higher economic relevance for beef cattle production systems than growing traits (Brumatti et al., 2011).

In this scenario, using genomic prediction methods for the identification of genetically superior animals becomes a promising strategy, allowing to achieve higher genetic gains by the improvement of prediction accuracy and decrease on generation intervals (Hayes et al., 2009). Notably, the success of genomic-assisted selection depends directly on the prediction accuracy, which is associated, among other factors, with the used statistical method and the genetic architecture of the interest traits.

## 1.1 General objective

The aim of this study was to assess the feasibility of applying machine learning methods in the genomic analysis of reproductive traits in Nellore cattle, in order to provide insights for future utilization as alternative methodologies in beef cattle breeding programs.

## 1.2 Specific objectives

- To compare the predictive performance of GBLUP (Genomic Best Linear Unbiased Predictor) and machine learning methods in simulated beef cattle populations presenting different degrees of dominance effects.
- To investigate the predictive ability of different machine learning methods for genome-enabled prediction of reproductive traits in Nellore cattle and to compare with the performance of parametric approaches (GBLUP and Bayesian least absolute shrinkage and selection operator - BLASSO).
- To perform a genome-wide association study (GWAS) using the Random Forest approach for scanning novel candidate genes for age at first calving in Nellore cattle.

# 2 Literature review

## 2.1 Machine learning for genomic data analysis in animal breeding

Machine learning (ML) theory refers to a branch of artificial intelligence that combines statistics, computer science and data mining principles aiming to learn inherent patterns and to predict (or classify) interest outcomes in complex and massive databases. ML techniques have become popular in many different areas over recent years and have been applied for optimizing different tasks in the industry, such as personalization of browser commercials and improvement of translating services (Pérez-Enciso, 2017).

Due to the natural heterogeneity regarding biological data (*e.g*. transcriptional profiles, metabolic pathways, genomic and other *omics* disciplines) machine learning properties have been widely explored with different purposes in bioinformatics and other biological sciences. For instance, in medicine, recent advances in neuroimaging coupled with high-throughput genotyping have offered new approaches to study brain disorders (Wang et al., 2012).

On the other hand, animal breeding theory is ruled by the linear model paradigm. Hence, the breeding value is a dominant concept in this area, since additive genetic variability is the main responsible for the genetic response to selection. Naturally, predicting breeding values for quantitative traits is a central problem in animal breeding theory. Therefore, despite their conceptual differences, animal breeding and ML share common objectives, such as prediction of interest variables (González-Recio et al., 2014; Pérez-Enciso, 2017).

The seminal work of Meuwissen et al. (2001) has revolutionized the animal breeding field. Currently, the application of high dimensional markers data in breeding programs is a feasible approach and different methodologies have been proposed to cope with statistical and computational issues (De Los Campos et al., 2013). At the same time, opportunities offered by the growing ensemble of phenotypic and genomic data as well as unstructured data emerged by modern biological information poses new challenges that may not be well addressed by standard parametric methodology.

Machine learning offers a new approach for some gaps faced by standard methods adopted in quantitative genetics. The flexibility of ML methods may provide a potential benefit when aggregating all kind of prior biological insights gleaned from designed experiments, along with genomics, transcriptomics, and other '*omics*' disciplines, for predicting breeding values of selection candidates and help understanding complex traits biology.

In this section, some popular machine learning approaches, namely, artificial neural networks, support vector machines and random forest are detailed, the basic theoretical backgrounds of each method are presented. Further, previous applications of such approaches in animal breeding research are discussed.

### *2.1.1 Artificial neural networks (ANN)*

Artificial neural networks are mathematical information processing systems developed to mimic the biological nervous system. Analogous to the human brain, in the ANN, the input information (*e.g.* markers data) is processed by interconnected artificial neurons (linear or non-linear computing units), able to learn complex hidden relationships between predictor variables and the target in an adaptive way by using some adequate learning algorithm (Bishop, 2006). The general structure of an artificial neuron, the processing unit of the ANN model, can be represented schematically in the following diagram form (Figure 1).

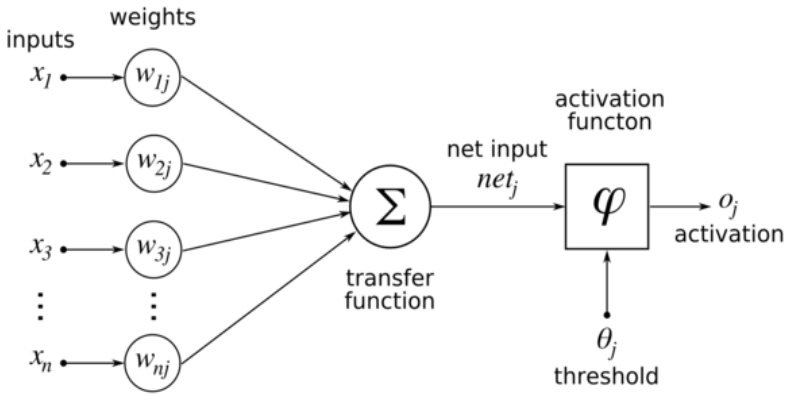


**Figure 1.** Diagram representation of an artificial neuron (Available at <https://m.tau.ac.il/~tsirel/dump/Static/knowino.org/wiki/Artificial_neural_network.html>)

As depicted in the above diagram, a typical artificial neuron, also termed perceptron, receives the input information $x_j$, which are individually weighted via synaptic weights $w_{km}$ (slope in the regression model) and, all obtained weights are summed up. After such a process, the resulting computation is then passed into a linear or non-linear activation function $\varphi(.)$ to produce the neuron output (Haykin, 2001).

Each artificial neuron is then interconnected to the others and organized in several layers forming an ANN (Hayes, 2001). Algebraically, an ANN can be viewed as a schematic of Kolmogorov's theorem for the representation of multivariate functions (Kurková, 1992; Pereira and Rao, 2009). These machine learning methods are an interesting alternative for genome-enabled prediction since they can act as universal approximators of complex functions (Gianola et al., 2011; Okut et al., 2013).

There are several architectures for an ANN regarding the number of hidden layers and the number of neurons in each of those layers and the type of activation function performed at each neuron. The two most used in genome-enabled prediction of complex traits are the so-called Feed-forward Multilayer Perceptrons (MLP) and Radial Basis Function Networks (RBFNN). This section will focus only on MLP features.

#### 2.1.1.1 Multilayer perceptron feed-forward neural networks (MLP)

The MLP is widely used for classification and regression tasks, including genome-enabled prediction (Gianola et al., 2011; Okut et al., 2013; Eheret et al., 2015). These models are termed feed-forward because the processed information always flows in one direction, *i.e.* the results from one layer form the input of the next layer. Because of the relative simplicity and ability to cover most of the problems, the single hidden layer feed-forward neural network is a frequently MLP architecture used for regression tasks (Hastie et. al. 2009). This model is formed by an input layer that receives data (*e.g.* markers matrix), one hidden layer containing the neurons linked to an output layer (Figure 2).

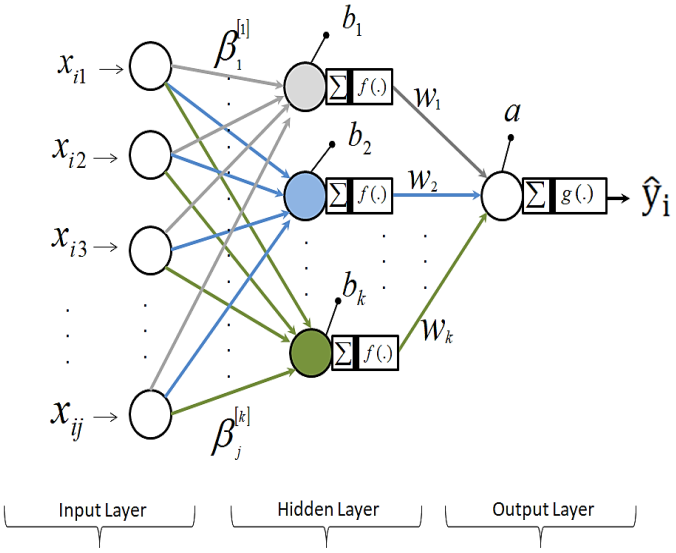


**Figure 2.** Schematic representation of the architecture of a single hidden layer neural network. $x_{ij}$ are the network inputs of the individual *i*; $\beta_j^{[k]}$ is a network weight for a given hidden layer, where *k* denotes the number of neurons in the hidden layer; $w_k$ is the network weight from the hidden to the output layer; $b_k$ and *a* are the biases in the hidden and output layers; *f(.)* and *g(.)* are activation functions in the hidden and output layers and $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ is the predicted value for the $i^{th}$ animal. Adapted from Eheret et al. (2015).

Algebraically, such ANN architecture can be viewed as a two-step regression (Hastie et al., 2009). In the first step the input variables $x_{ij}$ (*e.g.* SNP markers or any other covariate measured in the individuals, with *j* = 1,..,p ) of the animal *i* (for *i* = 1,..,n) are combined linearly with a vector of weights $\beta_j^{[k]}$ (with *k* = 1,..,*k*, regarding the k$^{th}$ hidden neuron in the net) plus an intercept (called bias in neural network terminology) $b_k$, the computations are transformed via a linear or non-linear activation function *f*(.), which can be neuron-specific or common to all neurons, resulting in the output of the $k^{th}$ neuron in the hidden layer for the $i^{th}$ individual (Eheret et al., 2015):

$$z_i^k = f\left(b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}\right) \qquad (2.1)$$

The resultant process produces *k* scores (where *k* is the number of neurons in the hidden layer) that are sent to the output layer and linearly combined to another vector of weights $w_k$ plus an intercept (*a*), and transformed by another activation function *g*(.), generally linear for quantitative responses, to calculate the predicted outcomes:

$$\hat{y} = g\{a + \sum_{k=1}^{s} w_k z_i^k\} \qquad (2.2)$$

The learning process of an MLP consists of finding the adequate set of weights and biases of the network which minimizes the differences between predicted and observed outputs, measured by an objective error function. The commonly adopted algorithm in such a process is the back-propagation of error (BP), viewed as a gradient descent method (Rojas, 1996). This is considered a supervised algorithm, in the sense, it needs several examples of the target variable in order to learn patterns.

In the BP network optimization, the typical function to be minimized is the sum of squared errors ($E_D$):

$$E_D\,(D|w, M) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2.3)$$

in which *D* represent the observed data used as input and target variables in the neural network, *w* represents the weights of the tuned *M* network architecture. Applying delta rule, the partial derivatives with respect to *w* are calculated in order to minimize such function (Rojas, 1996).

The back-propagation algorithm is flexible enough to cover high order interactions between predictor variables. However, as the number of predictor variables grows, such learning rule may suffer from overfitting, leading to predictions that are far beyond the range of the training data. Commonly, for genome-enabled prediction problems, the number of explanatory variables vastly exceeds the available observations ($p>>n$), in such a scenario, a neural network without practicing any regularization will probably collapse (Okut et al., 2013).

In general, there are two different ways to perform regularization in the neural network weights, the simplest is the early stop strategy. Briefly, such procedure consists of monitoring the weights and biases obtained with a gradient descent algorithm in the training data, substituting these parameter estimates in an independent validation set, for which target variables are also known. The iterative process stops when the observed error in the validation data set increases by certain epochs (Okut et al*.,* 2015). The network parameters which provide the smaller error in the validation set are returned as the most appropriate to predict not-yet observed data. Thus, early stopping imposes regularization by limiting the used weights in the network (Okut, 2016).

Another approach for restricting neural network connections strength is the Bayesian regularization. In Bayesian regularized ANN (BRANN) models, prior

distributions are imposed on the model parameters in order to penalize large weights aiming to achieve better generalization and smother mapping (Gianola et al., 2011). Thus, penalization parameters in BRANN regulate the trade-off between goodness-of-fit and model complexity of the network automatically. Similar to the traditional backpropagation, a gradient-based algorithm is used to optimize the earlier mentioned cost function which receives an additional term and can be considered as a penalized log-likelihood (Okut et al., 2013):

$$F = \beta E_D(D|w, M) + \alpha E_w(w|M) \tag{2.4}$$

in which $E_w(w|M) = \sum_{i=1}^{m} {w_i}^2$ is the sum of the square of the network weights, *m* is the total of weights, *D* stands for available training data, *M* represents a specific network architecture, $\alpha$ and $\beta$ are positive regularization parameters, that must be estimated on the training phase. The second term $\alpha E_w$ on the right-hand side of the equation is called weight decay, which penalizes larger weights and decreases the tendency of overfitting. The weigh decay coefficient $\alpha$ yields a nonlinear version of ridge regression. From a Bayesian perspective, the posterior distribution of *w* given $\alpha$, $\beta$, $D$, and $M$, assuming a Gaussian distribution for the noise on training data is (Gianola et al., 2011):

$$P(w|\alpha, \beta, D, M) = \frac{P(D|w, \beta, M)P(w|\alpha, M)}{P(D|\alpha, \beta, M)} \tag{2.5}$$

where $P(w|\alpha, M) = \left(\frac{\alpha}{2\pi}\right)^{\frac{m}{2}} \exp\left(-\frac{\alpha}{2}\mathrm{w'w}\right)$ is the prior distribution assigned for the neural network weights, $P(D|w, \beta, M)$ is the likelihood function of **w** and $P(D|\alpha, \beta, M)$ represents the marginal likelihood, which is a normalization factor and does not depend on **w**. The optimal solution for **w** is given by maximizing its posterior density, which is equivalent to minimizing the objective function $F = \beta E_D + \alpha E_w$ (Foresee and Hagan, 1997).

The tuning process of the parameters $\alpha$ and $\beta$ also receives a Bayesian treatment, with the following joint posterior density (Foresee and Hagan, 1997):

$$P(\alpha, \beta, |D, M) = \frac{P(D|\alpha, \beta, M)P(\alpha, \beta|M)}{P(D|M)} \tag{2.6}$$

Assuming uniform prior density for $P(\alpha, \beta|M)$ and then maximizing $P(\alpha, \beta, |D, M)$ is equivalent to maximization of the likelihood function $P(D|\alpha, \beta, M)$ which corresponds to the normalization factor in the $P(w|\alpha, \beta, D, M)$. According to Mackay (1992), such density has the following form:

$$P(D|\alpha,\beta,M) = \frac{P(D|w,\beta,M)P(w|\alpha,M)}{P(w|D,\alpha,\beta,M)} = \frac{Z_F(\alpha,\beta)}{(\pi/\beta)^{n/2}(\pi/\alpha)^{m/2}} \quad (2.7)$$

where n and m are the numbers of observations and parameters, respectively. The term $Z_F(\alpha,\beta)$ can be found applying Laplace approximation:

$$Z_F(\alpha,\beta) \propto \left|\mathbf{H^{MAP}}\right|^{-\frac{1}{2}} \exp\left(-\mathrm{F}(\mathrm{w}^{\mathrm{MAP}})\right) \quad (2.8)$$

where $\mathrm{H}^{\mathrm{MAP}}$ is the Hessian matrix of the objective function and MAP stands for *maximum a posteriori* estimates. As shown by Mackay (1992), the optimal values of $\alpha$ and $\beta$ at $\mathrm{w}^{\mathrm{MAP}}$ can be obtained as:

$$\alpha^{\mathrm{MAP}} = \frac{\gamma}{2E_w(\mathrm{w}^{\mathrm{MAP}})} \text{ and } \beta^{\mathrm{MAP}} = \frac{n-\gamma}{2E_D(\mathrm{w}^{\mathrm{MAP}})} \quad (2.9)$$

in which $\gamma = m - 2\alpha^{\mathrm{MAP}} tr(\mathbf{H^{MAP}})^{-1}$ is called the effective number of parameters in the neural network.

2.1.1.2 Practical considerations on the ANN's implementation

There are some aspects that must be considered before and during the training of a neural network model which will influence its predictive ability. Pre-processing of data, checking for inconsistency, is a basic step to avoid introducing incorrect and spurious data. In some cases, the reduction of dimensional space of features is also advisable. The principal component analysis is a common technique employed for such a task (Okut, 2016).

The number of hidden layers and the number of neurons at each layer must be determined in order to capture relevant signals from data and simultaneous preventing the overfitting. Therefore, the optimal number of neurons is a model selection issue. In general, a model with only a few neurons in the hidden layer may not address well the associations between the input and target variables. On the other hand, if too many neurons are assigned for a hidden layer, the network will tend to learn spurious signals, leading to the poor predictive ability of not observed data (Okut et al., 2013).

Usually, one must determine the optimal number of neurons empirically, by training several models and observing the set of parameters which provides the better predictive ability. In general, because its simplicity, enough flexibility, and faster training, most authors have considered for genome-enabled prediction, a single hidden

layer model (as depicted in Figure 2) with up to 10 hidden neurons (Gianola et al., 2011; Okut et al., 2013; Howard et al., 2014).

The activation function is another component of the neural network architecture, influencing the capability of mapping the problem. Generally, the nature of data or assumptions on the target variable distribution will guide the choice of the most suited activation function. Some commonly used activation functions on MLP models are the identity or linear, step or threshold, sigmoid, and hyperbolic tangent. A non-linear activation function provides flexibility for the neural network to approximate any complex function (Alados et al., 2004; Bishop, 2006).

Finally, normalization or standardization of the input and target variables on the same scale which the activation function lies (e.g. ranging from -1 to 1 for hyperbolic tangent or from 0 to 1 for sigmoid function) improves significantly convergence of learning algorithm by boosting its numerical stability (Okut, 2016). After the learning algorithm converges, the neural network outputs can be easily transformed into the original scale.

### *2.1.2 Support vector machines (SVM)*

Support vector machines (SVM) theory was first developed by Vapnik (1995) as a supervised learning algorithm for binary data classification. Due to their robustness and good generalization ability, the SVM approach gained popularity in machine learning research. Since their introduction, SVM theory has been extended and extensively applied in different tasks, such as classification, regression, novelty detection and feature reduction (Awad and Khanna, 2015).

#### 2.1.2.1 Margin maximization concept applied for binary classification

In order to introduce support vector machines as non-linear classifiers, one must understand the margin maximization concept. Suppose we have the training dataset $D$, comprising $p$ input vectors ($x_1$, $x_2$, $x_3$, ..., $x_p$, with $x_i \in \mathbb{R}^p$) with corresponding $N$ observed values assuming two possible classes, $y_i = +1$ or $y_i = -1$ for $i = 1,2,3, ..., N$. For sake of simplicity let us assume that the training data set is linearly separable, this

implies that one can define a linear hyperplane separating the two classes on the feature space. Considering a real-valued *p*-dimensional feature space, such hyperplane has *p-1* dimension embedded within the feature space $\mathbb{R}^p$. When *p*=2 this hyperplane is simply a one-dimensional straight line, which lies in the larger two-dimensional plane (Figure 3).

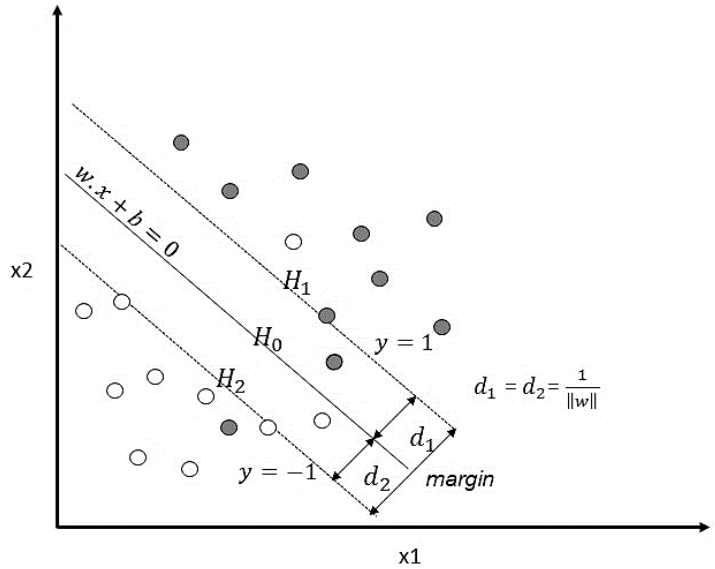


**Figure 3.** Example of an optimal hyperplane for linearly separable binary classes. White and gray circles represent datapoints for class -1 and 1, respectively; $d_1$ and $d_2$ are the distances of closest data points from the hyperplane. Adapted from Awad and Khanna (2015).

There are several possible hyperplanes that separate the data points, support vector machines aim to find the optimal separating hyperplane, which maximizes the margin of the training data. Intuitively, the margin is twice the distance between the hyperplane and the closest data points, called support vectors (SV) (Bishop, 2006). The margin defines how well two classes can be separated. A hyperplane $H_0$ separating the data can be described as:

$$w.x + b = 0 \tag{2.10}$$

where **w** is a vector of parameters of the hyperplane and **b** is a constant which moves out the hyperplane from the origin. As depicted in Figure 3, the selected hyperplane must meet the following constraints:

$$\begin{cases} x_i \,.\, w + b \geq +1 & \text{for } y_i = +1 \\ x_i \,.\, w + b \leq -1 & \text{for } y_i = -1 \end{cases} \tag{2.11}$$

If we multiply both sides of these equations by $y_i$, they can be combined into:

$$y_i\ (x_i\ .w + b)\ \geq\ 1 \quad \forall_i \tag{2.12}$$

Considering the closest points to the separating hyperplane (*i.e*. the support vectors), one can describe the two planes $H_1$ and $H_2$ in which these points lie as:

$$\begin{cases} x_i\ .w + b = +1 & \text{for } H_1 \\ x_i\ .w + b = -1 & \text{for } H_2 \end{cases} \tag{2.13}$$

Referring to Figure 3, we can define $d_1$ or $d_2$ as the perpendicular distance of any point lying in $H_1$ or $H_2$ from $H_0$, respectively. From vector geometry, one can determine the distance of a given point $x_i$ and $H_0$ as:

$$d_i\ (\mathbf{w}, b; x_i) = \frac{|(x_i\ .w+b)|}{\|\mathbf{w}\|} \tag{2.14}$$

Assuming $d_1$ and $d_2$ equidistant to $H_0$, the margin $M$ is given by (Gunn, 1998):

$$M = \min_{x_i:y_{i=-1}} \frac{|(x_i\ .w+b)|}{\|\mathbf{w}\|} + \min_{x_i:y_{i=1}} \frac{|(x_i\ .w+b)|}{\|\mathbf{w}\|} \tag{2.15}$$

$$M = \frac{1}{\|\mathbf{w}\|}\left(\min_{x_i:y_{i=-1}} |(x_i\ .w + b)| + \min_{x_i:y_{i=1}} |(x_i\ .w + b)|\right) \tag{2.16}$$

$$M = \frac{2}{\|\mathbf{w}\|} \tag{2.17}$$

Therefore, maximizing $M$, constraint to 2.12 is equivalent to finding:

$$\arg\min\|\mathbf{w}\| \qquad \text{subject to } y_i\ (x_i\ .w + b) - 1 \geq\ 0 \quad \forall_i \tag{2.18}$$

Such an optimization problem can be rewritten as (Bishop, 2006):

$$\arg\min\ \tfrac{1}{2}\|\mathbf{w}\|^2 \qquad \text{subject to } y_i\ (x_i\ .w + b) - 1 \geq\ 0 \quad \forall_i \tag{2.19}$$

The adopted form in 2.19 allows performing Quadratic Programming (QP) optimization (Hearst et al., 1998), useful for minimizing a quadratic function subject to a set of linear inequality constraints. In order to address the constraints of such a convex optimization problem, one can introduce Lagrange multipliers $\alpha_i \geq 0\ \forall_i$, resulting in the following function (Campbell, 2000):

$$L(w, b, \alpha) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i[\ y_i\ (x_i\ .w + b) - 1]$$

$$L(w, b, \alpha) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \textstyle\sum_{i=1}^{N} \alpha_i[\ y_i\ (x_i\ .w + b) + \sum_{i=1}^{N} \alpha_i] \tag{2.20}$$

The optimal solution is given by minimizing 2.20 with respect to $w$, $b$ and maximizing with respect to $\alpha_i$. This can be achieved by differentiating $L(w, b, \alpha)$ with respect to $w$ and $b$ and setting the derivatives equal to zero (Bishop, 2006):

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0\ \Rightarrow w = \textstyle\sum_{i=1}^{N} \alpha_i\ y_i x_i \tag{2.21}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (2.22)$$

Replacing 2.21 and 2.22 into 2.20 gives the *dual representation* of the maximum margin problem (Bishop, 2006):

$$\max \tilde{L}(\alpha) = \max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \quad (2.23)$$

subject to $\alpha_i \geq 0 \ \forall_i$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

After to compute the optimal value $\alpha^*$ of the dual problem, one can find $w^*$ by replacing $\alpha^*$ in 2.21:

$$w^* = \sum_{i=1}^{N} a_i^* \, y_i x_i \quad (2.24)$$

in turn, the parameter $b^*$ can be determined from a* and from the Karush-Kühn-Tucker (KKT) conditions, which generalizes the method of Lagrange multipliers, allowing inequality constraints (Awad and Khanna, 2015). Following the duality problem aforementioned, the KKT conditions provide (Awad and Khanna, 2015):

$$a_i^* [\, y_i \, (x_i \, . w^* + b^*) - 1] = 0, \forall_i = 1, \ldots, n \quad (2.25)$$

In 2.25, the term $a_i^*$ only assumes positive values for the training data points lying on the separating hyperplanes $H_1$ and $H_2$. Therefore, the support vectors are the data-points which satisfies $a_i^* > 0$, whereas for the remaining data $a_i^* = 0$ (Lorena and Carvalho, 2007). Thus, the parameter $b^*$ can be found by averaging over all SVs as follows:

$$b^* = \frac{1}{N_{SV}} \sum_{x_i \in S} (y_i - x_i . w^* ) \quad (2.26)$$

in which $N_{SV}$ represents the number of support vectors and $S$ the set of found SVs.

After computing $w^*$ and $b^*$, one can classify an unobserved data $x$ by solving:

$$sgn \, (\langle x . w^* \rangle + b^*) \quad (2.27)$$

in which $sgn\,(.)$ is the sign function, labeling the possible results as positive or negative. The classification is given by the dot product between the new predictive variable and the founded SVs. Therefore, all the relevant information contained in the training set can be summarized in terms of support vectors, hence, the name Support Vector Machine (Pontil and Verri, 1998).

Previously, as showed in Figure 3, it was presumed that all data points are completely separable, a problem known as hard-margin optimization. However, in practice, it may occur some data overlapping the hyperplane, in such cases, slack variables $\xi_i$ are introduced to the SVM cost function in order to relax the constraints in

*2.11*, allowing to maximizing the margin while softly penalizing points on the wrong side of the decision boundary (Hastie et al., 2009). The optimization problem can be rewritten as follows (Bishop, 2006; Awad and Khanna, 2015):

$$\arg\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i \quad (2.28)$$

$$\text{subject to } y_i\,(x_i\,.w + b) - 1 + \xi_i \geq 0 \quad \forall_i \quad (2.29)$$

where the constant $C > 0$ controls the trade-off between the slack variable penalty and the margin (Bishop, 2006). In other words, as *C* parameter increases, smaller is the margin of the hyperplane, in order to minimize the number of misclassified points, thus, reducing the training error, but increasing the model complexity. Conversely, as *C* value decreases, smoother the decision boundary, allowing a larger margin, although misclassifying more training points.

Reformulating the problem in *2.28*, subject to *2.29* gives the corresponding Lagrangian function (Bishop, 2006):

$$L(w, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}\alpha_i[\,y_i\,(x_i\,.w + b) - 1 + \xi_i] - \sum_{i=1}^{L}u_i\xi_i \quad (2.30)$$

Now, differentiating *2.30* with respect to w, b, and ξ and setting the derivatives to zero:

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N}\alpha_i\,y_i x_i \quad (2.31)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i y_i = 0 \quad (2.32)$$

$$\frac{\partial L}{\partial \xi} = 0 \Rightarrow \mathrm{C} = \alpha_i + u_i \quad (2.33)$$

Replacing these results in the Lagrangian function gives the *dual formulation* of the soft margin problem (Awad and Khanna, 2015):

$$\max \tilde{L}(\alpha) = \max \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j \quad (2.34)$$

which has the same formulation as in 2.23, except for the following constraints:

$$\sum_{i=1}^{N}\alpha_i y_i = 0.$$

$0 \leq \alpha_i \leq C$, $i = 1, 2, \ldots, N$.

2.1.2.2 The kernel trick

In order to account for non-linearities between input and response variables, the SVM algorithm may be benefited with the use of Kernel concept for mapping the input

variables into a higher dimensional Hilbert space $\mathcal{H}$ (referred as *kernel space or feature space*), where data are linearly separable (Awad and Khanna, 2015). Finding an optimal separating hyperplane in the feature space is equivalent to form a non-linear boundary in the original variable space that better fits the data distribution (Brereton and Lloyd, 2009). The mapping via kernels relies on the notion of similarity between data points, measured on the basis of the dot product. A kernel is a positive semidefinite matrix satisfying (Salcedo-Sanz et al., 2014):

$$\mathrm{K}(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \tag{2.35}$$

where $\mathrm{K}(x_i, x_j)$ is a given kernel matrix and the mapping function $\phi(x)$ belongs to the Hilbert space. Operating on the data only in terms of the inner products of mapped inputs allows representing the feature space without to explicitly compute $\phi(x)$, a technique is also known as the *kernel trick.* Some popular kernel functions include (Cristianini and Shawe-Taylor, 2000; Awad and Khanna, 2015):

a) Linear Kernel: $\mathrm{K}(x_i, x_j) = x_i^T x_j$

b) Polynomial: $\mathrm{K}(x_i, x_j) = \left(a x_i^T x_j + c\right)^q, q > 0$

c) Gaussian Radial Basis Function (RBF): $\mathrm{K}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

d) Hyperbolic Tangent (sigmoid): $\mathrm{K}(x_i, x_j) = tahn\left(\beta x_i^T x_j + \gamma\right)$

The *dual formulation* in 2.34 can be rewritten by replacing the scalar product of input vectors $x_i x_j$ with the kernel function as follows:

$$\max \tilde{L}(\alpha) = \max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i x_j) \tag{2.36}$$

subject to:

$$\sum_{i=1}^{N} \alpha_i y_i = 0.$$

$$0 \leq \alpha_i \leq C,\ i = 1,2,\ldots,N.$$

The choice of a suitable kernel function and its parameters play a crucial role in the SVM performance and it will depend on the data and problem at hand; however, a well-suited kernel function in the absence of prior knowledge and with good generalization capability is the RBF function (Awad and Khanna, 2015).

2.1.2.3 Support vector regression (SVR)

Analogously to the previous sections, SVM theory can be expanded for dealing with regression problems (Vapnik, Golowich and Smola, 1997). As in SVM, support vector regression (SVR) approach applies the basic principles of sparseness and structural risk minimization to find an optimal solution to the problem. The basic idea is to find a given function *f(x)* such that it minimizes some loss function. However, in SVR, the ε-insensitive loss function is adopted, which ignores predicted values that have a distance from the observed training data $y_i$ less than a given constant ε. For the linear case, the problem can be described graphically as in Figure 4, where data points lying inside the ε-tube are expected have no impact on the final solution and hence are not considered in the loss function (Alonso et al., 2013). Moreover, analogously to the soft margin in SVM for classification, one can introduce slack variables $\xi, \xi^*$ in the loss function, in order to penalize data points lying outside the ε-insensitive tube (Figure 4).

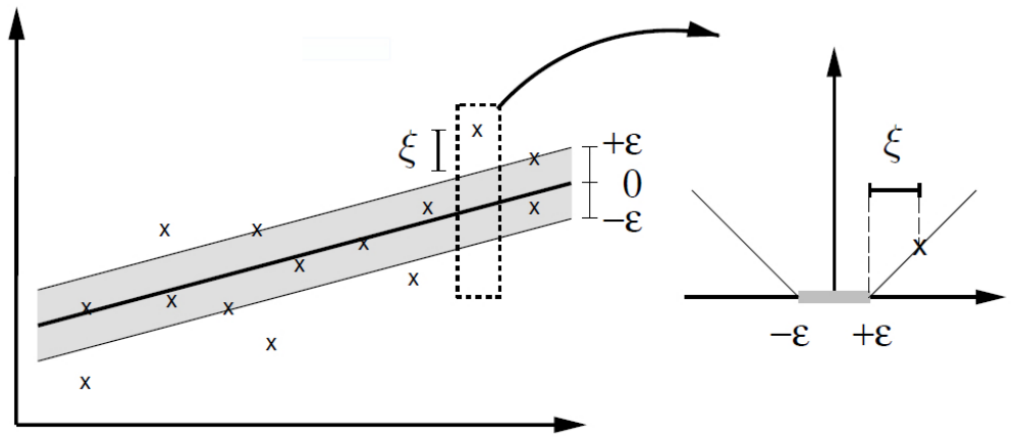


**Figure 4.** Geometrical interpretation of the ε-tube (left) in the linear support vector regression (SVR) model. All samples with absolute errors larger than the constant ε are penalized using the Vapnik's ε-insensitive loss function (right) (Alonso et al., 2013).

Taking into account the aforementioned, a generalized model for quantitative responses can be presented as:

$$f(x) = \langle w, \phi(x_i) \rangle + b \tag{2.37}$$

in which *w* is the weight vector of inputs, $\phi(x_i)$ represents the mapping via a kernel function into some feature space and *b* is the bias. The optimization problem consists

in to find the narrowest tube while minimizing the prediction error, this can be written as (Awad and Khanna, 2015):

$$\arg\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2.38}$$

$$\text{subject to:} \begin{cases} w^T\phi(x_i) + b - y_i \leq \varepsilon + \xi_i \ \forall_i \\ y_i - w^T\phi(x_i) + b \leq \varepsilon + \xi_i^* \ \forall_i \\ \xi, \xi^* \geq 0 \end{cases}$$

Similar to the soft-margin SVM, the constant C > 0 regulates the trade-off between minimizing model complexity and training errors. The larger the value of C, the smaller the tolerance with deviations larger than ε, minimizing the training error at the cost of augmenting the model complexity. As represented graphically in Figure 4, this corresponds to dealing with the so-called ε-insensitive loss function $|\xi|_\varepsilon$, described as (Smola and Schölkopf, 2004):

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| < \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{2.39}$$

The Lagrangian function then becomes:

$$L(w, b, \alpha, \ \alpha^* \xi, \ \xi^*, \lambda, \ \lambda^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) + \sum_{i=1}^{N}\alpha_i^*[y_i - w^T\phi(x_i) - b - \varepsilon - \xi_i^*] + \sum_{i=1}^{N}\alpha_i[w^T\phi(x_i) + b - y_i - \varepsilon - \xi_i] - \sum_{i=1}^{N}\lambda_i\xi_i + \lambda_i^*\xi_i^* \tag{2.40}$$

Optimal solutions are obtained minimizing the equation in 2.40 by taking its partial derivatives with respect to the primal variables ($w, b, \ \xi, \ \xi^*$) and setting them equal to zero, based on the KKT conditions it produces (Awad and Khanna, 2015):

$$\frac{\partial L}{\partial w} = 0 \ \Rightarrow w = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\,\phi(x_i) \tag{2.41}$$

$$\frac{\partial L}{\partial b} = 0 \ \Rightarrow \mathrm{b} = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) \tag{2.42}$$

$$\frac{\partial L}{\partial \xi} = 0 \ \Rightarrow \ \mathrm{C} = \ \alpha_i + \lambda_i \tag{2.43}$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \ \Rightarrow \ \mathrm{C} = \ \alpha_i^* + \lambda_i^* \tag{2.44}$$

Replacing these terms into equation 2.40, the dual form of the optimization problem can be written as:

$$\max -\frac{1}{2}\sum_{j=1}^{N_{SV}}\sum_{i=1}^{N_{SV}}(\alpha_i^* - \alpha_i)\left(\alpha_j^* - \alpha_j\right)\phi(x_i)^T\phi\left(x_j\right) - \ \varepsilon\sum_{i=1}^{N_{SV}}(\alpha_i^* - \alpha_i) + \ \sum_{i=1}^{N_{SV}}(\alpha_i^* - \alpha_i)\,y_i \tag{2.45}$$

subject to:

$$\sum_{i=1}^{N_{SV}}(\alpha_i^* - \alpha_i) \text{ and } \alpha_i, \alpha_i^* \epsilon\ [0, \mathrm{C}]$$

In equation 2.45, the Lagrange multipliers that are equal to zero correspond to the data inside the ε-tube, whereas the support vectors have non-zero values for the Lagrange multipliers. The final solutions depend only on the support vectors; hence the sparsity of the SVR solution.

### *2.1.3 Random forest (RF)*

Random Forest, early proposed by Breiman (2001), is one of the most popular ensembles learning algorithms for classification, regression, variable selection and other applications in data mining and machine learning (Genuer et al., 2010; Chen and Ishwaran, 2012). The basic concept of an RF model is to construct many decision trees on bootstrap samples from the original data set. The resulting estimate for each tree is averaged to obtain the final prediction. Such an approach termed bagging (Breiman, 1996), is useful to impose regularization, reducing the error prediction by a factor of the number of trees as dealing with correlation and interaction among variables (González-Recio and Forni, 2011; Chen and Ishwaran, 2012). Hence, RF offers a non-parametric approach, robust to overfitting and able to capture non-additive effects, providing useful features for analyzing genome-wide data.

#### 2.1.3.1 The RF algorithm

The basic unit of the RF model is a single decision tree, which can be applied to both classification and regression problems. Building a decision tree involves partitioning the original predictor space into a number of simplest regions. Typically, in the RF, the trees are grown using CART (Classification and Regression tree) methodology (Breiman et al., 1984), in which the input space is recursively partitioned using binary splits at each *internal node* of the tree until to achieve homogeneous or near homogeneous responses into the *terminal nodes* or *leaves* of the tree (Figure 5).

In the recursive binary splitting, a predictor variable $X_j$ and a cutpoint $t_k$ are selected, such that the resulted predictor space falls into subregions $\{X|X_j \leq t_k\}$ and $\{X|X_j > t_k\}$, which minimizes a given split function. Generally, the CART split criterion

is based on the Gini impurity (for classification) or the prediction squared error (for regression) (James et al., 2013).

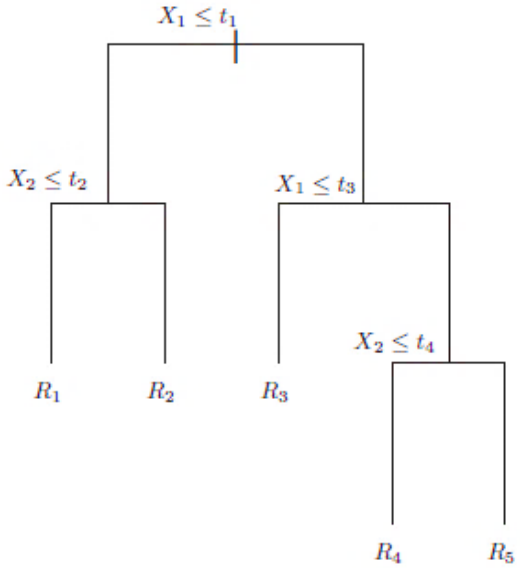


**Figure 5.** Example of a tree building from recursive binary splitting on a two-dimensional feature space. $X_1$ and $X_2$ are predictor variables (*e.g.* SNPs); $t_i$ stands for a binary splitting rule; $R_j$ (j = 1, 2, …, 5) represent the response regions on the terminal nodes (James et al., 2013).

For continuous responses, predictions of a tree-based model are generally performed using the mean of the training observations pertaining to each terminal node. Analogously, for discrete responses, classifications are performed as the most commonly occurring class of training observation in a particular terminal node (James et al., 2013).

However, unlike CART, in the Random Forest model, the trees are not grown using the entire predictive space, instead, a random sample of *m* predictors from the full set of *p* explanatory variables is chosen as split candidates for each individual tree. Besides, as mentioned before, prior to each tree construction a bootstrap sample is drawn from the original data. Therefore, in the RF, the building process of each tree involves a two-step randomization procedure, which decorrelates the trees, so that the resulting ensemble of trees is expected to have less variance (Chen and Ishwaran, 2012; Biau and Scornet, 2016).

The RF algorithm can be summarized in the following steps (González-Recio and Forni, 2011; Chen and Ishwaran, 2012):

1. Prior to building a tree, drawn a bootstrap sample with replacement from the whole dataset. Use the bootstrap dataset to build a classification or regression tree.
2. At each node of a given tree, draw randomly *mtry* (a constant defined by the user) variables from the whole predictor space. From the *mtry* predictors evaluate the possible variables and cut points and select those which minimize a given loss function (e.g. logit function, squared loss function, misclassification rate, Gini index, entropy).
3. Split the node into two child nodes and pass the observations according to the coordinates.
4. Repeat steps 2 - 3 until the terminal nodes of the tree have no more observations than the *nodesize* parameter. The predicted value is obtained by averaging the responses at the terminal nodes (for qualitative response is the majority vote for the outcome at the terminal nodes).
5. Repeat steps 1 - 4 *ntree* times with different bootstrap samples and *mtry* predictor variables sets for each tree node. A given observation may appear several times (or not) in each tree.
6. Aggregate the information from the ensemble of trees to compute final predictions by averaging all *ntree* predictors $\{f(X,\psi_b)\}_1^{Ntree}$ as:

$$\hat{y} = \frac{1}{Ntree}\sum_{b=1}^{Ntree} T(X,\psi_b) \tag{2.46}$$

where $\psi_b$ represents an individual *b* tree architecture in terms of the bootstrapped sample, split variables, cut point at each node and terminal node values.

For unobserved values, the prediction is obtained by passing down the predictor variables in the flowchart of each tree and the corresponding estimate at the terminal node is assigned as the predicted value. Predictions of each tree in the RF are averaged to compute the final prediction for unobserved data.

Another interesting feature of RF is the out-of-bag data (OOB), since each tree are grown with a bootstrap sample of the original data, neither all observations are used to fit each tree. The remaining samples that are not selected to fit the trees (roughly one-third of the total observations) are termed OOB samples. The OOB can be used as an internal validation set without needing to perform cross-validation, for

which the generalization error for an RF model can be computed (James et al., 2013). A common error measure adopted in regression problems is the mean squared error:

$$MSE_{OOB} = \frac{1}{N_{OOB}} \sum_{i=1}^{N_{OOB}} (y_i - \hat{y}_i)^2 \quad (2.47)$$

in which $N_{OOB}$ is the number of observations on the OOB samples, $\hat{y}_i$is the averaged prediction of the whole regression trees in the random forest and $y_i$ is the realized value of the *i*$^{th}$ OOB sample.

2.1.3.2 Random forest tuning parameters

There are some tuning parameters of the RF model that may influence its predictive ability, such as the number of trees to grow (*ntree*), the number of possible predictor variables randomly chosen for splitting at each node of each tree (*mtry*) and the number of observations at the terminal nodes (*nodesize*).

It can be noted that higher values for the *ntree* parameter will lead to an ensemble of trees with smaller variance. According to Biau and Scornet (2016), a large *ntree* does not lead to overfitting, albeit computational cost increases linearly as *ntree* increases. Therefore, these authors argue that the choice of *ntree* value should be a trade-off between the computational complexity and model accuracy. In practice, it is advisable to try increasing values of *ntree* until measures of interest (*e.g.* OOB prediction error) stabilize (Boulesteix et al., 2012).

Generally, the *mtry* parameter has the greatest impact in the OOB error, this parameter is related to the bias-variance trade-off of the RF model. Searching over fewer variables per node contributes to reducing the correlation between trees, and hence, the overall variance of prediction decreases. Conversely, this will decrease the accuracy of the individual trees, which leads to an increase in bias (Goldstein et al., 2010). As recommended by Breiman (2001), the default values for this parameter are $\sqrt{p}$ (for classification) and $p/3$ (for regression), where *p* represents the number of predictor variables (*e.g.* the total number of SNPs). Although these values seem to work well for several datasets, they might too small in the presence of a large number of noise predictors (Boulesteix et al., 2012). However, such behavior may be dependent on the problem at hand. For instance, using microarray data to classify cancer patients according to their genetic profiles, Díaz-Uriarte and Andrés (2006)

reported that the default setting of *mtry* was appropriate, according to the OOB error rate, although, in some cases, increasing *mtry* value has provided better performance.

In genome-wide data, the genetic background of the trait may influence the optimal choice for this parameter. For a trait with few major genes influencing the phenotype, small values for the *mtry* reduces the chances of selecting relevant variables as splitting candidates. On the hand, in the case of a trait with many informative variables (*e.g.* SNPs) with different signals, would be reasonable to choose a smaller *mtry* value, in order to allow the algorithm to test predictor variables with moderate effect, that might be masked by predictors with strong signals. Some authors have suggested a *mtry* value of *0.1p* for genome-wide association studies (Goldstein et al., 2010; Li et al., 2018).

Regarding the *nodesize* parameter, Boulesteix et al. (2012) argue that small values are necessary to avoid bias in the trees building. Generally, the values of 1 and 5, for classification and regression problems, respectively, are considered as default, although there is no solid theory supporting this choice (Biau and Scornet, 2016).

#### 2.1.3.3 Variable importance measures (VIM)

In the random forest algorithm, importance measures for each predictor variable can be internally computed. In this regard, there are two main variable importance measures (VIM): The Gini importance and permutation-based variable importance (Boulesteix et al., 2012). In classification models, the Gini VIM is directly derived from the splitting function called Gini index, which is based on the level of node impurity to determine the selected variable for splitting in descendent nodes during the trees building process (Breiman, 2001). The Gini VIM for a given variable can be computed as the sum of Gini index reduction (from parent to children node) for all nodes in which the interest predictor was used as a splitting node, scaled by the number of trees in the forest (Boulesteix et al., 2012; Chen and Ishwaran, 2012). The more informative is the predictor, the higher are the chances of it being selected as a splitting node, leading to a high Gini VIM value. Nevertheless, if predictor variables are categorical, Gini VIM tends to present bias in favor of those with more categories (Strobl et al., 2007).

The permutation-based VIM is computed by accounting how much prediction error increases when a given predictor variable is randomly permutated in the out-of-bag (OOB) data, whereas all other variables remain unchanged (Breiman, 2001). The difference between the OOB error of the data with random permutation and the OOB error without permuting the variable of interest averaged over all trees in the forest is considered the VIM of the variable. Variables with strong associations with the response variables are expected to have higher permutation-based VIM since permutating such variables would lead to an increase in the OOB error (Boulesteix et al., 2012; Chen and Ishwaran, 2012).

Accounting for VIM provides an approach to rank variables, which is useful for identifying a subset of relevant predictors. The RF importance scores may reflect both direct and interaction effects (Yao et al., 2013). This is an appealing feature for high-dimensional genomic data and has been considered as an alternative methodology for genome-wide association studies of human diseases (Goldstein et al., 2010) and economically important traits in livestock (González-Recio and Forni, 2011; Mokry et al., 2013; Yao et al., 2013).

### *2.1.4 Machine learning for genome-enabled prediction and classification of complex traits*

Genomic selection has gained much attention over recent years as a feasible strategy for accelerating the genetic improvement in livestock (Hayes et al., 2009; Meuwissen et al., 2013). For this purpose, several genome-enabled prediction methods in animal breeding commonly make use of shrinkage or regularization processes for imposing prior assumptions regarding the genetic architecture of complex traits (De Los Campos et al., 2013). These models are linear in their essence and typically cope only with the additive effects of genetic variants. Covering non-additive effects (*e.g.* epistatic effects) under a parametric paradigm is possible by taking appropriate contrasts in the regression model, at the cost of increased computational burden (Su et al., 2012). Alternatively, different machine learning models have been proposed as more flexible approaches to explore possible non-

linear effects for enhancing genome-wide prediction of complex traits (Gianola et al., 2006; Gianola et al., 2011; Long et al., 2010).

Among the earliest efforts for applying machine learning methods in the genome-enabled prediction of complex traits, Gianola et al. (2006) introduced a reproducing kernel Hilbert spaces model (RKHS) as a semi-parametric alternative to infer genetic values of quantitative traits. Posteriorly, support vector machine (SVM; Moser et al., 2009), artificial neural networks (ANN; Gianola et al., 2011) and random forest (RF; González-Recio and Forni, 2011) were also investigated. In general, the empirical evidence provided so far indicates that machine learning methods have the potential to achieve similar or superior results than the traditional linear models.

Gianola et al. (2011), argued that ANN is an interesting alternative to adaptatively accommodate high order non-linear interactions in complex traits. These authors investigated Bayesian regularized artificial neural network models (BRANN) with several architectures to predict fat, milk and protein yield of Jersey cows, as well as grain yield in wheat. The models were derived using either pedigree or genome-derived relationships matrices as the network input information, the prediction accuracy was assessed according to the number of neurons (1 up to 6) in the hidden layer and use of different activation functions. The authors reported that the ANN models with non-linear architectures tended to outperform the linear model with only one neuron in the hidden layer and a linear activation function (equivalent to a Bayesian Ridge Regression model), supporting the choice of a non-linear neural network with at least 2 neurons.

In Angus cattle, the BRANN model performed similarly to a Bayesian linear regression model to predict expected progeny difference for marbling score, for which only additive effects are expected to affect the outcome of interest. The average correlations between predicted and observed values in the testing set were 0.776 with BayesCpC and ranged from 0.776 to 0.807 with the BRANN (Okut et al., 2013).

More recently, Ehret et al. (2015) applied ANN with back-propagation for predicting daughter yield deviations (DYD) and yield deviations (YD) for fat, milk and protein yield of Holstein-Friesian and Fleckvieh cattle. Those authors compared the influence of different genomic structures as input information in the neural networks, the entire marker matrix (X), a genome-based relationship matrix (G) and principal

components matrix (UD). Results pointed out that dimension reduction methods enhanced the prediction performance for all studied traits while also decreasing the computational cost. However, the non-linear ANN did not outperform the prediction ability of the linear model, with average prediction accuracies in the best non-linear architecture varying between 0.35 and 0.68, according to the used data set.

As observed for other genome-enabled prediction models, the predictive ability of machine learning methods may depend on several factors such as the size of the training set, heritability magnitude, extent of linkage disequilibrium (LD) in the population and the underlying genetic architecture of the trait (González-Recio and Forni, 2011; Howard et al., 2014; Ghafouri-Kesbi et al., 2016; Naderi et al., 2016; Sadeghi et al., 2018). Different simulation studies support that machine learning enables more accurate predictions when the interest traits are affected by non-additive effects, especially when the underlying genetic architecture is mainly due to epistasis (Long et al, 2011a; Howard et al., 2014).

In practice, animal breeders do not know exactly the underlying genetic architecture of quantitative traits; therefore, the empirical results may vary remarkably. It can be observed in Tussel et al. (2013), where the predictive ability of the BRANN model varied drastically in the three pig lines, two purebreds, and a crossbreed line. Notably, the BRANN provided the worst results in the purebreds data-set and the best predictive ability in the crossbred dataset, where non-additive genetic effects such as dominance are expected to affect the trait. Similarly, Long et al. (2011b) notice that an SVM model with a Gaussian radial basis function or a linear kernel provided similar predictive ability as the BLASSO for predicting sire estimated breeding values in dairy cattle. On the other hand, when used to predict phenotypes in a wheat data, the SVM with radial basis kernel outperformed the linear models, showing clear superiority compared to the BLASSO in the situation in which phenotypes may be affected by markers with non-additive effects (Long et al., 2011b).

As in RKHS, the choice of a suitable kernel may impact the predictive performance of the SVM model. In this regard, the radial basis function seems to perform better than other kernels for genome-enabled prediction of complex traits (Long et al., 2011b; Kasnavi et al., 2017). The radial basis SVM model provides an appealing alternative for genome-enabled prediction of complex traits, with flexibility

enough to cope with non-linear effects and at the same time robust to prevent overfitting. For instance, Yao et al. (2016) reported that a self-training SVM model with radial basis kernel is feasible for enhancing the accuracy of genomic prediction for residual feed intake in dairy cattle with small reference populations.

Machine learning methods also have been demonstrated as efficient pre-screening tools of important markers for enhancing the prediction accuracy of genomic breeding values (Li et al., 2018). In a Brahman cattle population, Li et al. (2018) demonstrate that it is possible to reduce the high dimensionality associated with large genomic data by identifying a subset of significant SNPs using different ensemble learning methods. Those authors notice that using the 3,000 top SNPs identified by a random forest model or gradient boosting machine (GBM) to construct genomic relationship matrices provide similar accuracy for body weight prediction when compared to models using the whole SNP panel. The average prediction accuracy of GEBVs was 0.43 using all SNPs, 0.42 using the subset of SNPs identified by the RF algorithm and 0.46 using the subset identified with GBM (Li et al., 2018).

Most machine learning methods were first proposed for dealing with the labeling of discrete data. Therefore, ML methods can be extended in a straightforward manner for analyzing complex discrete traits using genomic information (González-Recio and Forni, 2011; Heuer et al., 2016; Naderi et al., 2016; Sadeghi et al., 2018). The area under the receiver operating characteristic curve (AUC) metric has been commonly used as a criterion for evaluating the prediction accuracy of binary traits, such as susceptibility to some disorder or survival status. This metrics is obtained by comparing true positive and false positive discovering at different thresholds and can be interpreted as the probability that a given model assigns a higher score for a susceptible individual than a non-susceptible when both are chosen randomly from the population. Values of AUC closer to 1 are indicative of a better prediction model (González-Recio and Forni, 2011; González-Recio et al., 2014).

Comparing different simulated scenarios, González-Recio and Forni (2011), reported that the machine learning methods (RF and GBM) performed better than threshold Bayesian regression models (Bayes A and LASSO) to analyze discrete data associated with a small number of additive QTLs. Further, in the previous study, the RF had the highest AUC at classifying scrotal hernia in different pig lines, showing

better performance for correctly identifying susceptive animals (González-Recio and Forni, 2011).

Recently, a simulation study evidenced that the calibration group design and genetic architecture may influence the efficiency of the RF model for correctly classify health and diseased cows based on markers information (Naderi et al., 2016). The authors simulated different scenarios regarding the heritability, number of QTL and LD levels as well as the disease incidence in the training sets. Results pointed out that an increasing number of diseased animals in the training set improved the RF model performance whereas decreasing heritability, the number of QTLs and LD levels impact negatively on the AUC metric obtained with RF (Naderi et al., 2016).

The use of genomic information for assessing phenotype performance such as disease susceptibility, at the same time exploring nonadditive effects based on SNP data, provides a promising strategy for improving herds management (Yin and König, 2016). In this regard, machine learning methods seem to be a good choice in the animal breeder toolkit.

### *2.1.5 Machine learning applications in genome-wide association studies*

The availability of high-throughput genomic technologies has offered great opportunities for unraveling the biological processes governing economically important traits. Notably, the use of genomic information for identifying regions potentially associated with a phenotype of interest has demonstrated a great efficiency for prospecting candidate genes for different complex traits in many domestic species (Signer-Hasler et al., 2012; Le et al., 2017; Martínez et al., 2017; Melo et al., 2017, Mucha et al., 2018). Such an approach is commonly termed as a genome-wide association study (GWAS).

Nevertheless, the identified marker-phenotype associations have been accounting only partially for the total genetic variance, even for highly heritable traits. This phenomenon, known as missing heritability, has been described for human complex diseases (Manolio et al., 2009), in plants (Brachi et al., 2011) and in livestock animals (Shin et al., 2015). Furthermore, one must highlight that most statistical approaches for GWAS are focused only on main additive effects for individual markers,

while several other sources may affect the phenotype variation such as gene-gene or gene-environment interactions. Notwithstanding, detecting SNP interactions and other non-linear effects in GWAS is extremely challenging, among other reasons, due to the curse-of-dimensionality problem, genetic heterogeneity and computational complexity (Gilbert-Diamond and Moore, 2011).

As discussed backward, machine learning techniques provide some interesting features for dealing with genomic data. In the ML models, minimal or no assumptions about causal mechanisms are assumed, for genome-wide association purposes, this implies that ML algorithms may offer opportunities to identify novel potential causal variants when the true nature of the underlying associations between phenotype and markers are unknown and complex. Notwithstanding, there are a rather limited number of empirical applications of machine learning for GWAS in the animal breeding literature. Among the studied ML methods, the Random forest (RF) has been the most applied approach, mainly due to its simplicity and ability to identify important variables in large datasets. Simulation studies support RF as a useful approach to pre-screen candidate genes in animal breeding, especially for traits with high heritability and QTLs presenting major effects (Minozzi et al., 2014; Naderi et al., 2016).

RF has been also successfully applied to real datasets for identifying important genomic regions associated with complex traits. For instance, using the RF approach, Li et al. (2014) identified 2 markers that are strongly associated with sheep coat pigmentation. The previous authors also performed a genome-wide association study for pregnancy status in cattle and found candidate genes with an important role in reproductive performance, embryo growth diversity, and male germ development in humans. In a Canchim beef cattle genotyped with a high-density SNP panel, Mokry et al. (2013) explored the RF algorithm for identifying genomic regions associated with backfat thickness. The RF approach identified 70 SNPs associated with the response variable. Subsequently, a stepwise regression was performed to select the most relevant markers identified by the RF approach, providing a final subset with 21 SNPs which explained 53,27% of the deregressed EBV variance for backfat thickness. Interestingly, most SNP identified were associated with fat-related QTL in their chromosome region (Mokry at al., 2013).

Recently, Li et al. (2018) assessed the efficiency of three tree-based ensemble methods, Random Forest (RF), Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XgBoost), to identify a subset of relevant SNPs for genomic prediction of live weight breeding values in a Brahman cattle population. All three methods identified as the most important an SNP mapped to the gene *BMPER* (BMP binding Endothelial Regulator) located on BTA4 which plays vital roles in adipocyte differentiation, fat development and energy balance in humans and mice (Zhao et al., 2015). Gene ontology enrichment analysis showed that genes closest to the top 3,000 SNPs identified from each method presented similar biological functions, involved in the developmental process, visual perception, nervous system development and cellular activity (Li et al., 2018).

The variable importance scores provided by tree-based ensemble learning methods such as RF and GBM are useful to select the most relevant predictors while implicitly incorporating interaction effects. Such an appealing attribute has been extensively used for exploring gene-gene interactions associated with complex traits in humans (García-Magariños et al., 2009; Jiang et al., 2009; Li et al., 2016). In animal breeding, results from simulated data seem to provide pieces of evidence on the RF capability for discovering markers with non-additive effects (e.g. dominance and epistasis effects) in genome-wide association studies (Waldmaan, 2016).

Exploring such a property, Yao et al. (2013) used an RF approach for identifying single nucleotide polymorphisms potentially presenting additive and epistatic effects associated with residual feed intake in dairy cattle. Possible pairwise epistatic interaction between SNPs was identified by analyzing the trees structures produced within the RF, where the most frequent pairs of descendent nodes in the trees were assumed as possible epistatic interactions. Afterward, a linear regression model using all possible interactions between the top 25 descendent pairs was performed to validate the significant epistatic interactions (Yao et al., 2013). Those authors noticed that many high scored SNPs in the RF approach had much lower ranking in the Bayesian LASSO model analysis, demonstrating that SNPs with relatively small additive genetic effects may contribute to the residual feed intake genetic variance through pairwise interaction with other SNP.

## 3 References

Alonso J, Castañon AR, Bahamonde A (2013) Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter. **Computers and Electronics in Agriculture** 91:116-120.

Alados I, Mellado JA, Ramos F, Alados-Arboledas L (2004) Estimating UV erythemal irradiance by means of neural networks. **Photochemistry and Photobiology** 80:351-358.

Awad M, Khanna R (1 Eds.) (2015) Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers: APRESS, 268p.

Biau G, Scornet E (2016) A Random Forest Guided Tour. **TEST** 25:197-227.

Bishop CM (Eds) (2006) Pattern Recognition and Machine learning: SPRINGER, 738p.

Brereton RG, Lloyd GR (2009) Support Vector Machines for classification and regression. **Analyst** 135:230-267.

Boligon AA, Albuquerque LG, Mercadante MEZ, Lobo RB (2010) Study of relations among age at first calving, average weight gains and weights from weaning to maturity in Nellore cattle. **Revista Brasileira de Zootecnia** 39:746-751.

Boulesteix A-L, Janitza S, Kruppa J, König IR (2012) Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. **Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery** 2:493–50.

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability in this field. **Genome Biology** 12:232.

Breiman L, Friedman JH, Olshen R, Stone C (Eds) (1984) Classification and Regression Trees: CHAPMAN AND HALL/CRC, 368p.

Breiman L (1996) Bagging predictors. **Machine Learning** 24:123-140.

Breiman L (2001) Random forests. **Machine Learning** 45:5–32.

Brumatti RC, Ferraz JBS, Eler JP, Formigonni IB (2011) Desenvolvimento de índices de seleção em gado de corte sob enfoque de um modelo bioeconômico. **Archivos de Zootecnia** 60:205–13.

Carvalheiro R (2014) Genomic selection in Nelore cattle in Brazil. In: PROCEEDINGS 10TH WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. Vancouver – Canada, 2014. Available in: < https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/258_paper_10329_manuscript_1314_0.pdf?sfvrsn=2>. Access on 15 oct. 2015.

Campbell C (2000) An introduction to kernel methods. In.: Howlett RJ, Jain LC (Eds), **Radial Basis Function Networks: Design and Applications**. Berlin: SPRINGER VERLAG, p.155–192.

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. **Genomics** 99:323-329.

Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge: CAMBRIDGE UNIVERSITY PRESS, 204p.

De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semiparametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetics Research** 92:295-308.

De Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics** 193: 327-345.

Dias LT, El Faro L, Albuquerque LG (2004) Estimativas de herdabilidade para idade ao primeiro parto de novilhas da raça Nelore. **Revista Brasileira de Zootecnia** 33:97-102.

Diaz-Uriarte R, Andrés SA (2006) Gene selection and classification of microarray data using random forest. **BMC bioinformatics** 7:1-13.

Ehret A, Hochstuhls D, Gianola D, Thaller G (2015) Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesan and German Fleckviech cattle. **Genetics Selection Evolution** 47.

Foresee FD, Hagan MT (1997) Gauss-Newton approximation to Bayesian learning. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, Houston, TX, USA, p.1930–1935.

García-Magariños M, López-de-Ullibarri I, Cao R, Salas A (2009) Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction. **Annals of human genetics** 73:360-369.

Genuer R, Poggi J-M, Tuleau-Malot C (2010) Variable selection using random forests. **Pattern Recognition Letters** 31:2225-2236.

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics** 173:1761-1776.

Gianola D, De Los Campos G (2008) Inferring genetic values for quantitative traits non-arametrically. **Genetics Research** 90:525-540.

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. **Genetics** 183:347-363.

Gianola D, Okut H, Weigel K, Rosa G (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with jersey cows and wheat. **BMC Genetics** 12.

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A (2016) Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science** 57:229-236.

Gilbert-Diamond D, Moore JH (2011) Analysis of Gene-Gene Interactions. **Current Protocols Human Genetics** 70:1.14.1-1.14.12.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: Methodological consideration & new findings. **BMC Genetics**11.

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. **Genetics Selection Evolution** 43.

González-Recio O, ROSA GJM, GIANOLA D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science** 166:217-231.

Gunn SR (1998) Support Vector Machines for Classification and Regression. University of Southampton, School of Electronics and Computer Science. Technical Report.

Hastie T, Tibshirani R, Friedman J (Eds) (2009) The elements of statistical learning: data mining, inference and prediction, New York: SPRINGER, 745p.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited Review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science** 92:433-443.

Haykin SS (Eds) (2001) Redes Neurais: Princípios e práticas. Brasil: BOOKMAN, 900p.

Heuer C, Scheel C, Tetens J, Kühn C, Thaller G (2016) Genomic prediction of unordered categorical traits: an application to subpopulations assignment in German Warmblood horses. **Genetics Selection Evolution** 48.

Hearst MA, Dumais ST, Osman E, Platt JC, Scholkopf B (1998) Support Vector Machines. **IEEE Intelligent Systems** 13:18-28.

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **Genes, Genomes, Genetics** 4:1027-1046.

James G, Witten D, Hastie T, Tibishirani R (2013) An Introduction to Statistical Learning: with applications in R. New York: SPRINGER, 426p.

Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case-control studies. **BMC Bioinformatics** 10(Suppl. 1): S65.

Kasnavi SA, Afshar MA, Shariati MM, Kashan NEJ, Honarvar M (2017) Performance evaluation of support vector machine (SVM)-based predictors in genomic selection. **Indian Journal of Animal Sciences** 87:1226-1231.

Kurková V (1992) Kolmogorov's theorem and multilayer neural networks. **Neural Networks** 5:501-506.

Le TH, Christensen OF, Nielsen B, Sahana G (2017) Genome-wide association study for conformation traits in three Danish pig breeds. **Genetics Selection Evolution** 49:12.

Li Y, Kijas J, Henshall M, Lehnert S, McCulloch R, Reverter A (2014). Using random forests (RF) to prescreen candidate genes: A new prospective for GWAS. In: Proc. 10th World Congress Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Li J, Malley JD, Andrew AS, Karagas MR, Moore JH (2016) Detecting gene-gene interactions using a permutation-based random forest method. **Biodata Mining** 9:14.

Li, B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genome Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics** 9.

Libbrecht MW, Noble WS (2015) Machine Learning Applications in Genetics and Genomics. **Nature Reviews** 16:321- 332.

Long N, Gianola D, Rosa GMJ, Weigel KA, Kranis A, González-Recio O (2010) Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics Research** 92:209-225.

Long N, Gianola D, Rosa GJM, Weigel KA (2011a) Marker-assisted prediction of non-additive genetic values. **Genetica** 139:843–854.

Long N, Gianola D, Rosa GJM, Weigel KA (2011b) Application of support vector regression to genome-assisted prediction of quantitative traits. **Theoretical and Applied Genetics** 123:1065-1074.

Lorena AC, Carvalho ACPLF (2003) Introdução às Máquinas de Vetores Suporte. **Revista de Informática Teórica e Aplicada** 14.

Mackay DJC (1992) Bayesian Interpolation. **Neural Computation** 4:415-447.

Manolio TA, Collins FS, Cox NJ et al. (2009) Finding the missing heritability of complex diseases. **Nature** 461:747-753.

Martínez R, Bejarano D, Gómez Y, Dasoneville R, Jiménez A, Even G, Sölkner J, Mészáros G (2017) Genome-wide association study for birth, weaning and yearling weight in Colombian Brahman cattle. **Genetics and Molecular Biology** 40:453–459.

Melo TP, Takada L, Baldi F, Oliveira HN, Dias MM, Neves HHR, Schenkel FS, Albuquerque LG, Carvalheiro C (2016) Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations. **BMC Genetics** 17:89.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157:1819-1829.

Meuwissen THE, Hayes BJ, Goddard ME (2013) Accelerating Improvement of Livestock with Genomic Selection. **Annual Review of Animal Biosciences** 1:221–237.

Minozzi G, Pedretti A, Biffani S, Nicolazzi EL, Stella A (2014) Genome-wide association analysis of the 16th QTL-MAS Workshop dataset using the Random Forest machine learning approach. **BMC Proceedings** 8(Suppl. 5): S4

Mokry FB, Higa RH, Mudadu MA, Lima AO, Meirelles SLC, Silva MVGB, Cardoso FF, De Oliveira MM, Urbinati I, Niciura SCM, Tullio RR, De Alencar MM, Regitano LCA (2013) Genome-wide association study for back fat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics** 14.

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution** 41(56):1-16.

Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, Cornington J (2018) Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. **Journal of Dairy Science** 101:2213-2225.

Naderi S, Yin T, König S (2016) Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. **Journal of Dairy Science** 99:7261-7273.

Ogutu JO, Piepho H, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. **BMC Proceedings** 5.

Okut H, Wu X, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D (2013) Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. **Genetics Selection Evolution** 45:1-13.

Okut H, Gianola D, Rosa GJM, Weigel K (2015) Evaluation of prediction ability of Cholesky factorization of genetic relationship matrix for additive and non-additive genetic effect using Bayesian regularized neural network. **IORE Journal of Genetics** 1:1-15.

Okut H (2016) Bayesian Regularized Neural Networks for Small *n* Big *p* Data. *In*: Rosa JLG. **Artificial Neural Networks – Models and Applications**. INTECH, p.27-48.

Pereira BDB, Rao CR (2009) **Data Mining using Neural Networks: A Guide for Statisticians**. 186p. Available in: <https://www.researchgate.net/profile/Basilio_Pereira/publication/266274091_Data_Mining_Using_Neural_Networks_A_Guide_for_Statisticians/links/54b903400cf269d8cbf729c4/Data-Mining-Using-Neural-Networks-A-Guide-for-Statisticians.pdf>. Access on: 27 May 2016.

Pérez-Enciso M (2017) Animal breeding learning from machine learning. **Journal of Animal Breeding and Genetics**, 134:85-86.

Pontil M, Verri A (1998) Support Vector Machines for 3D Object Recognition. **EEE transactions on pattern analysis and machine intelligence** 20:637-646.

Rojas R (1996) Neural Networks – A Systematic Introduction. Berlin: SPRINGER-VERLAG, 502p.

Sadeghi S, Rafat SA, Alijani S (2018) Evaluation of imputed genomic data in discrete traits using Random Forest and Bayesian threshold models. **Acta Scientiarum. Animal Sciences** 40.

Salcedo-Sanz S, Rojo-Álvarez JL, Martínez-Ramón M, Camps-Valls G (2014) Support vector machines in engineering: an overview. **Data Mining Knowledge Discovery** 4:234–267.

Shin D, Park K, Ka S, Kim H, Cho K (2015) Heritability Estimates Using 50k SNPs indicates Missing Problem in Holstein Breed. **Genomics & Informatics** 13(4):146-151.

Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, Rieder S (2012) A Genome-Wide Association Study Reveals Loci Influencing Height and Other Conformation Traits in Horses. **Plos One** 7:e37282.

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. **Statistics and Computing**. 14:199-222.

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. **BMC bioinformatics** 8.

Su G, Christensen OF, Ostersen T, Henryon M, Lund Ms (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. **Plos One** 7(9).

Tussel L, Pérez-Rodriguez P, Forni S, Wu X-L, Gianola D (2013) Genome-enabled methods for predicting litter size in pigs: a comparison. **Animal** 7(11):1739-1749.

Upstill-Goddard R, Eccles D, Fliege J, Collins A, (2012) Machine learning approaches for the discovery of gene-gene interactions in disease data. **Briefings in Bioinformatics Advance** 1-10.

Vapnik V (1995) The Nature of Statistical Learning Theory. New York: SPRINGER, 314p.

Vapnik V, Golowich S, Smola A (1997) Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. *In*: Mozer M, Jordan M, Petsche T (Eds.) **Neural Information Processing Systems**. Cambridge: MIT PRESS.

Waldmaan, P (2016). Genome-wide prediction using Bayesian additive regression trees. **Genetics Selection Evolution** 48(42):1-2.

Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen Li (2012) Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. **Bioinformatics** 28:127-136.

Yang P, Yang YH, Zhou BB, Zomaya AY (2010) A review of ensemble methods in bioinformatics. **Current Bioinformatics** 5:296–308.

Yao C, Spurlock DM, Armentano LE, Page Jr CD, Vandehaar MJ, Bickhart DM (2013) Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. **Journal of Dairy Science** 96:6716–6729.

Yao C, Zhu X, Weigel KA (2016) Semi-supervised learning for genomic prediction of novel traits with small reference populations: an application to residual feed intake in dairy cattle. **Genetics Selection Evolution** 48:1-9.

Yin T, König S (2016) Genomics for phenotype prediction and management purposes. **Animal Frontiers** 6(1):65-72.

Zhao C, Gui L, Li Y, Plath M, Zan L (2015). Associations between allelic polymorphism of the BMP Binding Endothelial Regulator and phenotypic variation of cattle. **Molecular and Cellular Probes** 29:358–364.

# CHAPTER 2 – Genome-wide prediction for complex traits under the presence of dominance effects in beef cattle simulated populations using GBLUP and Machine learning methods

**ABSTRACT –** The aim of this study was to compare the predictive performance of the Genomic Best Linear Unbiased Predictor (GBLUP) and machine learning methods (Random Forest - RF, Support Vector Machine - SVM and Artificial Neural Network - ANN) in simulated populations presenting different levels of dominance effects. Simulated genome comprised 50k single nucleotide polymorphism (SNP) and 300 QTL (Quantitative Trait Loci), both biallelic and randomly distributed across 29 autosomes. A total of six traits were simulated considering different values for the narrow and broad-sense heritability. In the purely additive scenario with low heritability ($h^2$ = 0.10), the predictive ability obtained using GBLUP was slightly higher than the other methods whereas ANN provided the highest accuracies for scenarios with moderate heritability ($h^2$ = 0.30). The accuracies of dominance deviations predictions varied from 0.180 to 0.350 in GBLUP extended for dominance effects (GBLUP-D), from 0.06 to 0.185 in RF and they were null using the ANN and SVM methods. Although RF has presented the higher accuracies for total genetic effect predictions, the MSE (Mean-squared error) values in such a model were higher than those observed in GBLUP-D at large additive and dominance variances. When applied to pre-screen important regions, the RF approach detected QTL of high additive and/or dominance effects. Among machine learning methods, only the RF was capable to cover implicitly dominance effects without increasing the number of covariates in the model, resulting in higher accuracies for the total genetic and phenotypic values as the dominance ratio increases. Nevertheless, whether the interest is to infer directly about dominance effects, GBLUP-D could be a more suitable method.

**Keywords:** artificial neural network, genomic selection, non-additive effects, random forest, support vector machine

## 1 Introduction

Genome-wide dense marker availability on a commercial scale has brought a new paradigm to animal breeding. The use of such information for genome-enabled prediction of breeding values has allowed accelerating the genetic gains by providing early and more accurate prediction than pedigree-based approaches (Meuwissen et. al., 2013). Nevertheless, whole-genome prediction models have been typically formulated considering only additive effects, ignoring possible non-additive relationships, for instance, dominance effects caused by allele interactions at the same locus. Including this effect in genomic evaluation has theoretical advantages such as exploring the specific combining ability for enhancing progeny performance and may increase the accuracy of breeding values prediction, avoiding an overestimation, especially if dominance variance ratio is large (Toro and Varona, 2010; Aliloo et al., 2016).

Variance component estimation of dominance effects using either pedigree or genomic-based analysis have been ranging from null to a substantial contribution to the total genetic variance of different complex traits (Fuerst and Sölkner, 1994; Rodriguez-Almeida et al., 1995, Van Tassel et al., 2000; Gallardo et al., 2010; Su et. al., 2012; Nagy et al., 2013; Aliloo et al., 2016). Nonetheless, accounting for non-additive effects may increase the model parameterization with the construction and inversion of large and dense matrices, leading to an intensive computational cost (Su et. al., 2012). In addition, although parametric models as Genomic Best Linear Predictor (GBLUP) and Bayesian regressions have been shown to be robust for genomic prediction, such models rely on strong assumptions that not always hold in practice (Okut et al., 2013).

Recently, machine learning theory has been expanded to a genomic prediction scope, mainly due to its theoretical flexibility to cope with complex relationships between markers and phenotypes. Such approaches can deal with the dimensionality problem in an adaptive way, without imposing any specific relationship between phenotypes and genotypes, providing appealing attributes that make them well suited for genomic data analysis (Gonzalez-Récio et al., 2014).

Previous studies using simulated data reported similar or better prediction ability for machine learning methods such as Support Vector Regression, Random Forest and Artificial Neural Networks when compared with GBLUP or Bayesian regression models (González-Recio and Forni, 2011; Ogutu et al., 2011; Howard et al., 2014, Ghafouri-Kesbi, et al., 2016). Nevertheless, comparisons have been performed mainly for scenarios under purely additive effects which may not represent real data conditions. Thus, the aim of this study was to compare the predictive performance of GBLUP and machine learning methods in simulated populations presenting different levels of dominance.

## 2 Material and methods

### 2.1 Simulated data

The simulation procedures were performed according to previous simulation studies considering dominance genetic effects (Toro et al., 2010; Nishio et al., 2014; Martini et al., 2017). Genotype data including markers and QTL (Quantitative Trait Loci) spread across the genome were simulated using QMSim software (Sargozalei and Schenkel, 2009). First, a historical population with 1,000 animals was simulated with random mating and constant population size during 1,000 generations and then gradually reduced to 100 individuals in additional 1,020 generations. This step aimed to create the linkage disequilibrium and to allow mutation-drift equilibrium establishment. Recurrent mutation process was assumed for both marker and QTL, with a mutation rate of $5x10^{-4}$. In order to expand the resultant population, the remained animals (50 males and 50 females) were randomly mated by an additional five generations assuming five offspring per dam and exponential growth of the number of dams.

Finally, 100 males and 1,500 females from the last generation of the expanded population were assumed to be the base population (G0) and additional five generations were simulated as recent population (G1 to G5), assuming 1 offspring per dam with equal probability of being male or female, resulting in a total of 9100 animals (G0 to G5). The replacement rates of sires and dams were kept constant at 60% and

20%, respectively. Minor allele frequencies of markers and QTL on the G0 population were set to be ≥ 0.05. The simulated genome comprised 50k single nucleotide polymorphism (SNP) markers and 300 QTL, both biallelic and randomly distributed across 29 autosomes, with a total length of 2,320 centimorgans.

The resultant simulated populations were used to model complex traits affected by purely additive effects or presenting different degrees of dominance. Simulations of genotypic values were performed in terms of breeding values and dominance deviations, with the assumption of Hardy-Weinberg equilibrium (Falconer and Mackay, 1996), as described in Vitezica et al (2013):

$$g = E\ (g) + \ z\alpha + wd,$$

with z and w coded as:

$$z_i = \begin{cases} (2 - 2p_j) \\ (1 - 2p_j) \\ -2p_j \end{cases} \text{ for genotypes } \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

and,

$$w_i = \begin{cases} -2q_j^2 \\ 2p_jq_j \\ -2p_j^2 \end{cases} \text{ for genotypes } \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

where $E\ (g) = 0,\ p_j$ and $q_j$ are the true alleles frequencies for $A_1$ and $A_2$ at the *jth QTL* and $\alpha_j = \ \mathrm{a}_j + \ d_j(q_j - p_j)$ is the allele substitution effect. The additive effect $\mathrm{a}_j$ for each QTL was sampled from a gamma distribution with shape and scale parameters of 0.42 and 1.66, respectively, with positive and negative signs drawn with equal chance. The dominance deviations $d_j$ were determined as $|\mathrm{a}_j|\delta_j$, where $\delta_j$ is the degree of dominance, which was initially drawn from a normal distribution *N*(0, 1). The additive and dominance effects for each QTL were scaled to achieve the desirable variances for each scenario.

Final additive ($\sigma_a^2$) and dominance ($\sigma_d^2$) variances were computed as (Falconer and Mackay, 1996):

$$\sigma_a^2 = \ \sum_{j=1}^{N_{QTL}} 2p_j(1 - p_j)\left\{a_j + (1 - 2p_j)d_j\ \right\}^2$$

and,

$$\sigma_d^2 = \sum_{j=1}^{N_{QTL}} \{2p_j(1-p_j)d_j\}^2$$

Consequently, the total genetic variance ($\sigma_g^2$) was partitioned as:

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2$$

Residual effects were sampled from a normal distribution, with $e \sim N(0, \sigma_e^2)$, and added to the total genetic effects in order to achieve a phenotypic variance of 1 for all scenarios. Therefore, the observed phenotypes were computed as:

$$y = E\ (g + e) + \ z\alpha + wd + e$$

Six traits were simulated considering different narrow-sense ($h^2$ = 0.10 and 0.30) and broad-sense ($H^2$ = 0.10, 0.15, 0.20, 0.30, 0.45 and 0.60) heritability values. After stochastic simulation procedure, 1500 animals were randomly selected from generation G1 to G4 to compose the reference population and 500 animals from G5 were randomly selected as testing population, for which only genotypes were assumed to be known. We repeated this procedure for 10 replicates in each scenario.

## 2.2 Prediction models

### *2.2.1 Genomic best-unbiased prediction (GBLUP)*

The general model for the Genomic Best Linear Unbiased Prediction (GBLUP) can be written as:

$$y = 1_n u + Zg + e$$

where *y* is the vector of phenotypes, $1_n$ is a vector of 1's, *u* is an overall mean, *Z* is an incidence matrix relating the animals to the additive effects, *g* is a vector of direct genomic breeding values and *e* is a vector of residuals. The GBLUP model extended for dominance effects (GBLUP-D) can be represented as follows:

$$y = 1_n u + Zg + Wd + e$$

in which *W* is an incidence matrix relating the animals to dominance deviations and *d* is a vector of genomic dominance deviations. Both g and *d* were assumed to be normally distributed with $g \sim N(0, G\sigma_a^2)$ and $d \sim N(0, D\sigma_d^2)$, where *G* and *D,* are genomic relationship matrices for additive and dominance effects, respectively, with $\sigma_a^2$ and $\sigma_d^2$

representing the respective variances of such effects. The G matrix was constructed as described in VanRaden (2008):

$$G = \frac{M_a M_a{}'}{\sum_{j=1}^{N_m} 2p_j(1-p_j)}$$

where $M_a$ is a *n x* $\mathrm{N_m}$ matrix (n is the number the individuals and $\mathrm{N_m}$ is the number of markers), in which the elements for the *i*$^{th}$ individual and *j*$^{th}$ marker are equal to $-2p_j$ , $1-2p_j$ and $2-2p_j$ for the aa, Aa and AA genotypes, respectively, with $p_j$ representing the expected frequencies of the allele A at the *j*$^{th}$ marker in the population. Similarly, the D relationship matrix was constructed following Vitezica et al. (2013):

$$D = \frac{M_d M_d{}'}{\sum_{j=1}^{N_m}\{2p_j(1-p_j)\}^2}$$

in which $M_d$ is a *n x* $N_m$ matrix with the elements coded as $-2p_j^2$, $2p_j(1-p_j)$ and $-2(1-p_j)^2$ for aa, Aa and AA genotypes, respectively. The GBLUP and GBLUP-D models were fitted using the *BGLR* package (Pérez and De Los Campos, 2014).

### *2.2.2 Random forest (RF)*

The Random Forest algorithm (Breiman, 2001) uses an ensemble of unpruned decorrelated decision trees, built from *B* bootstrap samples of the training data set and randomly selecting a subset of the original predictor variables as candidates for splitting tree nodes. This ensemble of *weak learners* can be used for prediction of an unobserved data by averaging all *B* predictors $\{T(x,\psi_b)\}_1^B$ as:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T(x,\psi_b)$$

where $\psi_b$ represents an individual *b* tree architecture in terms of split variables, cut point at each node and terminal node values. In an RF model, samples that are not selected (roughly one-third of the total observations) on *bootstrapping*, termed out-of-bag (OOB) samples, are used as internal validation from which the OOB error is computed. A common error measure adopted in regression problems is the mean square error of the OOB data:

$$MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

in which *N* is the number of observations in the OOB samples, $\hat{y}_i$ is the average of the predictions for the *i*$^{th}$ animal from trees in which it was OOB, and $y_i$ is the realized value of the animal.

In the RF approach, variable predictors can be ranked by the importance of their contributions to predictive accuracy. One of the possible ways for computing the variable importance measure (VIM) is by accounting how much OOB error increases when a given variable predictor (for example an SNP) is randomly permutated on the OOB data while all other predictor variables left unchanged. The relative variable importance can be calculated as the difference between the original predictive measure (without permuting the variable on the OOB sample) and that of the OOB with the permuted variable. This step is repeated for each covariate (SNP) and the decrease of accuracy is averaged over all trees in the random forest. Important variables for the outcome prediction are expected to have higher VIM since the permutation of such variables on the validation data will increase the prediction error. We implemented the RF model in the *randomForest* R package (Liaw and Wiener, 2013). The model parameters used in the present study were fixed as *ntree* = 1000 (number of trees to grow) and *mtry* = $\sqrt{p}$ (number of SNPs selected at each tree node) and *nodesize* as default.

### *2.2.3 Support vector machines (SVM)*

SVM utilizes linear models to implement non-linear regressions, by mapping the predictors in a feature space of different dimensions using kernels inner products, followed by linear regression on the resulting observed space. The general model can be viewed as (Hastie et al., 2009):

$$\hat{y} = \beta_0 + h(x)^T\beta$$

where $h(x)^T$ represents a linear or nonlinear transformation of the original input space featured by a given kernel function (*h*), here, the radial basis function, $\beta_0$ is a constant and $\beta$ are the weights for each variable on the feature space. For the risk minimization, we adopted the '***ε-insensitive***' loss function:

$$H(\beta_0, \beta) = \sum_{i=1}^{N} V_\varepsilon(y_i - f(x_i)) + \frac{C}{2}\|\beta\|^2,$$

in which:

$$V_\varepsilon(y_i - f(x_i)) = \begin{Bmatrix} 0 \text{ if } |y_i - f(x_i)| < \varepsilon, \\ |y_i - f(x_i)| - \varepsilon, \text{otherwise} \end{Bmatrix},$$

is a function which sets an insensitive tube around the residuals, ignoring the errors within the tube (less than ε), C is a regularization parameter that controls the trade-off between the complexity of the loss function and the training error, and $\|.\|^2$ denotes the norm under a Hilbert Space. According to Hastie et al., 2009 if $\hat{\beta}$ and $\hat{\beta}_0$ are the constants that minimize *H,* the solution function has the following form:

$$\hat{\beta} = \sum_{i=1}^{N}(\hat{\alpha}_i^* - \hat{\alpha}_i)x_i,$$

$$f(x) = \sum_{i=1}^{N}(\hat{\alpha}_i^* - \hat{\alpha}_i)K\langle x, x_i\rangle + \beta_0$$

where $\hat{\alpha}_i$, $\hat{\alpha}_i^*$ are positive weights given to each observation and the inner product $K\langle x_i, x_{i'}\rangle$ is an *N x N* symmetric and positive definite kernel matrix.

The parameters ε and C were defined as $3\sigma_e\left(\sqrt{\ln n/n}\right)$ and $\max(|\bar{y} - 3\sigma_y|, |\bar{y} + 3\sigma_y|)$, respectively while the kernel bandwidth was defined by the grid-search procedure on the training data. The *kernlab* R package (Karatzoglou et al., 2004) was used on the model construction.

### *2.2.4 Artificial neural network (ANN)*

A multilayer perceptron (MLP) neural network, with a single hidden layer and two neurons, was used in this study to predict the total genetic values. In order to reduce computational costs, only the top 1% SNPs ranked by importance scores of the RF algorithm were used as input variables. The model can be described as:

$$y_i = \sum_{k=1}^{s} w_k\, g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\, \beta_j^{[k]}\right) + e_i$$

where: $e_i \sim N(0, \sigma_e^2)$, *s* is the number of neurons, $w_k$ is the weight for the $k^{th}$ neuron, $b_k$ is the bias for the $k^{th}$ neuron, $\beta_j^{[k]}$ is the weight of the $j^{th}$ input to the net and $g_k(.)$ is a given activation function with $g_k(x) = \frac{\exp(2x)-1}{\exp(2x)+1}$.

The Gauss-Newton algorithm, implemented in the *brnn* R package (Pérez-Rodrigues and Gianola, 2013) was used to perform the weights updates. To allow better generalization for the ANN architecture, Bayesian regularization was used on the learning process. The objective function to be minimized is:

$$E_D(D|w,M) = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2;$$

where *D* denotes a given dataset used on the network, *w* are the weights, and *M* is a given network architecture, in terms of the number of neurons and activation functions. Bayesian regularization produces shrinkage of the parameters estimates in order to reduce its variance and, the objective function becomes:

$$F = \beta E_D(D|w,M) + \alpha E_w(w|M);$$

in which $E_w(w|M)$ is the sum of squares of the network weights and $\alpha$ and $\beta$ are positive regularization parameters.

### 2.3 Assessing prediction accuracy

In GBLUP and GBLUP-D models, the prediction accuracy of the additive (*a*), dominance (*d*) and total genetic effects (*g*) were measured as the Pearson correlation between predicted and true values ($r_{\hat{a},a}$, $r_{\hat{a},d}$, $r_{\hat{g},g}$, respectively). For the GBLUP-D model, the prediction of the total genotypic effect ($\hat{g}$) was calculated by summing up $\hat{a}$ and $\hat{d}$, whereas in GBLUP, $\hat{g}$ was equal to $\hat{a}$. In the machine learning methods (ML), no additive or dominance structures were imposed in the genotype matrix, so that it is not possible to compute directly the predicted values for *a*, *d* or *g*. However, in order to assess which effects (additive and/or dominance), the ML methods are capturing, the prediction accuracy was assessed by the correlation between the predicted responses ($\hat{y}$) and the true additive, dominance and total genetic values ($r_{\hat{y},a}$, $r_{\hat{y},d}$, $r_{\hat{y},g}$, respectively). Mean-Squared error (MSE) of the total genetic values predictions was also computed to compare the prediction ability of methods.

Since the true additive and dominance effects are unknown in real populations, there were also evaluated in this study, the correlations between the observed simulated phenotypes and the predicted responses, *i.e.* the estimated genomic breeding values in the GBLUP, the estimated total genetic values in the GBLUP-D and

the $\hat{y}$ for RF, SVM or ANN. All predictive ability metrics were averaged over the ten replicates in each scenario.

## 3 Results and discussion

### 3.1 Extent of linkage disequilibrium

Genomic prediction accuracy is highly dependent on the extent of linkage disequilibrium between markers and QTL on the population (Meuwissen et. al., 2001). The LD between marker and QTL (generally unobserved) can be viewed as the proportion of the variation caused by the alleles at a QTL, explained by the marker (Hayes, 2009). Thus, high LD between markers in a specific genomic region is expected to capture the signal of the alleles of each QTL in that region (Meuwissen et. al., 2001; Goddard, 2009).

In the present study, the simulations were performed in order to mimic the extent of LD in real beef cattle populations. In general, the LD measured by the $r^2$ statistic, considering adjacent markers, has been reported in the literature ranging from 0.17 to 0.31, for different breeds and panel densities (Lu et al., 2012, Espigolan et al., 2013; Pérez O'Brien et al., 2014; Fernandes Junior et. al., 2016). In our study, the average $r^2$ between adjacent markers across all replicates was equal to 0.24, thus within the interval reported by those authors.

The LD decay according to the distance range between markers across all simulated populations is in Figure 1. It is worth mentioning that such an LD decay pattern was similar to that observed by Lu et al. (2012), who worked with a 50k panel density in an Angus cattle population.

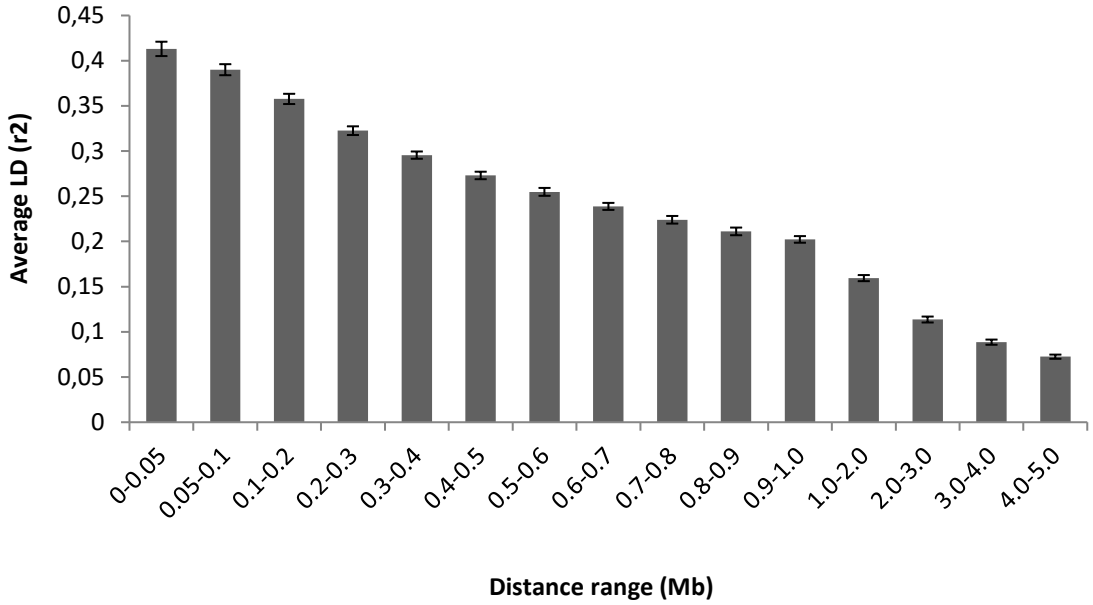


**Figure 1.** Linkage disequilibrium (LD) decay measured by $r^2$ statistic, according to different distance range. The results are presented as the overall average of 10 replicates.

### 3.2 Genomic additive and dominance relationship matrices

The histograms of the off-diagonal elements of additive and dominance relationship matrices for one specific replicate are depicted in Figure 2. G matrix off-diagonal elements presented an average of -0.0038 ± 0.04, with values ranging from -0.1715 to 0.759 and, for the D matrix, these presented mean equal to 0.0000197 ± 0.0212 with values ranging from -0.136 to 0.307. The standard deviation of the off-diagonal elements of G is approximate twice the observed for D, which is expected since the additive relationship matrix is supposed to be more informative than the dominance relationship matrix. Further, the means of the off-diagonal on both G and D matrices are very close to zero, indicative of Hardy-Weinberg equilibrium (HWE) on the simulated population (Vitezica et al., 2013). When the population is not in HWE $M_a$ and $M_d$ contrasts are not necessarily orthogonal, which violates the model assumptions and may cause bias on genetic parameter estimates (Bolormaa et al., 2015; Vitezica et al., 2013). However, a recent study proposes an approach to build non-additive genomic relationship matrices on populations deviating from HWE (Vitezica et al., 2017).

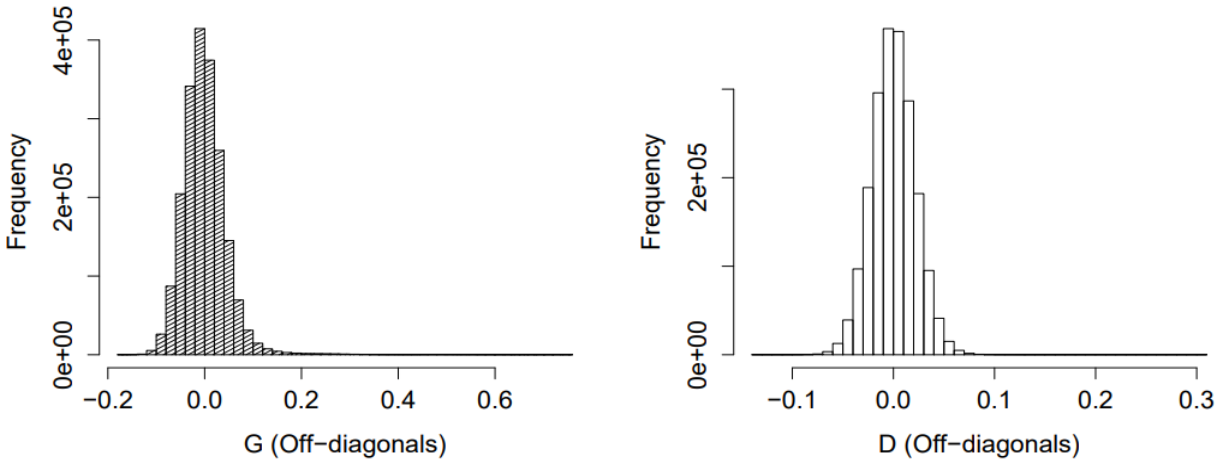


**Figure 2.** Histograms of the off-diagonal elements for additive (left) and dominance (right) genomic relationship matrices.

### 3.3 Prediction Accuracy

In the purely additive scenario with low heritability ($h^2$ = 0.10), GBLUP presented slightly better prediction accuracy of breeding values than machine learning methods, with accuracies ranging from 0.359 to 0.376 depending on the model (Table 1). The ANN model provided higher accuracies in the scenarios with moderate heritability ($h^2$ = 0.30), for which accuracy ranged from 0.563 to 0.635 (Table 1). It is possible that the ANN model has been benefited from the predictive variables selection performed *a priori* by the RF model and, such a result needs further investigation. Additionally, higher correlations between predicted and true breeding values across models were observed as the narrow-sense heritability increased (0.10 to 0.30), which was expected since, at increasing narrow-sense heritability, phenotypic variation is more explained by additive genetic effects.

In the absence of dominance effects, the accuracies of breeding values were slightly lower for the GBLUP-D method than for GBLUP (Table 1). However, this difference was more pronounced for the total genetic values prediction in the purely additive scenarios, as the inclusion of dominance in the absence of such effect is a confounding factor. On the other hand, as the dominance variance increased, the opposite was observed and total genetic predictions become more accurate in GBLUP including both additive and dominance effects (Table 1). Similar results have been reported in another simulation study (Nishio and Satoh, 2014).

**Table 1.** Prediction accuracies[1] and standard deviations obtained by GBLUP and different machine learning methods for simulated traits with different proportions of additive variance ($h^2$) and dominance deviations variance ($d^2$).

| Scenario | Effects[2] | Methods[3] | | | | |
|---|---|---|---|---|---|---|
| | | GBLUP | GBLUP-D | RF | SVM | ANN |
| $h^2$ = 0.10 $d^2$ = 0.00 | Add | 0.376±0.06 | 0.373±0.06 | 0.359±0.09 | 0.360±0.06 | 0.357±0.09 |
| | Dom | - | - | - | - | - |
| | Gen | 0.376±0.06 | 0.328±0.06 | 0.359±0.09 | 0.360±0.06 | 0.357±0.09 |
| $h^2$ = 0.10 $d^2$ = 0.05 | Add | 0.411±0.08 | 0.413±0.08 | 0.399±0.09 | 0.377±0.07 | 0.378±0.09 |
| | Dom | - | 0.180±0.06 | 0.062±0.04 | 0.004±0.03 | 0.007±0.03 |
| | Gen | 0.340±0.08 | 0.345±0.08 | 0.364±0.09 | 0.313±0.08 | 0.306±0.09 |
| $h^2$ = 0.10 $d^2$ = 0.10 | Add | 0.393±0.04 | 0.396±0.04 | 0.419±0.05 | 0.369±0.03 | 0.383±0.05 |
| | Dom | - | 0.237±0.09 | 0.134±0.10 | 0.018±0.03 | 0.001±0.07 |
| | Gen | 0.261±0.05 | 0.306±0.05 | 0.395±0.07 | 0.249±0.05 | 0.273±0.05 |
| $h^2$ = 0.30 $d^2$ = 0.00 | Add | 0.595±0.03 | 0.592±0.03 | 0.585±0.05 | 0.579±0.03 | 0.632±0.04 |
| | Dom | - | - | - | - | - |
| | Gen | 0.595±0.03 | 0.579±0.03 | 0.585±0.05 | 0.579±0.03 | 0.632±0.04 |
| $h^2$ = 0.30 $d^2$ = 0.15 | Add | 0.575±0.05 | 0.575±0.05 | 0.589±0.07 | 0.566±0.04 | 0.619±0.04 |
| | Dom | - | 0.286±0.07 | 0.163±0.06 | 0.016±0.05 | 0.041±0.07 |
| | Gen | 0.460±0.05 | 0.485±0.05 | 0.575±0.06 | 0.454±0.04 | 0.527±0.05 |
| $h^2$ = 0.30 $d^2$ = 0.30 | Add | 0.575±0.06 | 0.582±0.06 | 0.611±0.04 | 0.563±0.06 | 0.635±0.04 |
| | Dom | - | 0.350±0.05 | 0.185±0.06 | 0.021±0.04 | 0.053±0.06 |
| | Gen | 0.408±0.05 | 0.488±0.05 | 0.555±0.04 | 0.406±0.05 | 0.478±0.04 |

[1]Prediction accuracies for the breeding values (a), dominance deviations (d) and total genetic effects (g) were assessed as the Pearson correlation between predicted and true effects ($r(\hat{a}, a)$, $r(\hat{d}, d)$ and $r(\hat{g}, g)$, respectively) in GBLUP and GBLUP-D models and by the correlation between predicted responses ($\hat{y}$) and the true effects (a, d or g) for machine learning methods. Prediction accuracies are presented as the average of 10 replicates; [2]Add = additive effects, Dom = dominance effects and Gen = total genetic effects; [3]GBLUP = Genomic Best Linear Unbiased Predictor considering only additive effects, GBLUP-D = GBLUP considering both additive and dominance effects, RF = Random Forest, SVM = Support Vector Machine, ANN = Artificial Neural Network.

Regardless of the used method, there was an overall decrease in the accuracies of the total genetic predictions when dominance effects were present. These results are an indication that dominance effects may not be effectively accounted for in the prediction models as the additive effects are (de Almeida Filho et al., 2016), suggesting that a considerably larger data set is required to accurately predict the dominance deviations in comparison to the additive effects.

The accuracies of dominance deviation predictions ranged from 0.180 to 0.350 in the GBLUP-D and from 0.060 to 0.185 in the RF models, throughout the different scenarios. For the ANN and SVM models, those accuracies were close to zero (Table 1). Similarly, Nishio and Satoh (2014) have reported accuracy values for dominance deviations between 0.148 and 0.348, using the GBLUP-D model in a simulated population with five chromosomes.

As highlighted before, it is known that the genomic selection accuracy depends directly on the magnitude of LD, as a consequence, the proportion of additive variance explained by an observed marker decreases linearly as the $r^2$ between such marker and the causal variant decreases. For the dominance variance, such a relationship reduces by a factor of $r^4$, which implicates that much larger LD is necessary to detect dominance effects (Wei et al., 2014).

In the present study, the average LD measured by $r^2$ statistic reflects the pattern found in commercial beef cattle populations, thus, similar accuracies would be expected in real populations. Nonetheless, other aspects such as the trait architecture in terms of number and distribution of QTL effects, presence of high order non-linear marker relationships (*e.g.*, epistasis, genotype by environmental interactions, imprinting) and number of animals in the training set are expected to impact on the observed accuracies as well (Goddard et al., 2011).

Among machine learning methods, only the Random Forest algorithm was capable to capture implicitly the dominance signals (Table 1), although it is possible to construct specific dominance kernels for RKHS based models such as SVM. In the ANN, a more straightforward approach to model both additive and dominance effects would be to use directly the G and D matrices as input variables in the net architecture (Pérez-Rodrigues and Gianola, 2013). However, this approach would highly increase computational requirements (compared to models using only G), which would be unfeasible in practical applications. Further, there are no clear advantages of such ANN architecture (using G matrix) for genome-enabled prediction over a benchmark approach such as GBLUP (Howard et al., 2014; Ehret et al., 2015). Nevertheless, it is worth mentioning that ANN and SVM are powerful methods to cover other non-additives effects such as epistasis (Beam et al., 2014; Howard et al., 2014), not studied here.

The higher accuracies for dominance deviations predictions using the GBLUP-D model could be explained by the fact that this method handles such effect directly. In contrast, it can be noted that as the dominance variance increases, compared to the parametric models, the RF method predictions tended to present a higher correlation with the breeding values and, notably, with the genotypic values (Table 1). It is worth mentioning that RF predictions do not provide interpretable inferences about the additive or dominance effects since it combines all sources of genetic effects (additive and dominance in the present study) in a unique overall prediction. Nonetheless, the RF predictions would be useful to identify most productive animals or with susceptibility to diseases by capturing additive and non-additive signals.

Although the RF model has presented the highest accuracies for the total genetic predictions, in some cases, this method was associated with higher MSE values compared to those observed in GBLUP-D, particularly at a large additive and/or dominance variance (Table 2).

**Table 2**. Mean squared errors (MSE) and standard deviations of total genetic values predictions using GBLUP and different machine learning methods for simulated traits considering different levels of broad-sense heritability ($H^2$).

| Scenario* | Methods[1] | | | | |
|---|---|---|---|---|---|
| | **GBLUP** | **GBLUP-D** | **RF** | **SVM** | **ANN** |
| $H^2$ = 0.10 ($d^2$ = 0.00) | 0.09 (0.01) | 0.09 (0.01) | 0.09 (0.01) | 0.11 (0.01) | 0.23 (0.03) |
| $H^2$ = 0.15 ($d^2$ = 0.05) | 0.13 (0.01) | 0.13 (0.01) | 0.13 (0.01) | 0.15 (0.01) | 0.25 (0.04) |
| $H^2$ = 0.20 ($d^2$ = 0.10) | 0.20 (0.04) | 0.19 (0.01) | 0.18 (0.02) | 0.22 (0.02) | 0.30 (0.03) |
| $H^2$ = 0.30 ($d^2$ = 0.00) | 0.20 (0.01) | 0.21 (0.02) | 0.24 (0.02) | 0.20 (0.02) | 0.22 (0.02) |
| $H^2$ = 0.45 ($d^2$ = 0.15) | 0.34 (0.02) | 0.33 (0.03) | 0.35 (0.03) | 0.34 (0.02) | 0.36 (0.03) |
| $H^2$ = 0.60 ($d^2$ = 0.30) | 0.51 (0.04) | 0.46 (0.05) | 0.51 (0.03) | 0.51 (0.04) | 0.51 (0.04) |

*In the absence of dominance contribution to phenotype variance ($\overline{d^2}$), broad-sense heritability ($H^2$) equals to the narrow-sense heritability ($h^2$); [1]GBLUP = Genomic Best Linear Unbiased Predictor considering only additive effects, GBLUP-D = Genomic Best Linear Unbiased Predictor considering both additive and dominance effects, RF = Random Forest, SVM = Support Vector Machine, ANN = Artificial Neural Network; MSE is presented as the average of 10 replicates.

In turn, the ANN model presented poor predictive ability in the low heritability scenarios, with the highest MSE values, approximately twice those obtained in the other methods. Since the ANN model was built considering only the top 1% SNPs ranked with the RF algorithm, such a result may be partially due to the fact that, at low heritability levels, the power to detect relevant regions deeply decreases (van den Berg

et. al., 2013). In addition, it is known that ANN models are prone to over-fitting which affects its predictive ability and generalization capability (Lawrence et al., 1997).

Table 3 shows the average accuracy for the RF and GBLUP methods, considering the scenario with $h^2$ = 0.30 and $d^2$ = 0.15 and increasing the reference sample size. The accuracy of dominance effects improved from 0.181 to 0.418 for the GBLUP-D model when the number of animals in the reference population increased from 500 to 3,500.

**Table 3.** Average prediction accuracies[1] and standard deviations for a simulated trait with broad-sense heritability ($H^2$) equal to 0.45 ($h^2$=0.30 and $d^2$=0.15) according to the number of animals in the training set and the used method.

| Method[2] | Effects[3] | Training population size | | | |
|---|---|---|---|---|---|
| | | 500 | 1500 | 2500 | 3500 |
| GBLUP | Add | 0.423±0.06 | 0.575±0.05 | 0.673 ± 0.04 | 0.718 ± 0.03 |
| | Dom | - | - | - | - |
| | Gen | 0.341±0.05 | 0.460±0.05 | 0.546 ± 0.05 | 0.584 ± 0.04 |
| GBLUP-D | Add | 0.422±0.06 | 0.575±0.05 | 0.678 ± 0.04 | 0.722 ± 0.03 |
| | Dom | 0.181±0.10 | 0.286±0.07 | 0.379 ± 0.07 | 0.418 ± 0.06 |
| | Gen | 0.363±0.03 | 0.485±0.05 | 0.589 ± 0.05 | 0.634 ± 0.04 |
| RF | Add | 0.448±0.05 | 0.589±0.07 | 0.637 ± 0.04 | 0.651 ± 0.02 |
| | Dom | 0.056±0.06 | 0.163±0.06 | 0.171 ± 0.06 | 0.183 ± 0.06 |
| | Gen | 0.401±0.04 | 0.575±0.06 | 0.619 ± 0.05 | 0.635 ± 0.03 |

[1]Prediction accuracies for the breeding values (a), dominance deviations (d) and total genetic effects (g) were assessed as the Pearson correlation between predicted and true effects ($r(\hat{a}, a)$, $r(\hat{d}, d)$ and $r(\hat{g}, g)$, respectively) in GBLUP and GBLUP-D models and by the correlation between predicted responses ($\hat{y}$) and the true effects (a, d or g) for the Random Forest. Prediction accuracies are presented as the average of 10 replicates; [2]GBLUP = Genomic Best Linear Unbiased Predictor considering only additive effects, GBLUP-D = GBLUP considering both additive and dominance effects; RF = Random Forest; [3]Add = additive effects, Dom = dominance effects and Gen = total genetic effects.

The increase in the number of animals is probably related to an increase in the number of animals with heterozygous genotypes at each *locus*, improving the dominance deviation predictions. However, in the RF method, there was observed only a little improvement in the correlation between predicted values and dominance deviations. Consequently, despite the superiority of the RF over parametric methods to predict total genetic values, the differences observed in the prediction accuracies between GBLUP-D and RF decreased rapidly with the training population increasing.

With a training set of 3,500 animals, GBLUP-D and RF models presented similar accuracies for the total genetic effects, with values higher than those obtained by the GBLUP model (Table 3).

In practice, the true additive genetic and dominance effects are unknown on real populations, so the validation of genomic prediction models including dominance effects have been based, for instance, on the correlation between predicted genetic effects and individually adjusted phenotypes (Su et. al., 2012; Ertl et al., 2015; Aliloo et. al., 2016). Increasing the dominance variance, the accuracy derived by the correlation between phenotypes and the estimated total genetic effects also increases for GBLUP-D and RF models, considering a reference population with 1500 animals (Figure 3).

In the low narrow-sense heritability scenarios, as $d^2$ increased from 0 to 0.10, the prediction accuracies improved from 0.097 to 0.141 in GBLUP-D and from 0.105 to 0.145 in RF, respectively. In the moderate narrow-sense heritability scenarios, as $d^2$ increased from 0 to 0.30, the phenotype prediction accuracies improved from 0.322 to 0.375 in GBLUP-D and from 0.325 to 0.434 in RF, respectively (Figure 3).

There was no evident improvement in the phenotype prediction using GBLUP-D when broad-sense heritability was equal to 0.45 ($h^2 = 0.30$ and $d^2 = 0.15$) compared to the results obtained with GBLUP. This result is probably because the residual noise masks the total genetic predictions when accuracy improvement of such an effect is not substantial. However, as reported in Table 3, the accuracy of the total genetic prediction tends to improve by increasing the training sample size, thus, gains on the phenotype prediction accuracies are also expected.

The phenotype of an animal can be viewed as a combination of its total genetic merit and environmental deviations. Once the total genetic merit is a function of both additive (breeding values) and non-additive (dominance and epistasis) genetic effects (Falconer and Mackay, 1996), the assessment of future performance based on the total genetic merit instead of the breeding values is expected to identify more accurately the most productive animals. Such a strategy can be used to support culling decisions and to improve the overall herd production. In dairy cattle, Aliloo et al. (2016) reported better predictions of phenotypes including dominance effects on the genomic

analysis, albeit the observed differences were not significant, except for fat yield in the Holstein cows ($p < 0.01$).

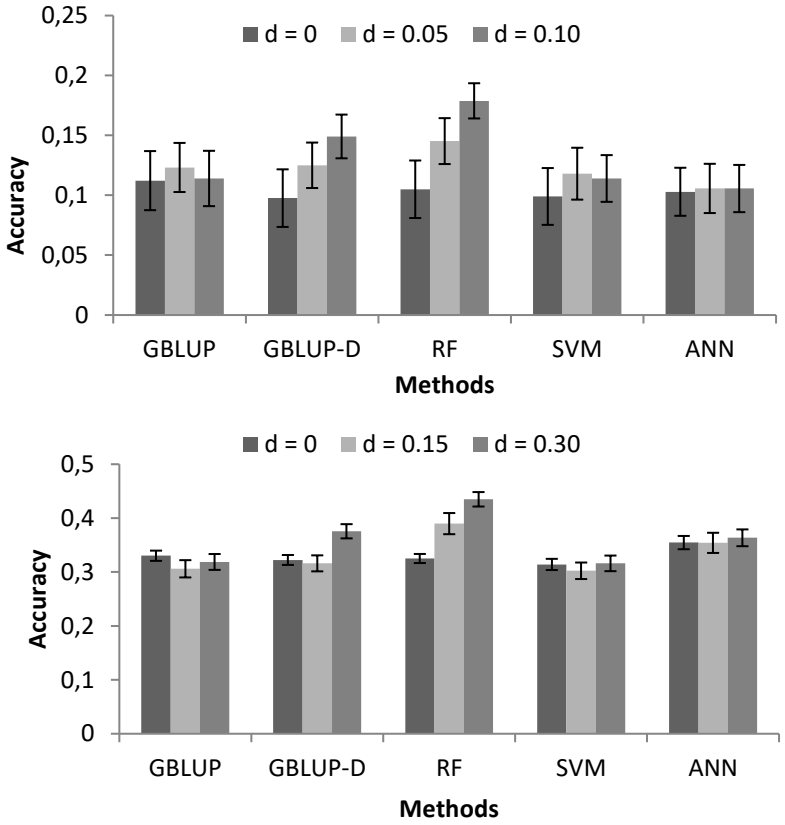


**Figure 3.** Average phenotype prediction accuracy using genomic best-unbiased predictor with or without including dominance effects (GBLUP and GBLUP-D, respectively) and different machine learning methods for simulated complex traits presenting low ($h^2$=0.10; above) or moderate ($h^2$=0.30; below) narrow-sense heritabilities and different dominance contributions to phenotype variation (d).

Another practical use of dominance information in a breeding program would be to explore mating allocation for a specific combining ability of the parents in order to maximize the offspring's productive performance. Previous studies have shown that an extra response is expected by the appropriate design of future mating pairs (Toro and Varona, 2010; Su et. al., 2012). However, predicting an animal performance with the RF method requires the knowledge of its realized genotype, thus, exploring mate allocation techniques may not possible with such an approach.

In practice, there are some limitations on using models considering both additive and non-additive genetic effects for genomic predictions. A common issue to be considered is the high computational cost, since accounting for every possible non-

additive effect rapidly increases the model parameterization. Although models based on genomic relationship matrices have relaxed those constraints, construction and inversion of such matrices are still challenging since both additive and non-additive genomic relationship matrices are dense.

Machine learning methods offer a general framework to cope with non-linear effects. In the present study, we provide insights into the behavior of ML methods when complex traits are affected by both additive and dominance genetic effects. In a general way, our results have pointed out the RF algorithm as an adequate approach to predict the total genetic values on the presence of dominance effects, without imposing any specific genetic structure on markers data. In addition, the RF method presented superior results in comparison to those obtained with the GBLUP approach and was comparable with the equivalent model expanded to account directly for the dominance deviations (GBLUP-D).

### 3.4 Association mapping with RF algorithm

In the RF approach, variable importance measures can be used to identify relevant regions affecting the traits of interest. In regression problems, importance scores are generally based on the average percentage increase in MSE when generating a prediction of the OOB data randomly permuting the $i^{th}$ variable of interest, whereas all others remain unchanged (Breiman, 2001). Since the tree structures generated by the RF algorithm are informative to explore the different types of relationships between the explanatory variables, the importance scores for SNPs can potentially reflect both additive and non-additive contribution to the phenotype prediction (Yao et al., 2013). However posterior analyses are necessary in order to assess the nature of identified effects.

The RF model provided reasonable importance scores for the markers, generally with stronger peaks near to regions presenting the most relevant QTL effects (Figure 4). This is indicative that RF is a promising alternative tool for pre-screening candidate genes, mainly with major effects. High importance scores were also assigned for regions showing a strong contribution to the dominance variance, although additive effects have contributed more effectively to the QTL detection (Figure

4). This is partially due to the genetic architecture of the trait considered, which presents higher variance for the additive than dominance deviations (*h2* = 0.30 and $d^2$ = 0.15), being more probable in real situations. Another reason is that dominance effects are more difficult to detect (Bolormaa et al., 2015).

Our results are in agreement with those from Waldmann (2016), using a simulated dataset, this author reported that RF detected all non-additive effects (both dominance and epistasis effects), although they were not well-separated from adjacent noise.

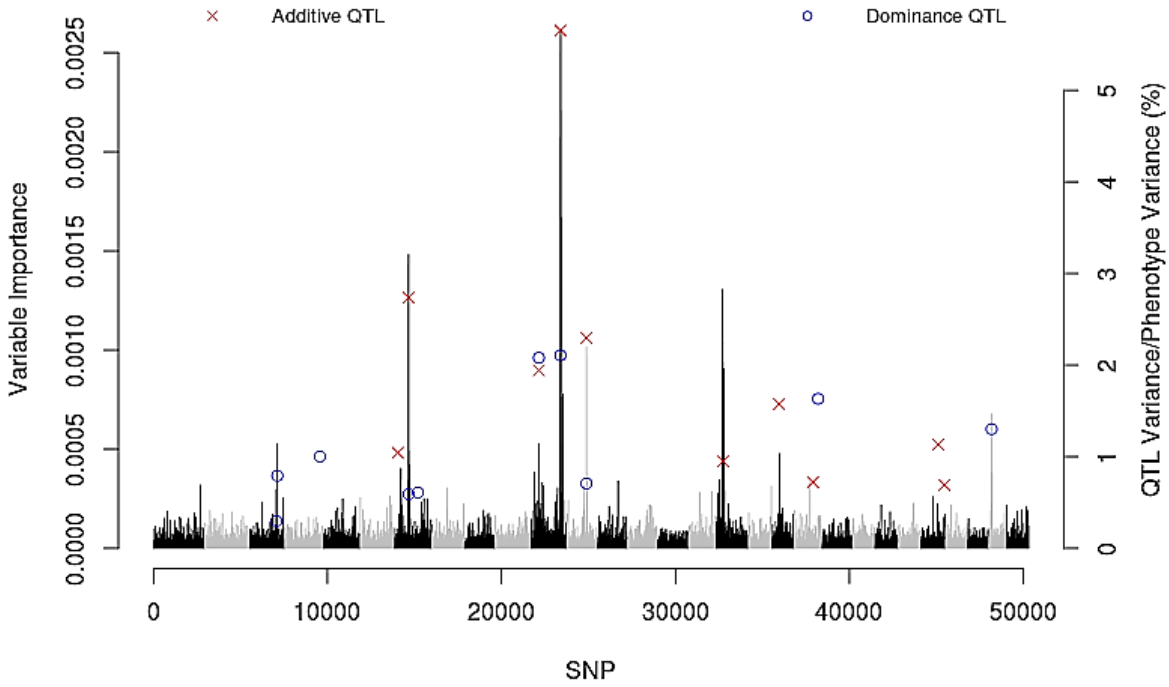


**Figure 4**. Variable importance measures (percent decrease in MSE) for SNPs, real QTL positions and percentage of phenotypic variance related to the simulated QTL presenting the top 10 additives (red x) and/or dominance (blue circle) effects across 29 autosomes.

RF also has been successfully applied to the genome-wide association on real livestock data. Examining the structures of individual trees within the RF, Yao et al. (2013) identified single nucleotide polymorphisms potentially presenting additive and epistatic effects associated with residual feed intake in dairy cattle. In a Canchim beef cattle population, the RF approach identified rather plausible genomic regions associated with backfat thickness, providing a set of SNPs explaining approximately 50% of the deregressed estimated breeding values variance (Mokry et al., 2013).

## 4 Conclusions

We have investigated the predictive ability of GBLUP and different machine learning methods in the presence of dominance effects. According to the found results, among machine learning methods, only the random forest method was capable to cover implicitly dominance effects without increasing the number of covariates in the model, providing higher accuracies for the total genetic values as the dominance ratio increases. Nevertheless, whether the interest is to infer directly about dominance effects, GBLUP-D could be a more suitable method.

## 5 References

Aliloo H, Pryce JE, González-Récio O, Cocks BG, Hayes BJ (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. **Genetics Selection Evolution** 48[8]:1-11.

Beam AL, Motsinger-Reif A, Doyle J (2014) Bayesian neural networks for detecting epistasis in genetic association studies. **BMC Bioinformatics** 15[368]:1-12.

Bolormaa S, Pryce JE, Zhang Y, Reverter A, Barendse W, Hayes BJ, Goddard ME (2015). Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. **Genetics Selection Evolution** 47[26]:1-12.

Breiman L (2001) Random Forests. **Machine Learning** 45:5–32.

de Almeida Filho JE, Guimarães JFR, Silva FF, de Resende MDV, Muñoz P, Kirst M, Resende Jr MFR (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**. 117:33-41.

Eheret A, Hochstuhl D, Gianola D, Thaller G (2015) Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. **Genetics Selection Evolution** 47[22]:1-9.

Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling C, Götz K (2014) Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. **Genetics Selection Evolution** 46[40]:1-10.

Espigolan R, Baldi F, Boligon AA, Souza FRP, Gordo DGM, Tonussi RL, et al. (2013) Study of whole-genome linkage disequilibrium in Nellore cattle. **BMC Genomics**.14:305.

Fernandes Junior GA, Rosa GJM, Valente BD, Carvalheiro R, Baldi F, Garcia DA, Gordo DGM, Espigolan R, Takada L, Tonussi RL, de Andrade WBF, Magalhães AFB,

Chardulo LAL, Tonhati H, Albuquerque LG (2016) Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genet Selection Evolution** 48[7]:1-8.

Falconer DS, Mackay TFC (1996) **Introduction to quantitative genetics**. 4th ed. Essex, UK: Longman.

Fuerst C, Sölkner J (1994) Additive and nonadditive genetic variances for milk yield, fertility, and lifetime performance traits of dairy cattle. **Journal of Dairy Science** 77:1114–1125.

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A (2016) Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science** 57:229-236.

Gallardo JA, Lhorente JP, Neira R (2010) The consequences of including non-additive effects on the genetic evaluation of harvest body weight in Coho salmon (*Oncorhynchus kisutch*). **Genetics Selection Evolution** 42[19]:1-8.

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. **Genetics Selection Evolution** 43.

González-Recio O, Rosa GJM, Gianola D (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science** 116:217-231.

Goddard M (2009) Genomic selection: prediction of accuracy and maximization of long term response. **Genetica** 136:245–257.

Goddard ME, Hayes BJ, Meuwissen THE (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of Animal Breeding and Genetics** 128:409–21.

Hastie TJ, Tibshirani R, Friedman J (2009). The elements of statistical learning. New York: **Springer**.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009). Invited review. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science** 92:433–43.

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **Genes Genomes Genetics** 4[6]:1027-1046.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab - An S4 Package for Kernel Methods. **Journal of Statistical Software** 11:1–20.

Liaw A, Wiener M (2002) Classification and regression by randomForest. **R News** 2:18-22.

Lawrence S, Giles CL, Tsoi AC (1997) Lessons in neural network training: Overfitting may be harder than expected, pp. 540–545 in Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97. AAAI Press, Menlo Park, California.

Lu D, Sargolzaei M, Kelly M, Li C, Voort GV, Wang Z, Plastow G, Moore S, Miller SP (2012) Linkage Disequilibrium in Angus, Charolais and Crossed beef cattle. **Frontiers Genetics** 152[3]:1-10.

Martini JWR, Gao N, Cardoso DF, Wimmer V, Erbe M, Cantet RJC, Simianer H (2017) Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). **BMC Bioinformatics** 18:3.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157:1819–29.

Meuwissen THE, Hayes BJ, Goddard ME (2013) Accelerating improvement of livestock with genomic selection. **Annual Review of Animal Bioscience** 1:221–37.

Mokry FB, Higa RH, Mudadu MA, Lima AO, Meirelles SLC, Silva MVGB, Cardoso FF, De Oliveira MM, Urbinati I, Niciura SCM, Tullio RR, De Alencar MM, Regitano LCA (2013) Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**. v.14, n.47.

Nagy I, Gorjanc G, Curik I, Farkas J, Kiszlinger H, Szendro ZS (2013) The contribution of dominance and inbreeding depression in estimating variance components for litter size in Pannon White rabbits. **Journal of Animal Breeding Genetics** 130, 303–311.

Nishio M, Satoh M (2014) Including dominance effects in the genomic BLUP method for genomic evaluation. **Plos One**. 9[1]:1-6.

Ogutu JO, Piepho H, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. **BMC Proceedings**. 5[11]:1-5.

Okut H, Wu X, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D (2013) Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. **Genetics Selection Evolution** 45:1-13.

Pérez-Rodriguez P, Gianola D (2013) Brnn: brnn (Bayesian regularization for feed-forward neural networks). Available at: <http://CRAN.R-project.org/package=brnn>. Access on: 12 May 2018.

Pérez P, De Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. **Genetics** 198:483–95.

Pérez O'Brien AM, Mészáros G, Utsunomiya YT, Sonstegard TS, Garcia JF, Tassell CPV, et al (2014) Linkage disequilibrium levels in Bos indicus and Bos taurus cattle

using medium and high-density SNP chip data and different minor allele frequency distributions. **Livestock Science**166:121–32

Rodriguez-Almeida FA, Van Vleck LD, Willham RL, Northcutti SL (1995) Estimation of non-additive genetic variances in three synthetic lines of beef cattle using an animal model. **Journal of Dairy Science** 73:1002–11.

Sargolzaei M, Schenkel FS (2009) QMSim: a large-scale genome simulator for livestock. **Bioinformatics** 25[5]:680-681.

Su G, Christensen OF, Ostersen T, Henryon M, Lund MS (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. **Plos One** 7:e45293

Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution** 42[33]:1-9.

Van Den Berg I, Fritz S, Boichard D (2013) QTL fine mapping with Bayes C(π): a simulation study. **Genetics Selection Evolution** 45[1]:1-11.

Van Tassell CP, Misztal I, Varona L (2000) Method R estimates of additive genetic, dominance genetic, and permanent environmental fraction of variance for yield and health traits of Holsteins. **Journal of Dairy Science** 83:1873–7.

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science** 91: 4414–4423.

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics** 195:1223–30.

Vitezica ZG, Legarra A, Toro M, Varona L (2017). Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. **Genetics**. 206:1297–1307.

Waldmann P (2016) Genome-wide prediction using Bayesian additive regression trees. **Genetics Selection Evolution** 48[42]:1-12.

Wei WH, Hemani G, Haley CS (2014) Detecting epistasis in human complex traits. **Nature Reviews Genetics** 15:722–33.

Yao C, Spurlock DM, Armentano LE, Page Jr CD, Vandehaar MJ, Bickhart DM (2013) Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. **Journal of Dairy Science** 96:6716–6729.

## CHAPTER 3 – Genome-enabled prediction of breeding values for reproductive traits in Nellore cattle using parametric models and machine learning methods

**ABSTRACT –** The aim of this study was to assess the predictive ability of different machine learning methods for genome-enabled prediction of reproductive traits in Nellore cattle. Data of Nellore cattle from commercial herds, born between 1984 and 2015 were used. The studied traits were Age at First Calving (AFC), Scrotal Circumference (SC), Early Pregnancy (EP) and Stayability (STAY). After quality control, the number of genotyped animals and SNP markers available were respectively, 2,342 and 321,419 (AFC), 4,671 and 309,486 (SC), 2,681 and 319,619 (STAY) and, 3,356 and 319,108 (EP). The machine learning methods studied were support vector regression (SVR), Bayesian regularized artificial neural network (BRANN) and random forest (RF). Results from machine learning methods were compared with that obtained using GBLUP and BLASSO parametric models. For the SVR, the influence of different kernel bandwidth parameter values on the model predictive ability was assessed. In the BRANN models, different numbers of neurons in the hidden layer (1 to 4 neurons) were examined to assess the best ANN architecture. Further, two genomic structures were used as input information in the BRANN model, the genomic relationship matrix (G) and the principal components scores matrix (PC). The predictive ability of the studied models was evaluated by a 5-fold cross-validation scheme. The values of the bandwidth parameter that maximized the prediction accuracy of the SVR model were 4.9, 3.7, 2.5 and 2.8 for AFC, SC, EP and STAY, respectively. Using G-matrix, it was observed that increasing the number of neurons (3 to 4) provided slightly better accuracy of prediction than the simplest network architectures (with 1 or 2 neurons), with accuracies ranging from 0.372 to 0.555, 0.256 to 0.268, 0.539 to 0.585, 0.473 to 0.517 for AFC, SC, EP and STAY, respectively. For all traits, prediction accuracies from the BRANN_PC were slightly higher than those observed with the BRANN_G model. The average accuracies were from low to moderate according to the trait and model considered, ranging between 0.555 and 0.625 (AFC), 0.268 and 0.359 (SC), 0.573 and 0.666 (EP) and, 0.517 and 0.618 (STAY). Mean-squared error (MSE) varied between 12455.4 and 13016.2 for AFC, 6.32 and 7.76 for SC, 0.111 and 0.154 for EP, and, 0.011 and 0.016 for STAY. The SVR provided slightly better accuracies than the parametric models for all traits, increasing the prediction accuracy for AFC around 5.1% and 3.7% compared to GBLUP and BLASSO models, respectively, and 7.2% for SC, 3.4% for EP and 5% for STAY comparing to both GBLUP and BLASSO. In contrast, the RF, BRANN_G and BRANN_PC models did not present competitive predictive ability compared to the benchmark approaches. Our results indicate that the support vector regression is a suitable method for the prediction of genomic breeding values for reproductive traits in Nellore cattle, presenting better predictive ability and computational time efficiency than the studied parametric approaches. Further, the optimal kernel bandwidth parameter in the SVR model was trait-dependent, thus, the correct pre-definition of this parameter in the training phase is advisable.

**Keywords:** artificial neural network, fertility traits, genomic selection, random forest, support vector regression

## 1 Introduction

Reproductive efficiency-related traits have a paramount role in the profitability of beef cattle production systems under tropical environments (Brumati et al., 2011). Nonetheless, pedigree-based genetic improvement of fertility in beef cattle is challenging, among other reasons, due to the low heritability estimates, sex-limited expression and late phenotype measuring (Boligon et al., 2010; Johnston, 2014; Biscarini et al., 2015). Consequently, heifers' reproductive performance remains little explored as selection criteria in beef cattle breeding schemes (Johnston 2014).

The use of high-throughput genomic technologies in animal breeding has provided new opportunities for overcoming some of the aforementioned drawbacks. Genomic selection (GS) has become a promising approach for accelerating genetic improvement in breeding schemes through gains in the accuracy of breeding values for young animals with no available records and the decrease in generation intervals (Meuwissen et al., 2013). Thus, GS provides a feasible alternative to the traditional pedigree-based methodology for reproductive performance improvement in beef cattle populations.

Reproductive efficiency-related traits are potentially influenced by several genes with smalls effects, albeit non-additive genetic effects may have significative importance in cattle populations (Wall et al., 2005; Palucci et al., 2007). Conversely, most of the popular genome-enabled prediction methodologies, such as genomic BLUP and Bayesian regressions, assume only additive inheritance, while ignoring possible complex nonlinear associations between markers and phenotypes (*e.g.* dominance, epistasis, genotype by environment interaction). Incorporating such complex gene actions may enhance the model predictive ability (Gianola et al., 2011). However, due to a vast increase in model parameterization and prohibitive computational costs, parametric methods generally offer limited flexibility for dealing with non-linear effects in high dimensional genomic data.

In this regard, there has been a growing interest over recent years in using machine learning (ML) methods as an alternative approach to the standard parametric models for genome-enabled prediction of complex traits (Okut et al., 2013; Eheret et al., 2015; Naderi et al., 2016; Li et al., 2018). ML methods are capable to capture

hidden relationships between genotypes and phenotypes in an adaptative manner, without imposing any specific model. This implies that no prior assumptions regarding the underlying genetic architecture of the trait of interest are required. These appealing features provide the ML methods higher flexibility to cope with non-linear relationships on high-dimensional genomic data (Gonzalez-Récio et al., 2014).

Some popular ML methodologies are the artificial neural networks (ANN), support vector machines (SVM) and random forest (RF). Different simulation studies support that these methods provide similar prediction accuracy compared to linear parametric models under purely additive scenarios (Ghafouri-Kesbi, et al., 2016) and better predictive accuracy for traits affected by non-additive effects (Long et al, 2011a; Howard et al., 2014). Nevertheless, there are few empirical applications of ML methods to genomic data analysis in beef cattle populations. Therefore, the aim of this study was to assess the predictive ability of different machine learning for genome-enabled prediction of reproductive traits in Nellore cattle.

## 2 Material and methods

### 2.1 Phenotypic data and pedigree-based analysis

Data of Nellore cattle from commercial herds located in the southeast, midwest and northeast regions of Brazil, born between 1984 and 2015, were used in this study. The farms from which data were collected are part of the DeltaGen®, Paint® (CRV Lagoa) and CIA. de Melhoramento® breeding programs that integrate the Aliança Nellore database. The studied traits were age at first calving (AFC), scrotal circumference (SC), early pregnancy (EP) and stayability (STAY). The AFC was defined as the difference in days between the date of first calving and the dam birth date. SC was measured at yearling in centimeters. The EP was defined as a binary trait, attributing a value of 1 (failure) for heifers that calved after 31 months of age and 2 (success) for heifers that calved before 31 months of age, given that the heifer has been challenged early, between 14 and 18 months of age. Similarly, STAY was also considered as a binary trait, in which the value 2 (success) was assigned to dams that remained in the herd for at least 65 months with a minimum of three successful calvings, otherwise, the value 1 was assigned to represent failure. This criterion was

adopted based on the minimal number of calves and time frame needed to cover the breeding and rebreeding costs with the cows (Van Melis et al., 2007).

Contemporary groups (CG) for AFC and SC were defined by the combination of the herd, year and season of birth, and management group at weaning and yearling. For EP the CG was defined as the herd, year and season of birth of the cow, whereas for STAY the CG considered the herd, year and season of birth, and current herd of the cow. Animals with records for AFC and SC outside of 3.5 standard deviations (SD) from the CG overall mean were removed from the database. In the binary traits, CG showing no variability were removed. Further, for all traits, CG with fewer than four observations were removed from the final database. Table 1 summarizes the descriptive statistics and final data structure used in the variance component analysis for the studied traits.

**Table 1.** Descriptive statistics, number of sires ($N_{Sires}$), dams ($N_{Dams}$) and contemporary groups (CG) for age at first calving (AFC), scrotal circumference (SC), early pregnancy (EP) and stayability (STAY) measured in Nellore cattle.

| Trait | N[1] | Mean | Min[2] | Max[3] | $N_{Sires}$ | $N_{Dams}$ | CG |
|---|---|---|---|---|---|---|---|
| AFC (days) | 202,059 | 1,037.3 | 625 | 1,275 | 4,078 | 155,300 | 9,980 |
| SC (cm) | 448,473 | 26.33 | 15 | 40 | 6,698 | 314,668 | 16,055 |
| EP (%) | 166,877 | 23.30 | - | - | 3,425 | 126,986 | 737 |
| STAY (%) | 139,980 | 35.00 | - | - | 2,810 | 108,337 | 936 |

[1]N: number of animals with records; [2]Min: minimum; [3]Max: maximum

Prior to the genome-based analyses, a mixed animal model approach was used in order to remove the influence of the environmental effects of the target variables. The variance components were estimated by using a linear animal model for AFC and SC and a threshold animal model for STAY and EP. Single-trait analyses were performed for all traits, the general model can be described as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e},$$

in which **y** is the vector of observed records; **β** is a vector of systematic effects; **a** is the vector of random animal effects, assuming to follow a normal distribution $N(0, A\sigma_a^2)$, where **A** is the numerator relationship matrix and $\sigma_a^2$ is the additive variance; **e** is the vector of random residual effects, assuming to follow a normal distribution $N(0, I\sigma_e^2)$ where **I** is an identity matrix and $\sigma_e^2$ is the residual variance; **X** and **Z** are incidence

matrices relating **y** to **β** and **a**, respectively. The vector of systematic effects included the contemporary groups for all traits and the linear effect of age as a covariate for SC. The number of animals included in the pedigree matrix was 329,297 (AFC), 825,548 (SC), 229,812 (STAY) and 271,596 (EP). Variance components for AFC and SC were estimated using the restricted maximum likelihood (REML) method. The threshold models were implemented by Bayesian inference, in which the observed response in the categorical scale was assumed to be linked to an underlying continuous variable following a normal distribution:

$$\mathbf{U}|\boldsymbol{\theta} \sim \mathbf{N}(\mathbf{W}\boldsymbol{\theta}, \mathbf{I}\boldsymbol{\sigma}_{\mathbf{e}}^{\mathbf{2}}),$$

in which **U** is the vector of response variables in the underlying scale with order *r x 1*; $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{a}')$ is a *r x 1* vector of location parameters, and **W** is an *r x s* known incidence matrix. In the threshold analyses, the residual variance was fixed to 1 for ensuring identifiability in the likelihood function (Gianola and Sorensen, 2002). For the Bayesian models, the total length of the Gibbs chains was 1,100,000, with a burn-in period of 100,000 chains and thinning interval at every 100 iterations. All analyses were performed using BLUPF90 family programs (Misztal et al., 2016).

The results obtained with the single-trait animal model analyses were used to compute the target variables for genomic prediction. For the AFC and SC, the response variables were set as the phenotype adjusted for fixed effects (Y*) whereas for the binary traits (EP and STAY), the expected breeding value (EBV) was adopted as the target variable. Although ML can be extended for discrete traits in a straightforward manner (González-Recio and Forni, 2011; Naderi et al., 2016), this study focuses only on the application of ML methods for genome-enabled regression.

### 2.2 Genotypic data

Genotype data were available for 7,687 Nellore cattle (1,013 bulls, 2,434 dams, and 4,242 progeny), considering all studied traits. The animals were genotyped with the Illumina® BovineHD Beadchip (HD; Illumina, Inc., San Diego, CA, USA) and the GeneSeek® Genomic Profiler Indicus HD (GGP75Ki; Neogen Corporation, Lincoln, NE, USA) panels, with approximately 777,000 and 75,000 SNPs distributed throughout the genome, respectively. The genotypes obtained with GP75Ki (n = 3,570) were imputed to HD panel using the FImpute v2.2 software (Sargolzaei et al., 2014)

considering both genotypic and pedigree data, with an expected accuracy higher than 0.97 (Carvalheiro et al., 2014). The genotype files quality control (QC) was performed by an iterative process using the R software (R Development Core Team, 2011). First, non-autosomal, unmapped and duplicated SNPs were removed from the original genotypic data file. Further, SNP markers with a call rate lower than 0.98, minor allelic frequency (MAF) lower than 0.05 and with a p-value less than $10^{-5}$ for the Hardy-Weinberg equilibrium test were removed from the data. For the samples QC, it was adopted as excluding criterion, animals with a call rate lower than 0.90. The iterative process was performed until no further animals or SNP failed in the QC criteria. After QC, the total number of genotyped samples and SNP markers available in the final data were, respectively, 2,342 and 321,419 (AFC), 4,671 and 309,486 (SC), 2,681 and 319,619 (STAY), and 3,356 and 319,108 (EP). The final data structure used in the genome-based analyses is presented in Table 2.

**Table 2.** Descriptive statistics for the response variables used in the genomic analyses.

| Trait[1] | Type[2] | Genotyped animals | | | Mean (SD[3]) | Min[4] | Max[5] |
|---|---|---|---|---|---|---|---|
| | | Males | Females | Total | | | |
| AFC | Y* | - | 2,342 | 2,342 | -7.35 (113.36) | -303.39 | 352.51 |
| SC | Y* | 4,671 | - | 4,671 | 0.63 (2.59) | -9.87 | 10.92 |
| STAY | EBV | 916 | 1,765 | 2,681 | 0.09 (0.13) | -0.79 | 0.66 |
| EP | EBV | 942 | 2,404 | 3,356 | 0.30 (0.44) | -1.23 | 2.30 |

[1]AFC: Age at first calving (days), SC: Scrotal circumference (cm), EP: Early pregnancy, STAY: Stayability; [2]Y*: phenotype adjusted for the fixed effects, EBV: expected breeding value; [3]SD: standard deviation; [4]Min: minimum; [5]Max: maximum

## 2.3 Genome-enabled prediction models

### *2.3.1 Linear genome-enabled prediction models*

In this study, the genomic best linear unbiased prediction (GBLUP) and Bayesian least absolute shrinkage and selection operator (BLASSO) were employed as benchmark models using the Bayesian generalized linear regression (BGLR) package (Pérez and De Los Campos, 2014). These models are well-documented

methodologies for GS, presenting conceptual differences regarding the prior assumptions assigned to the marker effects.

For GBLUP, the following general model in matrix notation can be assumed:

$$\mathbf{y}^* = \mathbf{1_n}\mu + \mathbf{Zg} + \mathbf{e},$$

in which $\mathbf{y}^*$ is the vector of response variables (Y* or EBV); $\mathbf{1_n}$ is a *n x 1* vector of 1 s; $\mu$ represents the overall mean; **Z** is the incidence matrix relating the animals to the direct genomic breeding values (DGV); **g** is the vector of DGV, assuming to follow a normal distribution $N(0, G\sigma_g^2)$, where **G** is the marker-based genomic relationship matrix and $\sigma_g^2$ is the markers genetic variance; **e** is the vector of random residual effects, assuming to follow a normal distribution $N(0, I\sigma_e^2)$, where **I** is an identity matrix and $\sigma_e^2$ is the residual variance. The G-matrix was calculated according to VanRaden (2008):

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

in which **W** represents the genotypes matrix (coded as 0 for AA, 1 for AB or 2 for BB) with *n x m* dimension (n = number of animals and m = number of markers); **P** is a matrix with the expected frequencies for the second allele at each locus, calculated as $2p_j$ and, $p_j$ is the frequency of the second allele for the *j*$^{th}$ SNP in the population. The allele frequencies were estimated from the available data.

For BLASSO, the general model has the following form:

$$\mathbf{y_i^*} = \mu + \sum_{\mathbf{j=1}}^{\mathbf{m}} \mathbf{x_{ij}}\boldsymbol{\alpha}_\mathbf{j} + \mathbf{e_i},$$

where $\boldsymbol{y_i^*}$ correspond to the observed response variable for the *i*$^{th}$ individual; $\mu$ represents the overall mean; $x_{ij}$ is the observed genotype covariate at the *j*$^{th}$ SNP for the *i*$^{th}$ individual; $\alpha_j$ is the allele substitution effect associated with the *j*$^{th}$ SNP and, $e_i \sim N(0, \sigma_e^2)$ is a residual term. The prior distributions assigned to the marker effects have the following form (Park and Casella, 2008):

$$p(\alpha_j \mid \tau_j) \sim N(0, \tau_j^2) \text{ and } p(\tau_j^2 \mid \lambda^2) \sim Exponential\ (\lambda^2), \text{ with } j = 1, \ldots, m$$

It can be shown that the prior marginal distribution assigned for each SNP effect is double exponential (DE), with the regularization parameter $\lambda^2$ following a gamma prior distribution $p(\lambda^2) \sim Gamma\ (r, \delta)$, where *r* and $\delta$ are the rate and shape

parameters, respectively (Park and Casella, 2008). The DE prior distribution has a higher mass centered at zero and ticker tails compared to the normal distribution, which implies a stronger shrinkage of markers with small effects (De Los Campos et al., 2013). The posterior inferences for the BLASSO model were obtained via Gibbs sampling algorithm. The model hyper-parameters were set to default so that the regression model reflect *a priori* 50% of the response variables variance.

*2.3.2 Machine learning methods*

2.3.2.1 Support vector regression (SVR)

Support vector machines (SVM), developed by Vapnik (1995), is originally a supervised learning technique for solving binary classification problems pertaining to the general category of kernel methods. Support vector regression (SVR) is a special case of SVM for dealing with quantitative responses and the general model formulation is given by (Hastie et al., 2009):

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{b} + \langle \mathbf{w}^\mathbf{T} \phi(\mathbf{x}_\mathbf{i}) \rangle,$$

in which $\mathbf{w}^\mathbf{T}$ is the vector of unknown regression weights, $\phi(\mathrm{x_i})$ represents a linear or non-linear mapping of the input vector, and **b** is the model bias. Commonly, in the SVR, the optimization problem is given by minimizing the following restricted loss function (Awad and Khanna, 2015):

$$\frac{\mathbf{1}}{\mathbf{2}}\|\mathbf{w}\|^\mathbf{2} + \mathbf{C}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\boldsymbol{\xi}_\mathbf{i} + \boldsymbol{\xi}_\mathbf{i}^{\,*}),$$

$$\text{subject to:} \begin{cases} \mathbf{w}^\mathbf{T}\phi(\mathbf{x}_\mathbf{i}) + \mathbf{b} - \mathbf{y}_\mathbf{i} \leq \boldsymbol{\varepsilon} + \boldsymbol{\xi}_\mathbf{i} \ \forall_\mathbf{i} \\ \mathbf{y}_\mathbf{i} - \mathbf{w}^\mathbf{T}\phi(\mathbf{x}_\mathbf{i}) + \mathbf{b} \leq \boldsymbol{\varepsilon} + \boldsymbol{\xi}_\mathbf{i}^{\,*} \ \forall_\mathbf{i}, \\ \boldsymbol{\xi}_\mathbf{i}, \boldsymbol{\xi}_\mathbf{i}^{\,*} \geq \mathbf{0} \end{cases}$$

in which, C is a positive regularization parameter that controls the trade-off between model complexity and training errors, and $\xi_i$ and $\xi_i^{\,*}$ are slack variables attributed to errors larger or below than a given constant ε. This is equivalent to dealing with the so-called ε-insensitive loss function ($L_\xi$), represented as:

$$\mathbf{L}_\boldsymbol{\xi} = \begin{cases} \mathbf{0} & \mathbf{if}\ |\boldsymbol{\xi}| < \boldsymbol{\varepsilon} \\ |\boldsymbol{\xi}| - \boldsymbol{\varepsilon} & \mathbf{otherwise} \end{cases},$$

where $\boldsymbol{\xi} = \mathbf{y}_\mathbf{i} - \hat{\mathbf{y}}_\mathbf{i}$. In the "ε-insensitive" loss function, the absolute errors smaller than a predefined constant ε are ignored, while the errors larger than ε are penalized.

Further, introducing Lagrange multipliers, the optimization task can be conveniently expressed as a quadratic programming problem for which the final solutions assumes the form (Hastie et al., 2009):

$$\mathbf{f(x)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} (\boldsymbol{\alpha}_\mathbf{i}^* - \boldsymbol{\alpha}_\mathbf{i})\, \mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) + \mathbf{b},$$

where $\boldsymbol{\alpha}_\mathbf{i}^*$ and $\boldsymbol{\alpha}_\mathbf{i}$ are positive weights associated with each observation, and $\mathbf{K(x_i, x_j)} = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle_{\mathcal{H}}$ is an *N x N* symmetric and positive definite matrix. Note that operating on the data only in terms of the inner products allows representing the feature space without explicitly to compute $\phi(\mathbf{x})$, requiring only the *kernel* matrix*.* In the SVR solution only a subset of weights $(\boldsymbol{\alpha}_\mathbf{i}^* - \boldsymbol{\alpha}_\mathbf{i})$ are different from zero, and the associated observations with nonzero coefficients are called support vectors.

The kernel used in the SVR model was the scaled Gaussian Radial Basis Function (RBF), which has the following form:

$$\mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{exp}\left(-\boldsymbol{\theta} \left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2 / \mathbf{p}\right),$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are the vector of genotypes for animals *i* and *j,* respectively; $\|.\|$ denotes the Euclidean distance between the genotypes; p is the $\mathbf{x_i}$ (or $\mathbf{x_j}$) dimension (number of SNPs) and $\theta$ is the user-defined kernel bandwidth parameter. Therefore, in the SVR model, three parameters need to be tuned, C, ε, and $\theta$. A grid-search procedure under the 3-dimensional parameter space (C, ε, and $\theta$) is commonly applied to find the optimal values.

However, in large datasets the computational burden of this process is prohibitive. Hence, as proposed by Cherkasky and Ma (2004) the parameters C and ε were defined directly from the training data as $C = \max(|\bar{y} - 3\sigma_y|, |\bar{y} + 3\sigma_y|)$ and ε = $3\sigma_y\left(\sqrt{\ln n/n}\right)$, where $\bar{y}$ and $\sigma_y$ are the mean and standard deviation of the response values, respectively, and $n$ is the training sample size. In turn, the kernel bandwidth was defined by a grid-search procedure on the training data with values ranging between 0.1 and 10.

The SVR model was implemented in the *kernlab* package (Karatzoglou et al., 2004). In order to reduce the computational cost in the SVR model, the scaled RBF kernel matrix was built externally with a proper R script and provided as a user-defined kernel matrix in the *ksvm( )* function.

2.3.2.2 Bayesian regularized artificial neural network (BRANN)

Artificial neural networks (ANN) are machine learning methods inspired by the human nervous system, especially in the neurons connectivity idea. There are different ANN types concerning the number of neuron layers, activation function and the number of neurons. In the present study, it was used a single hidden layer feed-forward ANN, formed by an input layer that receives input data (*e.g.* SNPs, genomic relationship coefficients, environmental data), one hidden layer containing the neurons (processing units) and one output layer (see Figure 1).

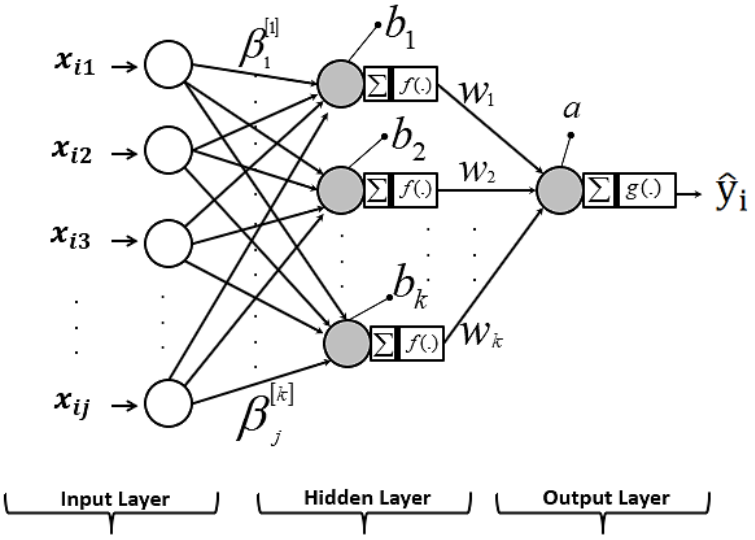


**Figure 1.** Schematic representation of the architecture of a single hidden layer neural network. $x_{ij}$ are the network inputs (here the genomic covariates) of the individual *i*; $\beta_j^{[k]}$ is a network weight for a given hidden layer, where *k* denotes the number of neurons in the hidden layer; $w_k$ is the network weight from the hidden to the output layer; $b_k$ and $a_k$ are the biases in the hidden and output layers; *f(.)* and *g(.)* are activation functions in the hidden and output layers and $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted value for the $i^{th}$ animal. Adapted from Eheret et al. (2015).

Algebraically, such an ANN architecture can be viewed as a two-step regression (Hastie et al., 2009). First, the input covariates $x_{ij}$ of an individual *i* receives a given weight $\beta_j^{[k]}$ connecting to each of *k* (*k = 1, 2, …, k*) neurons in the hidden layer, these weights are combined linearly and summed up with appropriates neuron-specific biases $b_k$ to compute a linear score for the neuron *k* as $\mathbf{z}_\mathbf{i}^{[\mathbf{k}]} = \mathbf{b}_\mathbf{k} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{\beta}_\mathbf{j}^{[\mathbf{k}]} \mathbf{x}_\mathbf{ij}$ (Hastie et al., 2009). The neuron-specific *k* scores are then mapped via some linear or

nonlinear activation function as $\mathbf{f_k(z_i^{[k]})}$. Here, the hyperbolic tangent activation function was adopted: $\mathbf{tahn\,(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}}$. The output layer receives the *k* mapped scores which are linearly combined using the output layer weights $\mathbf{w_k}$, plus the output bias parameter **a**. Finally, the resulting score is mapped by the output activation function $\mathbf{g_k(.)}$, here an identity function, to yield the predicted phenotype of individual *i* (for *i* = 1, …, n):

$$\mathbf{\hat{y}_i = g_k\left[a + \sum_{j=1}^{k} w_k f_k(z_i^{[k]})\right]}.$$

The weights optimization for an ANN can be achieved by minimizing the sum of squared prediction errors (Okut et al., 2013):

$$\mathbf{E_D(D|w, M) = \sum_{j=1}^{n} (\hat{y}_i - y_i)^2},$$

in which **D** denotes the available data, **w** is the vector of network weights, and **M** represents a specific network architecture in terms of the number of neurons, activation functions and number of layers. Gradient-based algorithms are commonly employed in weights optimization. However, as the number of predictor variables increases, the number of parameters to be estimated increases as well, and the ANN model is more probably to introduce overfitting, leading to poor predictive ability. Bayesian regularization aims to alleviate this problem by imposing prior distributions on the model parameters in order to produce shrinkage toward more plausible values (Okut et al., 2013). In the Bayesian regularized ANN (BRANN) model, the cost function receives additional terms and can be considered as a penalized log-likelihood (Gianola et al., 2011):

$$\mathbf{F(\theta) = \beta E_D(D|w, M) + \alpha E_w(w|M)},$$

in which $\mathbf{E_w(w|M)}$ is the sum of squares of the network weights, $\alpha = \frac{1}{2\sigma_\theta^2}$ is a positive regularization parameter for weights and biases and $\beta = \frac{1}{2\sigma_e^2}$ is the regularization parameter for residuals. In the Bayesian perspective, the optimal solution for *w* is given by maximizing its conditional posterior density $p(w|\alpha, \beta, D, M)$, which is equivalent to minimizing the regularized cost function $\mathbf{F(\theta)}$ (Gianola et al., 2011).

Different neural networks, regarding the number of neurons in the hidden layer, varying from 1 up to 4, were examined to assess the ANN architecture with the best predictive ability. Further, two genomic structures were used as input information in the BRANN, the marker-based genomic relationship matrix (G) or the principal components scores matrix (PC), hereinafter called BRNN_G and BRNN_PC models, respectively. For data dimension reduction, while minimizing information losses, the first 500 principal components scores (roughly 90% of the original genotypes matrix variability) were maintained in the PC matrix. The BRNN_G and BRNN_PC models were fitted using the *brnn* package (Pérez-Rodrigues and Gianola, 2013).

2.3.2.3 Random forest (RF)

The RF model uses an ensemble of decision trees that are grown partitioning the original predictor space into a number of simplest regions (Breiman 2001). Each tree uses a bootstrap sample from the training data and randomly selected *mtry* predictors from the full set of *p* predictor variables as splitting candidates in the tree nodes. This two-step randomization procedure decorrelates the trees so that the resulting ensemble of trees is expected to have less variance (Chen and Ishwaran, 2012).

In each tree node, among all *mtry* randomly selected predictors, the algorithm decides the variable and split point with lower risk, using a given loss function as criteria (e.g., squared loss function for regression problems). After grown all individual trees, the RF aggregate the information from the ensemble of trees to compute final predictions as (Hastie et al., 2009):

$$\hat{\mathbf{y}} = \frac{\mathbf{1}}{\mathbf{Ntree}} \sum_{\mathbf{b=1}}^{\mathbf{Ntree}} \mathbf{T}(\mathbf{X}, \boldsymbol{\psi}_{\mathbf{b}}),$$

where $\psi_b$ represents an individual *b* tree architecture in terms of the bootstrapped sample, split variables, cut point at each node and terminal node values. For unobserved values, the prediction is obtained by passing down the predictor variables in the flowchart of each tree and the corresponding estimate at the terminal node is assigned as the predicted value. Predictions of each tree in the RF are averaged to compute the final prediction for unobserved data (James et al., 2013).

In previous analyses, there was no observed dramatic change in the predictive ability by using different values for the RF parameters (data not shown). Therefore, it was decided to set the RF parameters as *ntree* = 1000 (number of trees to grow) and *mtry* = $\sqrt{p}$ (number of SNPs selected at each tree node), and *nodesize* as default for all traits. The RF model was implemented in the R package *randomForest* (Liaw and Wiener, 2002).

### 2.4 Cross-validation and comparison criteria

The predictive ability of the studied models was evaluated by a 5-fold cross-validation scheme, in which the original data sets were randomly divided into five groups with approximately equal size. The genotypes and phenotypes from four groups were used as training data for fitting the models, whereas the left-out group was used as a validation set for testing the model predictive ability. Therefore, the studied models were fitted five times, treating a different group as a validation set at each round, for which the response variable was omitted, and the remaining groups as the training set. The adopted criteria for comparing the predictive ability among models were the prediction accuracy (ACC) derived by the correlation between observed (Y* or EBV) and predicted variables (GEBV), and the mean squared error (MSE). In the traits for which the EBV was set as the target variable (STAY and EP), the ACC was considered as the simple linear correlation between observed and predicted values, $r(EBV, GEBV)$. On the other hand, for the traits where the adjusted phenotype (Y*) was set as the target variable (AFC and SC), the ACC was defined as $r(Y^*, GEBV)/h$, in which *h* is the square root of the estimated heritability of the trait.

## 3 Results and discussion

### 3.1 Heritability estimates

Low heritability estimates were found for AFC, and STAY, whereas moderate to high heritability values were estimated for EP and SC (Table 3). In general, these findings are in agreement with previous results reported in literature for beef cattle populations, which ranged between 0.10 and 0.21 for AFC (Boligon et al., 2010; Boligon and Albuquerque, 2011; Cavani et al., 2015; Terakado et al., 2015; Garcia et

al., 2016; Claus et al., 2017), 0.37 and 0.42 for SC (Van Melis et al., 2010; Terakado et al., 2015; Irano et al., 2016), 0.30 and 0.45 for EP (Boligon and Albuquerque, 2011; Irano et al., 2016) and between 0.10 and 0.11 for STAY (Van Melis et al., 2010; Cavani et al., 2015; Teixeira et al., 2017).

The low heritability estimates for AFC and STAY indicate that these traits are highly influenced by effects not included in the models (*e.g.,* feed nutritional apport, management conditions and, non-additive effects). Generally, the farm management decisions impose limitations on the AFC and STAY expression, which inflates the residual variance, decreasing the heritability estimates. Therefore, individual selection for AFC and STAY is expected to provide small genetic gain. On the other hand, heritability estimates for SC and EP suggest that a substantial proportion of the phenotypic variation can be explained by the additive genetic variance.

**Table 3.** Additive ($\boldsymbol{\sigma_a^2}$) and residual ($\boldsymbol{\sigma_e^2}$) variance components and heritability ($\boldsymbol{h^2}$) estimates for reproductive traits in Nellore cattle.

| Trait | $\boldsymbol{\sigma_a^2}$ | $\boldsymbol{\sigma_e^2}$ | $\boldsymbol{h^2}$ **(S.E)**[1] |
|---|---|---|---|
| Age at first calving (days) | 476.50 | 5,179 | 0.08 (0.005) |
| Scrotal circumference (cm) | 2.84 | 3.71 | 0.43 (0.007) |
| Early pregnancy (%) | 0.43 | 1.00 | 0.30 (0.012) |
| Stayability (%) | 0.10 | 1.00 | 0.09 (0.009) |

[1]S.E: heritability standard error

### 3.2 Influence of the bandwidth parameter in the SVR prediction accuracy

The values of the bandwidth parameter (θ) that maximized prediction accuracy of the SVR were 4.9, 3.7, 2.5 and 2.8 for AFC, SC, EP and STAY, respectively. For AFC, small values of θ tend to cause more instability in the SVR model accuracy than large values (Figure 2). Within the range of 2.2 and 5.8 for θ, the SVR model presented relatively stable prediction accuracies for SC, showing less influence of the bandwidth parameter on this trait. Prediction accuracies for EP and STAY presented similar patterns throughout different values examined for θ, with prediction accuracies falling drastically for $\theta > 4$.

The results in our study suggest that there are no global optimal values for the bandwidth parameter, which are clearly trait-dependent. However, there are empirical pieces of evidence of poor prediction accuracy when adopting extreme values for θ. Similar results were previously reported for kernel-based genomic prediction models for milk traits and daughter pregnancy rate in a Jersey population (De Los Campos et al., 2010). Further, it was noticed that the RBF kernel presented a higher risk of overfitting when considering the EBV as the target variable (EP and STAY), probably due to the expected additive nature underlying the response variables. In this case, a linear kernel would be a reasonable choice.

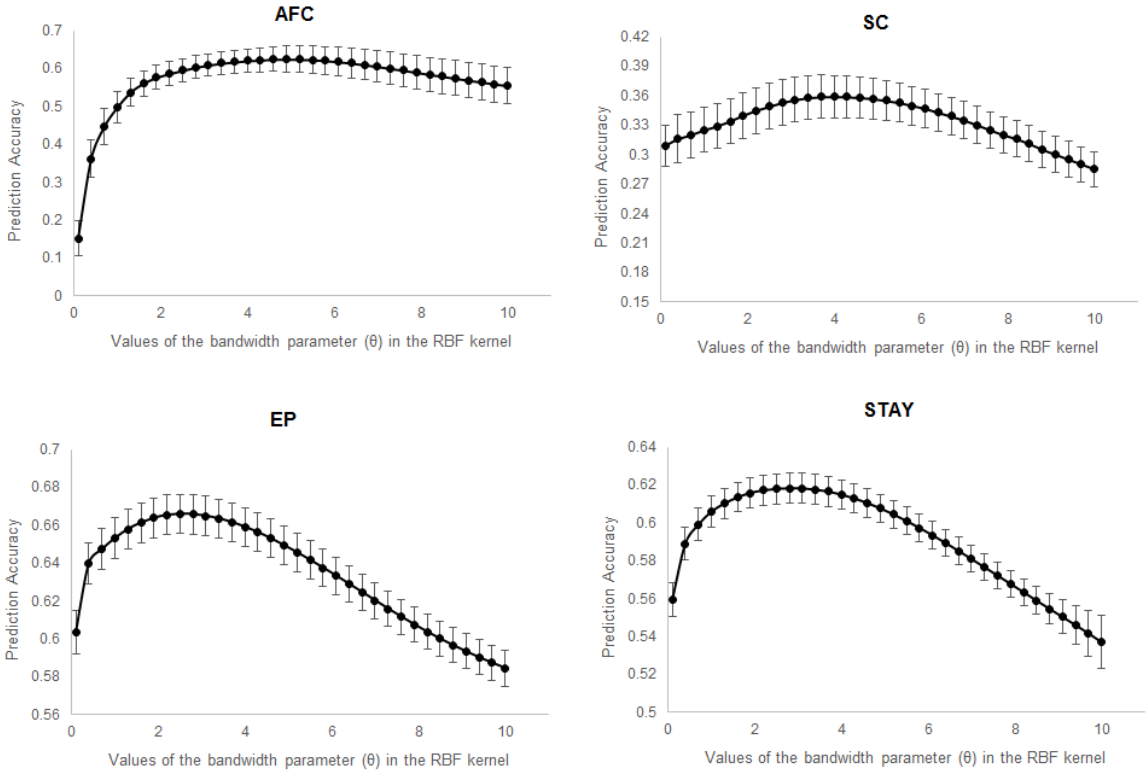


**Figure 2.** Accuracy of prediction (standard errors) in the SVM model according to different bandwidth parameters used in the RBF Kernel.

As observed in the SVR, the appropriate choice of the bandwidth parameter in the Gaussian RBF kernel plays a crucial role in the model prediction accuracy. This parameter defines how training data points influence the number of support vectors (Budiman et al., 2017). Low values for θ define a Gaussian function with high variance, which implies that two data points are considered similar in the feature space even if

they present high Euclidean distance. In turn, the similarity between two data points on the feature space decreases as θ increases as well as the risk of overfitting. Intuitively, the bandwidth parameter determines the similarity structure of the kernel matrix, *e.g.* defining an extremely low value for θ, the kernel matrix off-diagonal elements assume values close to 1. In contrast, too high values for θ, provide a kernel matrix with all off-diagonal elements close to zero, approaching an identity matrix (Figure 3).

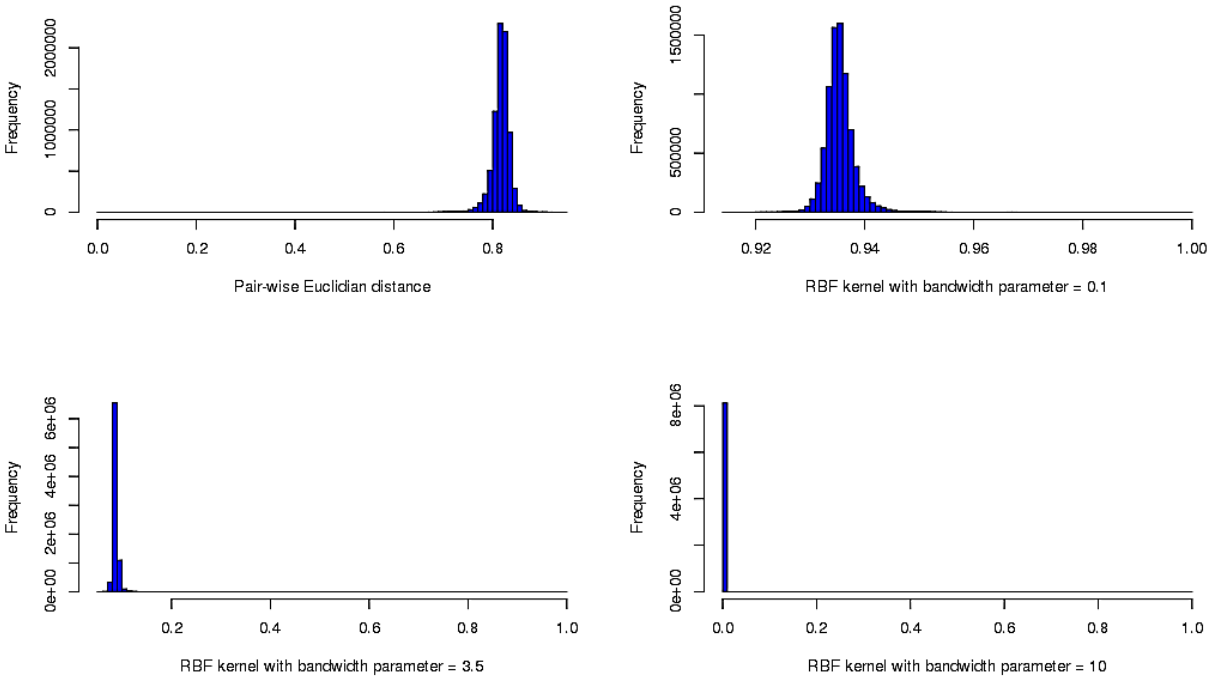


**Figure 3**. Histogram of the scaled pairwise Euclidean distances between training points considering the individuals with genotype and phenotype available for scrotal perimeter (top left). Influence of different values (0.1, 3.5 and 10) of the bandwidth parameter on the similarity between off-diagonal elements in the Kernel matrix (top right, bottom left and bottom right, respectively). Scaled Euclidian distances were computed as $d_{ij} = \|x_i - x_j\|/\sqrt{p}$ and, the Gaussian RBF kernel off-diagonal elements were obtained as $K_{ij} = exp\{-\theta {d_{ij}}^2\}$.

### 3.3 ANN architectures and type of genomic input

The observed differences in the prediction accuracy for Bayesian regularized artificial neural networks (BRANN), regarding the number of neurons in the hidden layer, were more pronounced using the G matrix than when using the PC matrix as

input information (Figure 4). For BRANN_G, it was observed that increasing the number of neurons (with 3 or 4) provided slightly higher prediction accuracy than the simplest network architectures (with 1 or 2 neurons), ranging from 0.372 to 0.555, 0.256 to 0.268, 0.539 to 0.585, 0.473 to 0.517 for AFC, SC, EP and STAY, respectively (Figure 4). This is probably due to the Bayesian regularization procedure that imposes shrinkage of negligible neuron weights, providing final models with a similar number of effective parameters (Gianola et al., 2011). These results are in agreement with those reported by Eheret et al. (2015), who notice that increasing the number of neurons up to 6 in ANN models with back-propagation, provided better results for genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle.

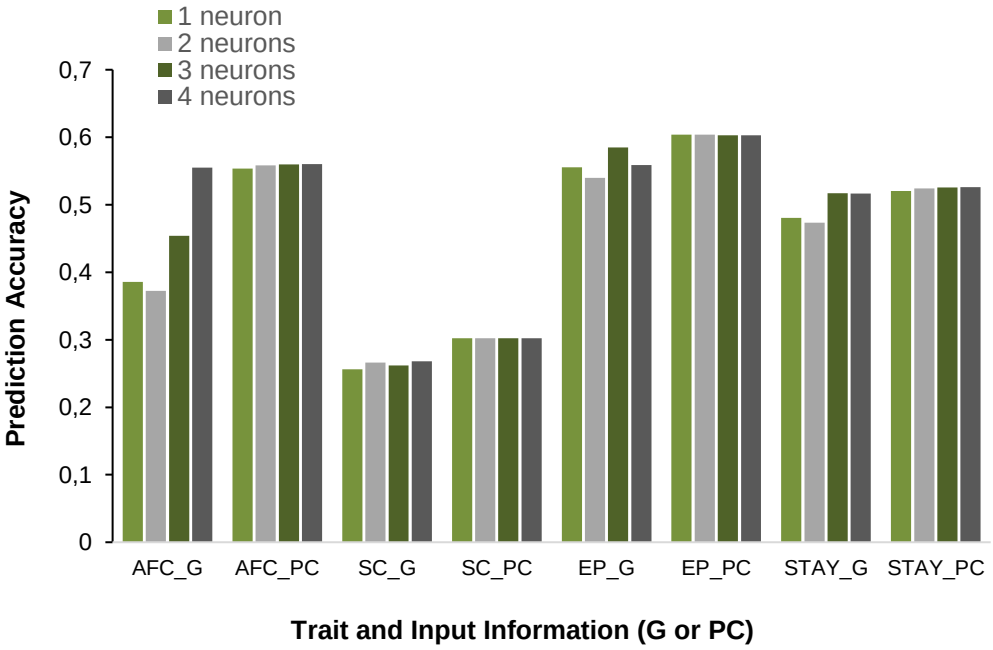


**Figure 4.** Accuracy of prediction for Bayesian neural networks (BRANN) according to the input information (Genomic-based relationship matrix – G or Principal Components Matrix – PC) and number of neurons (1 up to 4) for age at first calving (AFC), scrotal circumference (SC), early pregnancy (EP) and, stayability (STAY).

Alternatively, using the PC matrix as input information, the number of neurons in the hidden layer had little impact on the BRANN prediction accuracies for all traits (Figure 4). This result indicates that the neural network architecture is a secondary issue when the number of predictor variables is smaller than the sample size. Furthermore, the prediction accuracy for BRANN_PC was slightly higher than the

observed in the BRANN_G models for all traits, especially for SC and EP (Figure 4). Similar results were observed with different types of ANN applied for genome-enabled prediction in cattle (Eheret et al., 2015) and pigs (Tussel et al., 2013). Probably, dimensionality reduction and elimination of ambiguous explanatory variables in the PC matrix provides higher numerical stability in the BRANN model than when using the G matrix.

### 3.4 Predictive ability of models and computing time

In general, the average prediction accuracy (ACC) was from low to moderate according to the trait and model considered, ranging between 0.555 and 0.625 for AFC, 0.268 and 0.359 for SC, 0.573 and 0.666 for EP and, 0.517 and 0.618 for STAY. Similarly, the MSE varied remarkably across the studied models, with differences ranging between 12,455.4 and 13,016.2 for AFC, 6.32 and 7.76 for SC, 0.111 and 0.154 for EP, and, between 0.011 and 0.016 for STAY (Table 4).

In spite of presenting the highest heritability estimate, prediction accuracies for SC were lower than those for AFC, EP, and STAY (Table 4). Similar results were observed by Zhang et al. (2014) for female reproductive traits in Brahman cattle, which reported higher prediction accuracy using the GBLUP model for lifetime weaning rate ($h^2$ = 0.13) comparing to traits presenting higher heritability estimates, such as postpartum anestrus interval ($h^2$ = 0.51) and postpartum ovulation status ($h^2$ = 0.62).

These results may be explained by the fact that differences in prediction accuracies depend on many factors other than heritability magnitude, such as the type of response variable (Fernandes Júnior et al., 2016), relatedness between training and validation populations (Lorenz and Smith, 2015), training population design (Naderi et al., 2016), and the trait genetic architecture (Mehrban et al., 2017).

Further, despite the conceptual differences in the *a priori* assumptions for the marker effects, there were no observed major differences in terms of predictive ability between the linear genome-enabled prediction models. Except by the minimal differences found for AFC, the GBLUP and BLASSO models presented practically the same prediction accuracies and MSE for the studied traits (Table 4). This finding is in agreement with the literature, that commonly has been reporting only small differences

in the accuracies between GBLUP and Bayesian regression models in empirical applications (Maltecca et al., 2012; Chen et al., 2014; Mehrban et al., 2017).

Possibly, data-inherent features such as the extent of linkage disequilibrium, effective population size, insufficient sample size for inferring a large number of coefficients (the $n < p$ problem) and the complex nature of the studied traits (*i.e.* a large number of small effect QTLs) impose limitations for observing greater differences in prediction accuracies between models (De Los Campos et al., 2013; Mehrban et al., 2017). Nonetheless, when few QTLs explain a sizable proportion of the additive genetic variance, as in the case of fat percentage in dairy cattle, Bayesian regression models performing variable selection or differential shrinkage of marker effects are expected to present better predictive ability than GBLUP (Colombani et al., 2013; De Los Campos et al., 2013).

The machine learning methods showed remarkable differences in predictive abilities, with the SVR providing the highest prediction accuracies and lowest MSE for all traits (Table 4). The lowest predictive abilities were observed with the BRANN_G model for AFC, SC, and STAY whereas the RF performed worst for EP (Table 4). Furthermore, the SVR provided slightly better accuracies than the parametric models (GBLUP and BLASSO) for all traits, increasing the prediction accuracy for AFC around 5.1% and 3.7% compared to GBLUP and BLASSO models, respectively. Likewise, there was noticed an increase of 7.2% for SC, 3.4% for EP and, 5% for STAY comparing to both GBLUP and BLASSO. Lowest MSE values were also observed for the SVR over linear models (GBLUP and BLASSO) for all studied traits (Table 4). Similarly, Moser et al. (2009) reported that SVR provided the highest accuracies among five methods for predicting genomic breeding values of dairy bulls.

In contrast, the RF, BRANN_G and BRANN_PC models did not present competitive predictive ability compared to the benchmark approaches, presenting lower prediction accuracies and higher MSE than GBLUP and BLASSO for all traits (Table 4), possibly due to an over-fitting in the training phase. Since the MSE takes into account both accuracy and bias of the prediction performance one could infer that the SVR, GBLUP, and BLASSO models were less biased than the RF, BRANN_G, and BRANN_PC models.

**Table 4.** Prediction accuracy (ACC), mean squared error (MSE) and standard deviations (SD) for age at first calving (AFC), scrotal circumference (SC), early pregnancy (EP), and stayability (STAY) obtained with different genome-enabled linear prediction models and machine learning methods in a five-fold cross-validation scheme.

| Trait[1] | Model[2] | Tuning Parameters[3] | ACC[4] (SD) | MSE (SD) |
|---|---|---|---|---|
| AFC (days) | GBLUP | - | 0.593 (0.08) | 12,503.8 (402.9) |
| | BLASSO | - | 0.602 (0.08) | 12,488.9 (402.9) |
| | SVR | $K_\theta$ = 4.9 | **0.625 (0.07)** | **12,455.4 (410.5)** |
| | BRANN_G | hidden neurons = 4 | 0.555 (0.13) | 13,016.2 (1,034.1) |
| | BRANN_PC | hidden neurons = 4 | 0.562 (0.12) | 12,722.5 (546.4) |
| | RF | *mtry* = 567, *ntree* = 1,000 | 0.555 (0.11) | 12,543.3 (456.7) |
| SC (cm) | GBLUP | - | 0.333 (0.05) | 6.37 (0.24) |
| | BLASSO | - | 0.333 (0.05) | 6.38 (0.24) |
| | SVR | $K_\theta$ = 3.7 | **0.359 (0.05)** | **6.32 (0.24)** |
| | BRANN_G | hidden neurons = 4 | 0.268 (0.04) | 7.76 (1.52) |
| | BRANN_PC | hidden neurons = 1 | 0.302 (0.05) | 6.46 (0.27) |
| | RF | *mtry* = 556, *ntree* = 1,000 | 0.285 (0.05) | 6.50 (0.29) |
| EP (%) | GBLUP | - | 0.643 (0.02) | 0.116 (0.009) |
| | BLASSO | - | 0.643 (0.02) | 0.116 (0.009) |
| | SVR | $K_\theta$ = 2.5 | **0.666 (0.02)** | **0.111 (0.009)** |
| | BRANN_G | hidden neurons = 3 | 0.585 (0.06) | 0.154 (0.039) |
| | BRANN_PC | hidden neurons = 2 | 0.604 (0.02) | 0.127 (0.010) |
| | RF | *mtry* =565, *ntree*= 1,000 | 0.573 (0.03) | 0.143 (0.008) |
| STAY (%) | GBLUP | - | 0.587 (0.02) | 0.012 (0.001) |
| | BLASSO | - | 0.587 (0.02) | 0.012 (0.001) |
| | SVR | $K_\theta$ = 2.8 | **0.618 (0.02)** | **0.011 (0.001)** |
| | BRANN_G | hidden neurons = 3 | 0.517 (0.01) | 0.016 (0.001) |
| | BRANN_PC | hidden neurons = 4 | 0.526 (0.03) | 0.013 (0.001) |
| | RF | *mtry* = 565, *ntree*= 1,000 | 0.543 (0.02) | 0.013 (0.001) |

[1]The response variables used in the genome-enabled prediction models were the phenotype adjusted for the fixed effects (Y*) for AFC and SC, and the expected breeding value (EBV) for EP and STAY; [2]GBLUP = Genomic best linear unbiased predictor; BLASSO = Bayesian least absolute shrinkage and selection operator; SVR = Support vector regression; BRANN_G and BRANN_PC = Bayesian regularized artificial neural networks using genomic relationship matrix (G) and principal components matrix (PC) as input variables, respectively; RF = Random forest; [3]$K_\theta$ = value for the bandwidth parameter used in the radial basis Kernel matrix; hidden neurons = number of neurons in the hidden layer; *mtry* = number of random selected predictor variables and *ntree* = number of trees in the RF model; [4]ACC = $r(Y^*, GEBV)/h$ for AFC and SC and $r(EBV, GEBV)$ for EP and STAY, where GEBV is the genome-enabled predicted response and *h* is the square root of the heritability estimate. The highest ACC (SD) and lowest MSE (SD) across methods in the different traits are in bold faces.

Regarding the mean computational time, the BRANN_PC (0.32 min) was the most efficient method (Table 5), obviously, due to the data dimension reduction performed by the principal component analysis. Considering the different datasets, the SVR took approximately 2.05 min per iteration in the cross-validation scheme to complete the analyses, demanding considerably less time than the GBLUP (33.26

min), RF (51.22 min), BLASSO (1039.31 min) and BRANN_G (1436.30 min) models (Table 5). The machine used for implementing the models for EP and SC had an Intel Xeon 3.5 GHz CPU, whereas for AFC and STAY had an Intel Xeon 2.6 GHz CPU. Therefore, comparisons among traits need certain caution.

It was expected that the GBLUP and SVR models were computationally more efficient since these methods are based on the genomic-based relationship and Euclidean distance matrices, respectively, which reduces the number of unknown parameters to be solved to only $n + 1$. Further, it must be highlighted that in the present study, the computational burden for the SVR regarding the grid-search for the model hyper-parameters and the kernel matrix computation were alleviated as proposed by Cherkasky and Ma (2004) and using an R proper script, available upon request. The less time-efficient methods were BLASSO and BRANN_G (Table 5). The computational burden of the Bayesian model is mainly due to the high dimension of the genotype matrix and the MCMC iteration process. For the artificial neural networks, the computational complexity increases as the number of hidden neurons increases, for instance, an ANN with 3,000 input variables, 1 hidden layer, 4 neurons in the hidden layer and 1 neuron in the output layer has a total of 12,009 unknown parameters.

**Table 5.** Average computation time (in minutes) to complete each iteration of the cross-validation scheme according to the different genome-enabled prediction methods.

| Trait[1] | Methods[2] | | | | | |
|---|---|---|---|---|---|---|
| | **GBLUP** | **BLASSO** | **SVM[3]** | **RF** | **BRNN_G** | **BRNN_PC** |
| AFC | 20.46 | 907.15 | 1.32 | 47.72 | 450.77 | 0.12 |
| SC | 58.00 | 1016.00 | 3.78 | 89.60 | 1837.00 | 0.06 |
| EP | 28.14 | 1266.00 | 1.52 | 47.73 | 1458.00 | 0.28 |
| STAY | 26.46 | 968.10 | 1.57 | 24.31 | 1999.42 | 0.81 |
| Average | 33.26 | 1039.31 | 2.05 | 51.22 | 1436.30 | 0.32 |

[1]AFC = age at first calving (days), SC = scrotal circumference (cm), EP = early pregnancy (%) and, STAY = stayability (%); [2]GBLUP = genomic best linear unbiased predictor; BLASSO = Bayesian least absolute shrinkage and selection operator; SVR = support vector regression; BRANN_G and BRANN_PC = Bayesian regularized artificial neural networks using genomic relationship matrix (G) and principal components matrix (PC) as input variables, respectively; RF = random forest; [3]The Kernel matrix building and grid-search procedure were optimized using an proper R script

Empirical results obtained from applications of ML methods for genome-enabled prediction of beef cattle economic importance traits are scarce in the literature.

However, previous studies demonstrate that the performance of ML models may vary across different species and traits, but, generally, present similar or better predictive ability than linear genomic models (Long et al., 2011b; Okut et al., 2013; Tussel et al., 2013; Eheret et al., 2015).

One of the possible reasons for the differences observed in the predictive ability between ML and parametric models relies on the trait genetic architecture. For instance, Long et al. (2011b) notice that the SVR model with different kernels (linear or non-linear) and loss functions had similar performance compared to the BLASSO for predicting sires' PTAs for milk yield. Conversely, in the same study, the non-linear SVR model outperformed the parametric model for wheat yield prediction, showing clear superiority when the trait of interest may be affected by non-additive marker effects (Long et al., 2011b). Such a trend has been supported by simulation studies, which showed that machine learning methods enable more accurate predictions when the interest traits are affected by non-linear effects, especially due to epistasis (Long et al, 2011a; Howard et al., 2014).

In our study, the type of pseudo-phenotype used in the genomic-based analyses varied according to the trait. For EP and STAY the response variable was the EBV, which is expected to reflect only the additive inheritance for all markers. In contrast, for AFC and SC, the response variables were analyzed as pre-adjusted phenotypes for fixed effects, in this case, one could expect to some extent non-linear dependence between the markers and target variables (e.g., dominance or epistasis). Nevertheless, there were no observed different patterns in terms of prediction ability for the parametric and machine learning models across traits, albeit a better generalization capability of RF and BRANN_G methods was expected for AFC and SC.

Other factors such as the size of training set, heritability magnitude, extent of linkage disequilibrium (LD) in the population and the tuning parameters process may impact the performance of ML models as well (González-Recio and Forni, 2011; Ghafouri-Kesbi et al., 2016; Naderi et al., 2016; Sadeghi et al., 2018). Additionally, simulation studies reported that the ML models tend to present worst behavior when the traits are affected by many loci with small effects (González-Recio and Forni, 2011; Ghafouri-Kesbi et al., 2016), which seem to be the case in the present study.

It should be noticed that the ML algorithms discussed here have some differences regarding the data treatment, which may impact directly the model predictive ability. For instance, the SVR is based on single strong learners and, can be considered a special case of the RKHS (Reproducing Kernel Hilbert Space) model. Conversely, the BRANN and RF are based on the idea of implementing many naive functions together. While ANN type models attribute weights for each base learner, the RF aggregates the global commentary between them for providing an averaged final prediction (González-Recio et al., 2014).

The empirical results from this study indicate that the SVR model is suitable for genome-enabled prediction of reproductive traits in Nellore cattle. However, because of the dual formulation approach used in solving the SVR optimization problem, all markers are equally weighted. Therefore, a pre-screening of important markers is expected to improve the SVR prediction accuracy, this needs further investigation. Additionally, in the present study, we have used the RBF kernel, which enables the SVR model handling with complex relationships between the response variable and the markers. Several other kernels (linear or nonlinear) exist and the choice of a suitable mapping function may depend on the nature of the problem considered.

In the ANN-based models, the basis learners are adapted from the data, acting as universal approximators of complex functions (Okut et al., 2013; Eheret et al., 2015). However, in high dimensional data, complex neural network architectures may lead to over-fitting in the training set, leading to poor generalization capability. In the present study, we have employed a Bayesian regularization for attenuating such a problem (Gianola et al., 2011). Nonetheless, the BRANN models, especially using the G matrix, did not overcome the parametric models for predicting the reproductive traits. Possibly, the genetic nature of the studied traits is mainly due to additive effects, for which linear prediction methods are well suited.

One must highlight that ML presents some limitations, such as the lack of interpretability in genetic terms of the obtained solutions, providing a “black box” behavior (González-Recio et al., 2014). Additionally, ML requires an adequate preprocessing of the data and are sensible to the user-defined parameters during the training phase, which increases the computational requirements. Still, machine learning methods offer some appealing attributes under a GS context. An advantage

over parametric models is that most of ML approaches are model-free, which implies that they do not require any prior assumption about the mechanisms governing the problem, enabling to capture possible nonlinear relationships between phenotype and genotypes in a flexible manner (Long et al, 2011a; Long et al, 2011b; Howard et al., 2014). This is particularly interesting in breeding schemes where the prediction of phenotypic performance is of primary importance, for which models incorporating non-additive effects (e.g., epistasis) should perform better than models considering only additive effects.

## 4 Conclusions

Our results indicate that the support vector regression is a suitable method for the prediction of genomic breeding values for reproductive traits in Nellore cattle, presenting better predictive ability and computational time efficiency than the studied parametric approaches. Further, the optimal kernel bandwidth parameter in the SVR model was trait-dependent, thus, the correct pre-definition of this parameter in the training phase is advisable.

## 5 References

Awad M, Khanna R (1 eds.) (2015) Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers: **APRESS**, 268p.

Biscarini F, Nicolazzi EL, Stella A, Boettcher PJ, Gandini G (2015) Challenges and opportunities in genetic improvement of local livestock breeds. **Frontier Genetics** 6:1–7.

Boligon AA, Albuquerque LG, Mercadante MEZ, Lôbo RB (2010) Study of relations among age at first calving, average weight gains and weights from waning to maturity in Nellore cattle. **Revista Brasileira de Zootecnia** 39(4):746-751.

Boligon AA, Albuquerque LG (2011) Genetic parameters and relationships of heifer pregnancy and age at first calving with weight gain, yearling and mature weight in Nelore cattle. **Livestock Science** 41:12-16.

Breiman L (2001) Random Forests. **Machine Learning**, 45:5–32.

Brumatti RC, Ferraz JBS, Eler JP, Formigonni IB (2011) Desenvolvimento de índices de seleção em gado de corte sob enfoque de um modelo bioeconômico. **Archivos de Zootecnia** 60:205–13.

Budiman F, Suhendra A, Agushinta D, Tarigan A (2017) Determination of SVM-RBF Kernel Space Parameter to Optimize Accuracy Value of Indonesian *Batik* Images Classification. **Journal of Computer Science** 13:590-599.

Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien AMP, Sölkner J, McEwan JC, Van Tassell CP, Sonstegard TS, Garcia JF (2014) Accuracy of genotype imputation in Nellore cattle. **Genetics Selection Evolution** 46:69.

Cavani L, Garcia DA, Carreño LOD, Ono RK, Pires MP, Farah MM, Ventura HT, Millen DD, Fonseca R (2015) Estimates of genetic parameters for reproductive traits in Brahman cattle breed. **Journal of Animal Science** 93:3287-3291.

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. **Genomics** 99:323-329.

Chen L, Vinsky M, Li C (2014) Accuracy of predicting genomic breeding values for carcass merit traits in Angus and Charolais beef cattle. **Animal Genetics** 46:55-59.

Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. **Neural Network** 17(1):113–126

Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V, Robert-Granié C (2013) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in the French Holstein and Montbéliarde breeds. **Journal of Dairy Science** 96(1):575-591.

De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Vazquez AI, Allison DB (2010) Semi-parametric Marker-enabled Prediction of Genetic Values using Reproducing Kernel Hilbert Spaces methods. **In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.** Leipzig, Germany.

De Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics** 193:.327-345.

Claus, LAM, Júnior CK, Roso VM, Borges HF, Barcellos JOJ, Ribeiro ELA (2017) Genetic parameters of age at first calving, weight gain, and visual scores in Nelore heifers. **Revista Brasileira de Zootecnia** 46(4):303-308.

Eheret A, Hochstuhl D, Gianola D, Thaller G (2015) Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. **Genetics Selection Evolution** 47[22]:1-9.

Fernandes Júnior GA, Rosa GJM, Valente BD, et al. (2016) Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution** 48:1-8.

Garcia DA, Rosa GJM, Valente BD, Carvalheiro R, Albuquerque LG (2016) Comparison of models for the genetic evaluation of reproductive traits with censored data in Nellore cattle, **Journal of Animal Science** 94(6):2297-2306.

Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A (2016) Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science** 57:229-236.

Gianola D, Sorensen D (2002) Bayesian analyses of linear models. In: Gianola D, Sorensen D, editors, Likelihood, Bayesian, and MCMC methods in quantitative genetics. **Springer-Verlag**, New York, NY. 287–326.

Gianola D, Okut H, Weigel K, Rosa G (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with jersey cows and wheat. **BMC Genetics** 12:1-14.

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. **Genetics Selection Evolution** 43:7.

González-Recio O, Rosa GJM, Gianola D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science** 116:217-231.

Hastie TJ, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, Inference and Prediction. **Springer Series in Statistics**, 764p.

Howard R, Carriquiry AL, Beavis, WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3** 4[6]:1027-1046.

Irano N, de Camargo GMF, Costa RB, Teralkado AP, Magalhães AF, Silva RM, Dias MM, Bignardi AB, Baldi F, Carvalheiro R, de Oliveira HN, de Albuquerque LG (2016) Genome-wide Association Study for Indicator Traits of Sexual Precocity in Nellore Cattle. **Plos One** 11(8):e0159502.

James G, Witten D, Hastie T, Tibishirani R (2013) An Introduction to Statistical Learning: with applications in R. New York: SPRINGER, 426p.

Johnston DJ (2014) Genetic Improvement of Reproduction in Beef Cattle In: **Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, Vancouver**.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab - An S4 Package for Kernel Methods. **Journal of Statistical Software**, 11:1–20.

Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genome Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontier Genetics** 9:1-20.
Liaw A, Wiener M (2002) Classification and regression by randomForest. **R News** 2:18-22.

Long N, Gianola D, Rosa GJM, Weigel KA (2011a). Marker-assisted prediction of non-additive genetic values. **Genetica** 139:843–854.

Long N, Gianola D, Rosa GJM, Weigel KA (2011b) Application of support vector regression to genome-assisted prediction of quantitative traits. **Theoretical and Applied Genetics** 123:1065-1074.

Lorenz AJ, Smith KP (2015) Adding Genetically Distant Individuals to Training Populations Reduces Genomic Prediction Accuracy in Barley. **Crop Science** 55:2657-2667.

Maltecca C, Parker KL, Cassady JP (2012) Application of multiple shrinkage methods to genomic predictions. **Journal of Animal Science** 90:1777-1787.

Mehrban H, Lee DH, Moradi MH, Ilcho C, Naserkheil M, Ibáñez-Escriche N (2017) Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. **Genetics Selection Evolution** 49:1-13.

Meuwissen THE, Hayes BJ, Goddard ME (2013) Accelerating improvement of livestock with genomic selection. **Annual Review of Animal Bioscience** 1:221–237.

Misztal I, Tsuruta S, Lourenco D, Masuda Y, Aguilar I, et al. (2016). Manual for BLUPF90 Family of Programs. University of Georgia, Athens, GA.

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution** 41(56):1-16.

Naderi S, Yin T, König S (2016) Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. **Journal of Dairy Science** 99:7261-7273.

Okut H, Wu X, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D (2013) Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. **Genetics Selection Evolution**. 45:1-13.

Palluci V, Schaeffer LR, Miglior F, Osborne V (2007) Non-additive genetic effects for fertility traits in Canadian Holstein cattle. **Genetics Selection Evolution** 39:181-193.

Park T, Casella G (2008) The Bayesian Lasso. **Journal of the American Statistical Association** 103:681–686.

Pérez-Rodriguez P, Gianola D (2013) brnn: brnn (Bayesian regularization for feed-forward neural networks). Available at: <http://CRAN.R-project.org/package=brnn>, Access on 13 Feb. 2019.

Pérez P, De Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. **Genetics** 198:482-495.

R Development Core Team (2011) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. R Core Team.

Sadeghi S, Rafat SA, Alijani S (2018) Evaluation of imputed genomic data in discrete traits using Random Forest and Bayesian threshold models. **Acta Scientiarum. Animal Sciences** 40.

Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. **BMC Genomics** 15:1-12.

Teixeira DBA, Fernandes Júnior GA, Silva DBS, Costa RB, Takada L, Gordo DGM, Bresolin T, Carvalheiro R, Baldi F, Albuquerque LG (2017) Genomic analysis of stayability in Nellore cattle. **Plos One** 12(6):e0179076.

Terakado APN, Boligon AA, Baldi F, Silva JAIIV, Albuquerque LG (2015) Genetic associations between scrotal circumference and female reproductive traits in Nelore cattle. **Journal of Animal Science** 93:2706-2713.

Tusell L, Pérez-Rodríguez P, Forni S, Wu XL, Gianola D (2013) Genome-enabled methods for predicting litter size in pigs: a comparison. **Animal** 7(11):1739–1749.

Van Melis MH, Eler JP, Oliveira HN, Rosa GJM, Silva JAIIV, Ferraz JBS, Pereira E, (2007) Study of stayability in Nellore cows using a threshold model. **Journal of Animal Science** 85:1780–1786.

Van Melis MH, Eler JP, Rosa GJM, Ferraz JBS, Figueiredo LGG, Mattos EC, Oliveira HN (2010). Additive genetic relationships between scrotal circumference, heifer pregnancy, and stayability in Nellore cattle. **Journal of Animal Science** 88(12), 3809-3813.

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science** 91: 4414–4423.

Vapnik V (1995) The nature of statistical learning theory, 2nd edn. **Springer**, New York.

Wall E, Brotherstone S, Kearney JF, Woolliams JA, Coffey MP (2005) Impact of Nonadditive Genetic Effects in the Estimation of Breeding Values for Fertility and Correlated Traits. **Journal of Dairy Science** 88:376-385.

Zhang YD, Johnston DJ, Bolormaa S, Hawken RJ, Tier B (2014) Genomic selection for female reproduction in Australian tropically adapted beef cattle. **Animal Production Science** 54:16-24.

# CHAPTER 4 – Genome-wide association study for age at first calving in Nellore cattle using the Random Forest approach

**ABSTRACT –** The aim of this study was to perform a genome-wide association study (GWAS) using the Random Forest approach for scanning candidate genes for age at first calving in Nellore cattle. Data from Nellore cattle, born between 1984 and 2015 and raised in different commercial herds, located in the southeast, midwest and northeast regions of Brazil were used. Predicted breeding values were considered as the response variables. The remaining number of animals and SNPs after quality control were 3,369 and 320,940, respectively. The assessed values for the $M_{try}$ parameter (*i.e*. the number of SNPs to search at each node) were 1, $\sqrt{p}$, $0.01p$ and $0.1p$, in which $p$ represents the total number of SNPs. The RF parametrization which produced the lowest mean squared error in the out-of-bag data ($MSE_{OOB}$) was maintained for further analysis. After defining the best RF parameters configuration, we run five independent analyses with different initialization seeds for the algorithm. In order to improve the stability of the GWAS results, the importance scores of each SNP were averaged over the five independent RF analyses to compute the final importance scores. Results pointed out that the out-of-bag prediction error stabilized around 200 trees and the value of 1,000 was retained as a reliable size for the $N_{tree}$. Further, using 10% of the total number of markers provided the lowest prediction error. A total of 118 SNPs associated with age at first calving (AFC) were identified. The relevant SNPs were located over eight autosomes (BTA 3, 5, 10, 11, 18, 21, 25 and, 27). In total, 23 non-overlapping genomic regions embedded 172 candidate genes for AFC. Genomic regions previously associated with fertility and growth traits in Nellore cattle were found in the present study, which reinforces RF effectiveness for pre-screening candidate regions associated with complex traits. The RF-based genome-wide scan and functional analysis highlighted candidate genes with key roles in fertility, including embryo pre-implantation and development, embryonic viability, male germinal cell maturation, and pheromone recognition. Results reported in the present study are expected to enhance the knowledge of biological mechanisms regulating the expression of age at first calving in Nellore cattle.

**Keywords:** beef cattle, candidate genes, ensemble learning, fertility traits, non-parametric methods

## 1. Introduction

Adaptation to tropical environments and resistance to parasites are attributes that make Nellore cattle an important genetic resource for Brazilian pasture-based beef production systems. Notwithstanding, *Bos indicus* breeds generally present lower reproductive efficiency compared to taurine cattle (Abeygunawardena and Dematawewa, 2004; Sartori et al., 2010), which restricts the selection pressure on replacement heifers. It is known that the efficiency of reproductive performance is intimately associated with beef cattle industries' profitability since a large proportion of the production system costs are due to the cow's maintenance in the herd (Malhado et al., 2013). Hence, attaining high fertility rates is a key issue for reducing costs in beef production systems.

Age at first calving (AFC) is one of the most common selection criteria for fertility in beef cattle breeding programs, among other reasons, because it is easily measured and contributes directly to the reduction on generation intervals. The better understanding of genetic mechanisms underlying AFC plays a paramount role in assisting breeding programs to design efficient strategies to enhance fertility rates in Nellore cattle populations. Advances and cost reduction of high-throughput genotyping technologies have popularized genome-wide associations studies (GWAS), which has contributed for revealing several candidate genes for reproductive performance in beef cattle over recent years (Melo et al., 2017; Teixeira et al., 2017; Nascimento et al., 2018). Generally, methodologies employed for scanning genomic regions associated with quantitative traits in livestock involve additive parametrization of explanatory genetic variants by using either genomic-based relationships matrices or Bayesian regression models (Schmid and Bennewitz, 2017).

Alternatively, some studies have been focused on applying machine learning methods (ML) to identify potential causal variants using genome-wide data, especially for human diseases (Szymczak et al., 2009; Goldstein et al., 2012). ML requires minimal or no assumptions about the biological mechanisms governing complex traits, which allows capturing hidden patterns from high dimensional data (Libbrecht and Noble, 2015). This implies that ML offers a general framework for unrevealing potential novel causal variants when the true genetic nature underlying the associations between phenotype and markers are unknown and complex. For this purpose, the

random forest (RF) is one of the most popular learning algorithms. Importantly, RF variable importance measures provide an intuitive and straightforward approach to select and rank relevant predictors (*e.g.* single nucleotide polymorphisms - SNPs), while adaptatively dealing with correlation and interaction among variables (Chen and Ishwaran, 2012; Yao et al., 2013). These appealing features may contribute to enhance the knowledge about biological mechanisms involving complex traits. Nevertheless, applications of the random forest to identify genomic regions for reproductive traits in beef cattle are still scarce. The aim of this study was to perform a GWAS using the Random Forest approach for scanning candidate genes for age at first calving in Nellore cattle.

## 2 Material and methods

### 2.1 Animals and phenotypes

Phenotype and pedigree data used in this study were obtained from the Alliance Nellore database, which integrates information of Nellore cattle raised in different commercial herds, located in the Southeast, Midwest and Northeast regions of Brazil. Animals were born between 1984 and 2015. These herds adopt reproductive managements with an anticipated breeding season occurring between February and April, with approximately 60 days length, in which heifers between 14 and 18 months of age are exposed to reproduction for identifying sexually precocious animals. Heifers that did not conceive in the anticipated breeding season participate along with the other dams in the regular breeding season occurring between November and January.

In this study, the Age at First Calving (AFC) was adopted as a fertility-related trait, obtained as the difference in days between the date of first calving and the dam birth date. The contemporary groups (CG) were formed by animals born in the same herd, year and season, and raised in the same management group at weaning and yearling. In the data filtering, animals with records deviating ± 3.5 standard deviations from the CG mean were excluded from the dataset. Further, CG with less than five observations were not considered. A mixed-effects model approach was used to remove environmental influence from AFC, considering a single-trait linear animal model. The model included contemporary groups as fixed effects and, the additive polygenic effects and residuals as random effects. The number of animals included in

the additive relationship matrix was 329,297. Variance components were estimated by Restricted Maximum Likelihood (REML) using BLUPF90 family programs (Misztal et al., 2016). Predicted breeding values were considered as response variables in posterior genomic analyses.

## 2.2 Genotype file and quality control

The total genotyped population was composed by 8,666 Nellore cattle (1,128 bulls, 2,737 cows and 4,801 progeny) which were initially genotyped with either the Illumina BovineHD panel (HD; 4625 samples) or with the GeneSeek Genomic Profiler Indicus HD (GGP75Ki; 4041 samples), with approximately 777,000 and 75,000 SNPs, respectively, distributed throughout the genome. The lower density panel (GCP75Ki) was imputed to HD using the FImpute v2.2 software (Sargolzaei et al., 2014), considering all genotyped animals and pedigree information, with an expected accuracy higher than 0.97 (Carvalheiro et al., 2014). After imputation procedure, only genotyped samples with available EBV for AFC (1027 bulls and 2342 cows) were kept. Because of the low EBV accuracy, the progeny data were not considered in the genome-wide association analyses. For the genotypes quality control (QC), were discarded SNPs non-autosomal, unmapped or duplicated and SNPs with call rate < 0.98, minor allelic frequency (MAF) < 0.05 and with a p-value for the Hardy-Weinberg equilibrium test lower than $10^{-5}$. Only samples with a call rate higher than 0.90 were maintained in the genotypic data. The genotypes file filtering was performed using the R software (R Development Core Team, 2011). After quality control, the numbers of animals and SNPs retained were 3,369 and 320,940, respectively.

## 2.3 Genome-wide association analysis with Random Forest (RF)

### *2.3.1 RF algorithm description*

The random forest (RF) is a machine learning method that aggregates complementary information from an ensemble of classification or regression trees trained on different bootstrap samples drawn from the original data set (Breiman 2001). Briefly, let $y_{(nx1)}$ be a vector of observations for a given trait and $X_{(n\,x\,p)}$ the markers matrix, with *n* representing the number of available samples and *p* the number of SNPs.

Initially, a bootstrap sample is drawn from this data set and used for training an individual classification or regression tree. At each node of this given tree, a subset of $M_{try}$ variables are drawn randomly from the overall *p* SNPs and evaluated using a recursive binary splitting rule, for which the best predictor variable $X_j$ (with j = 1, 2, 3, …,$M_{try}$) and the threshold value $t_k$ are those which minimize a given loss function. For continuous responses, the squared loss function is commonly adopted. The tree node is partitioned according to the coordinates $\{y|X_j \leq t_k\}$ and $\{y|X_j > t_k\}$ originating two child nodes, which are also partitioned using the same splitting rule (evaluating different $M_{try}$ markers at each node). This process is repeated until the tree reaches terminal nodes with homogenous or near homogenous responses (Chen and Ishwaran, 2012). The predicted outcomes of the tree are the most frequent class (for categorical responses) or the average observation (for continuous responses) at terminal nodes. Finally, many trees are built using $N_{tree}$ different bootstrap samples, following the same steps. The information from the ensemble of trees is aggregate for computing final predictions:

$$\hat{y} = \frac{1}{Ntree}\sum_{b=1}^{Ntree} T(X, \psi_b),$$

where $\psi_b$ represents an individual tree architecture in terms of the bootstrap sample, SNPs selected at each node and terminal node responses.

A particularity of the RF is the out-of-bag data (OOB), which corresponds to the animals not included (roughly 1/3) in the bootstrap sampling for building a specific tree. The OOB can be used as an internal validation set for each tree and the generalization error of the RF model can be computed (James et al., 2013). The mean squared error is generally used as the loss function:

$$MSE_{OOB} = \frac{1}{N_{OOB}}\sum_{i=1}^{N_{OOB}}(y_i - \hat{y}_i)^2,$$

in which $N_{OOB}$ is the number of observations in the OOB samples, $\hat{y}_i$ is the average of predictions for the *i*$^{th}$ animal computed from trees in which this animal was OOB, and $y_i$ is the realized value. The $MSE_{OOB}$ is considered an internal validation of the prediction error and can be used for tuning the RF parameters.

An appealing feature of the RF is that it can provide variable importance measures (VIM) for each explanatory variable. The most frequently used measure is the permutation-based VIM, which can be internally computed for the *j*$^{th}$ SNP as the

average difference between the $MSE_{OOB}$ when the SNP of interest was randomly permuted in the OOB data and the $MSE_{OOB}$ obtained without permutation, considering all trees. SNPs with higher VIM are suggestive of having an association with the phenotype of interest, since permutating a relevant SNP is expected to increase the OOB prediction error (Mokry et al., 2013; Yao et al., 2013). For an SNP that has no association with the response variable, the permutation-based scored is expected to be approximately zero. Similarly, negative importance scores are an indicator that the permutation of the SNP on the OOB data provided lower generalization error; therefore, this SNP does not have importance for prediction. Generally, the absolute value of the most negative permutation based-score is used as a threshold to differentiate SNPs with real or spurious signals (Yao et al., 2013).

### *2.3.2 RF implementation*

The genome-wide association analysis was performed using the *randomForest* package (Liaw and Wiener, 2002) available for the R software (R Development Core Team, 2011). Because of the $MSE_{OOB}$ converged rapidly in previous analyses, the parameter $N_{tree}$ (*i.e.* the number of trees to grow) was fixed to 1000. The assessed values for the $M_{try}$ parameter (*i.e*. the number of SNPs to test at each node) were 1, $\sqrt{p}$, $0.01p$ and $0.1p$, in which *p* represents the total number of SNPs. The *nodesize* parameter (*i.e.* the maximum number of observations at the terminal nodes) was set to default (*nodesize* = 5) in all analyses. The parametrization that produced the lowest final $MSE_{OOB}$ was maintained for further analysis. After defining the best RF parameters configuration, we run five independent analyses with different initialization seeds for the algorithm. In the RF genome-wide association analysis, a standardized importance factor for each SNP was computed by dividing its original permutation-based score ($\%IncMSE_{SNPj}$) by the absolute value of the most negative importance score (Szymczak et al., 2016):

$$f_{SNPj} = \frac{\%IncMSE_{SNPj}}{|min\,\%IncMSE_{SNP}|}$$

In order to improve the stability of the GWAS results, the importance scores of each SNP were averaged over the five independent RF analyses to compute the final

importance scores. For reducing the false-positive discovering probability, we set the threshold $f_{SNPj} \geq 3$ to identify the SNPs with the strongest signals, as suggested by Szymczak et al (2016).

### 2.4 Identification of candidate genes and enrichment analysis

The identification of candidate genes flagged by SNPs previously selected in the RF analysis was performed using the genome data viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/?org=bos-taurus) from the National Center for Biotechnology Information (NCBI), considering the UMD3.1 (Zimin et al., 2009) as the reference map. For gene annotation, it was considered a 500 Kb window (SNP location ± 250 Kb) harboring each SNP with $f_{SNPj} \geq 3$. For SNPs with overlapping windows, it was considered as reference location only the SNP with the highest importance factor ($f_{SNPj}$). In order to provide more insights regarding the biological processes that the annotated genes are involved, the DAVID database (Huang et al., 2009a; Huang et al., 2009b), ClueGo program (Shannon et al., 2003) and Cytoscape plug-in (Bindea et al., 2012) were used for performing functional analysis.

## 3 Results and discussion

In the present study, the expected breeding values (EBVs) of 1027 sires and 2342 dams with available genotypes were used as response variables in the RF-based genome-wide association study. The EBV for both sires and dams showed an approximately normal distribution, with average values of -0.35 ± 18.3 and -4.3 ± 15.2 days, respectively (Figure 1). On the other hand, average EBV accuracies were higher for sires (0.57 ± 0.22) than for dams (0.46 ± 0.09). The adoption of EBVs instead of deregressed proofs (dEBV) as response variables, was due to the low accuracy for the breeding value predictions. In this case, the parental contribution removal would incorporate too much noise during the deregression process. In this scenario, some authors advocate that EBVs would be a reasonable choice for genome-enabled analysis (Morota et al., 2014; Fernandes Junior et al., 2016). Furthermore, preliminary analyses pointed out that the RF model fitted data better when using the EBVs as

response variables rather than dEBVs or fixed effects adjusted phenotypes (data not shown).

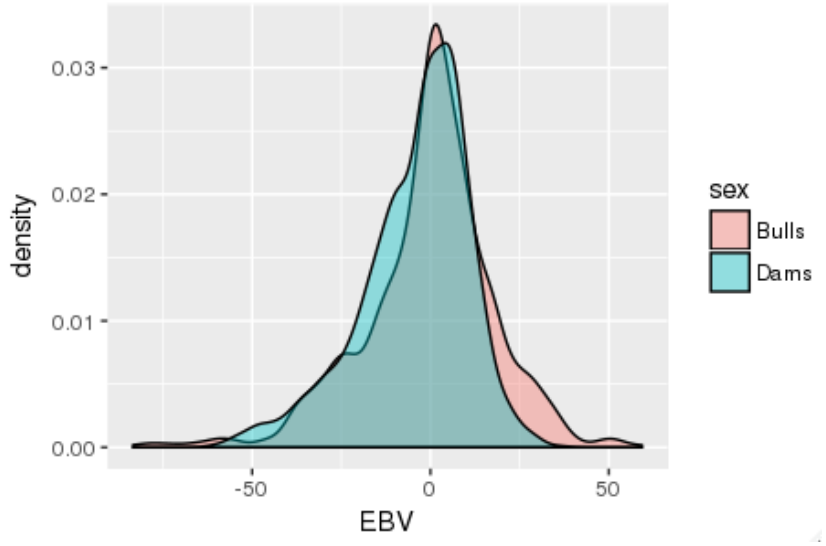


**Figure 1.** Density plot of expected breeding values (EBV) for age at first calving in Nellore cattle, according to the sex category.

The influence of RF parameters on the model predictive performance is presented in Figure 2, it can be seen that the out-of-bag prediction error stabilizes around 200 trees and 1,000 trees were used as a reliable size for the $N_{tree}$ parameter. Among the assessed values for $M_{try}$ (number of SNP randomly analyzed per tree node), the single-marker analysis ($M_{try}$ = 1) produced the worst predictive performance, whereas values $\sqrt{p}$, $0.01p$ and $0.1p$ gave similar results, with $M_{try}$ = $0.1p$ providing slightly lower OOB prediction error (Figure 2). This parameter controls the trade-off between bias and variance, impacting directly in the RF accuracy, with higher $M_{try}$ leading to greater sparsity of variable importance measures (Goldstein et al., 2012). Results pointed that using 10% of the total number of markers was the most appropriate choice, this is in agreement with other genome-enabled studies using human and cattle data (Goldstein et al., 2010; Li et al., 2018). Therefore, the genome-wide analyses were performed using $M_{try}$ = $0.1p$ and $N_{tree}$ = 1,000 for all five RF replicates.

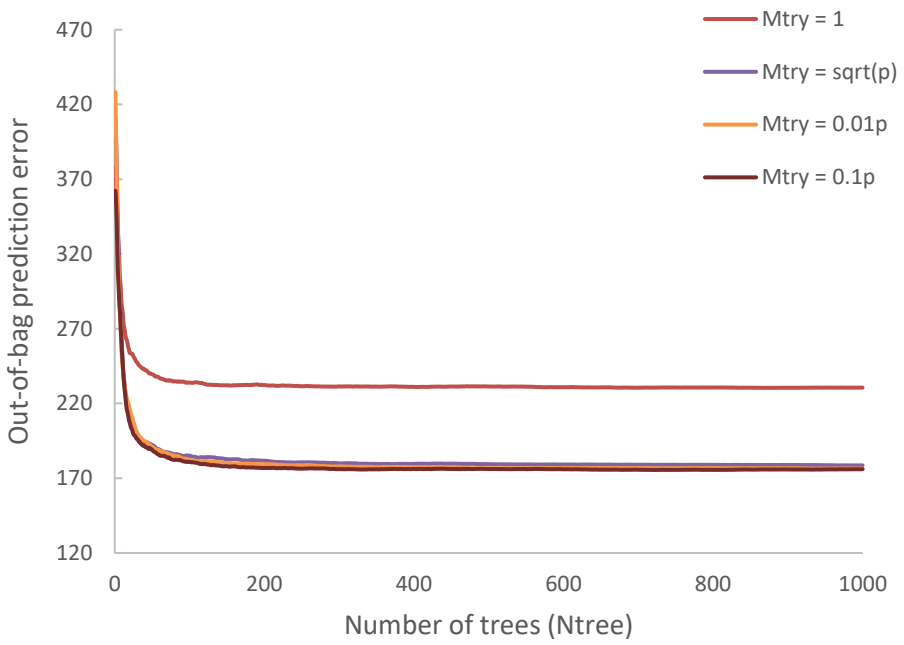


**Figure 2.** Influence of the random forest parameters ($M_{try}$ and $N_{tree}$) in the out-of-bag prediction error for age at first calving in Nellore cattle.

The genome-wide association analysis, using the RF approach, allowed the identification of 118 SNPs associated with age at first calving (AFC). The most important SNPs ($f_{SNPj}$ ≥ 3) were located over eight autosome chromosomes (BTA 3, 5, 10, 11, 18, 21, 25 and, 27) (Figure 3). The genomic regions surrounding the relevant SNPs (250 Kb downstream or upstream) were investigated, in total, 23 non-overlapping genomic regions embedded 172 candidate genes for AFC (Table 1).

For the BTA3, the RF analysis pointed three lead SNPs associated with AFC, located at 0.28 - 1.18 Mb regions, and harboring a total of 15 candidate genes (Table 1), some associated with neuronal functions and germinal cells maturation. For instance, the *TBX19* (T-Box 19) encodes an essential transcription factor (TPIT) for anterior pituitary POMC (proopiomelanocortin) cells expression in mice (Lamolet et al., 2001). Mutations on *TBX19* were related to congenital isolated adrenocorticotrophic hormone (ACTH) deficiency in humans (Couture et al., 2012). Presumably, *TBX19* gene might have importance for female fertility traits, since anterior pituitary also synthesizes reproductive hormones such as FSH and LH, which integrates the hypothalamus-pituitary axis regulation of the estrous cycle in females. Another gene located at BTA3 was the adenylated cyclase 10 (*ADCY10*), which has a critical role in

sperm maturation in the epididymis. It was noticed that splicing errors in this gene were responsible for bovine spermatozoa subfertility (Noda et al., 2013). In the present study, these three lead SNPs found at BTA3 encompasses a genomic region with 1Mb length, previously associated with muscling and conformation scores in Nellore cattle (Carreño et al., 2019). This is indicative that this QTL plays a pleiotropic role in Nellore cattle, affecting reproductive, growth and muscle development traits.

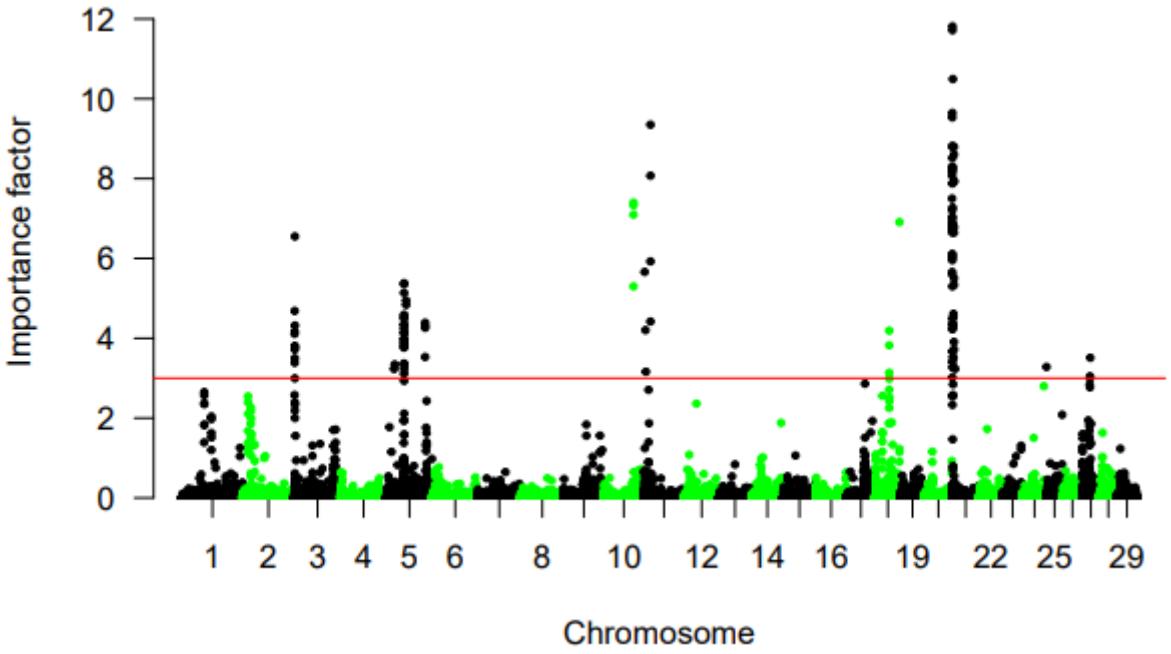


**Figure 3.** Manhattan plot considering the relative importance scores for each SNP averaged in five independent Random Forest analyses for age at first calving (AFC) in Nellore cattle. Negative importance scores were plotted as zero. The red line corresponds to the threshold value for SNP selection.

Four regions were associated with AFC on BTA5, with lead SNPs located at 19.89 Mb, 46.35 Mb, 51.72 Mb and, 101.74 Mb (Table 1). The ATPase, Ca++ transporting plasma membrane 1 (*ATP2B1*) ends approximately 218 Kb downstream the marker *BovineHD0500005765* (19.89 Mb). This same gene was located in the vicinities of single nucleotide polymorphisms significantly associated with calving interval in Italian Holstein Cattle (Minozzi et al., 2013). The ubiquitin Specific Peptidase 15 (*USP15*) gene was annotated on BTA5 near to 51.72 Mb. In mice, its orthologous is expressed in the developing acrosomal cap of spermatids in the testes (Crimmins et al., 2009). Further, the *FAM19A2* gene, found in the same genomic region, was

previously identified using Bayesian inference within a 1 Mb length window (51.9 – 52.3 Mb) explaining 1.78% of the additive genetic variance for weight gain from birth to weaning in Nellore cattle (Terakado et al., 2017).

Another strong candidate gene for AFC on BTA5 is the *DPPA3* (Developmental pluripotency-associated 3), which was highly expressed in the oocyte of human primordial follicles (Markholt et al., 2012) and in female mice embryonic gonads at 18.5 days after breeding (Small et al., 2005). This is a maternal effect gene that regulates normal development in mice during the embryo preimplantation stage. It has been detected in primordial germ cells, oocytes, preimplantation embryos and pluripotent cells (Payer et al., 2003). Similarly, *NANOG* and its paralog *NANOGNB* are highly expressed during embryo preimplantation stages in humans, mice, and cows (Dunwell and Holland, 2017). Therefore, it is clear that these pluripotency cell-associated neighboring genes (*DPPA3*, *NANOG,* and *NANOGNB*) have a synergic role on bovine embryo pre-implantation process, probably also coordinating cells differentiation after embryo fertilization (see Dunwell and Holland, 2017).

Besides *ADCY10* and *USP15,* expressed in male germ cells, other genes identified in the present study were previously related to bull fertility traits, such as *FUT8* (located on BTA10), significantly associated with sire conception rate (Rezende et al., 2018) and *NOB1* and *NFTA5* genes (located on BTA18), found in whole-exome sequencing of bulls divergent for fertility (Whiston, 2017). These findings corroborate the favorable genetic correlations between male and female reproductive traits reported for beef cattle (Terakado et al., 2015). Furthermore, some genes identified on BTA25 may have a deleterious role on male andrological parameters, such as the Calpain-15 (*CAPN15*), which has a causal variant affecting cryptorchidism susceptibility in rats (Barthold et al., 2016). On the other hand, the *AXIN1* has been shown to act as a suppressor of testicular germ cell tumors (Xu et al., 2017).

**Table 1.** Relevant SNPs identified in the Random Forest analysis for age at first calving (AFC) in Nellore cattle.

| SNP Name | BTA | Position (Mb) | Candidate genes within ± 250 Kbp interval | $f_{SNPj}$ |
|---|---|---|---|---|
| BovineHD0300000050 | 3 | 0.28 | *LOC790004, TBX19, SFT2D2, TIPRL, GPR161* | 4.11 |
| BovineHD0300000124 | 3 | 0.59 | *DCAF6, MPC2, ADCY10, LOC107131946* | 4.68 |
| BovineHD0300000291 | 3 | 1.18 | *MPZL1, RCSD1, LOC104971405, CREG1, CD247, POU2F1* | 6.55 |
| BovineHD0500005765 | 5 | 19.89 | *ATP2B1* | 3.23 |
| BovineHD0500013322 | 5 | 46.35 | *DYRK2, LOC101906308* | 5.35 |
| BovineHD0500014865 | 5 | 51.72 | *USP15, FAM19A2* | 4.94 |
| BovineHD0500029155 | 5 | 101.74 | *A2ML1, MIR2284R, MFAP5, RIMKLB, TRNAE-UUC, AICDA, APOBEC1, GDF3, DPPA3, LOC512775, NANOGNB, NANOG, SLC2A3* | 4.26 |
| BovineHD1000022233 | 10 | 77.84 | *FUT8, TRNAC-GCA* | 7.39 |
| BovineHD1100000982 | 11 | 2.78 | *KANSL3, FER1L5, LOC1O4973281, LMAN2L, CNNM4, CNNM3, ANKRD23, ANKRD39, SEMA4C, FAM178B, ACTR1B, COX5B, LOC783105* | 5.66 |
| BovineHD1100001777 | 11 | 4.94 | - | 4.20 |
| BovineHD1100002271 | 11 | 6.14 | *LOC107132896, NPAS2, RPL31, TRNAE-UUC, TBC1D8, CNOT11, RNF149, CREG2, RFX8* | 3.16 |
| BovineHD1100005574 | 11 | 18.17 | *LOC100140559, LOC100849080* | 9.35 |
| BovineHD1800011017 | 18 | 36.73 | *LOC100847995, SNTB2, VPS4A, COG8, TRNAE-UUC, TMED6, NIP7, TERF2, CYB5B, TRNAD-GUC, NFAT5, WWP2, NQO1, NOB1* | 4.18 |
| BovineHD1800018414 | 18 | 63.61 | *MBOAT7, RPS9, TMC4, TSEN34, PRPF31, LOC101905804, LENG1, TFPT, LOC101905303, LOC107131472, CNOT3, TARM1, NDUFA3, OSCAR, LOC107131475, LOC107131473, LOC107131474, LOC107131476, LOC107131477, LOC107131469, LOC107131468, LOC107131467, LOC107131465, NLRP13, NLRP8, NLRP5, ZNF444, LOC512150, LOC790201, LOC100140659, LOC768229, ZNF787, LOC101903510, LOC100139334, LOC530319, LOC783493, LOC532036, LOC782638, LOC100848202, LOC783562* | 6.90 |
| BovineHD2100000002 | 21 | 0.0087 | *SNRPN, SNURF* | 7.50 |
| BovineHD2100000071 | 21 | 0.645 | *NDN, MAGEL2, MKRN3* | 9.53 |
| BovineHD2100000134 | 21 | 1.13 | *LOC100336464, LOC101908683* | 11.81 |

**Table 1.** Relevant SNPs identified in the Random Forest analysis for age at first calving (AFC) in Nellore cattle. (Continued)

| | | | | |
|---|---|---|---|---|
| BovineHD2100000346 | 21 | 2.22 | *LOC107131587, LOC784898, LOC100848941, LOC101907203, UBE3A* | 10.49 |
| BovineHD2100000555 | 21 | 3.62 | *LOC100849023* | 4.61 |
| BovineHD2100000879 | 21 | 5.06 | *GABRG3* | 8.79 |
| BovineHD2100001732 | 21 | 8.01 | *SYNM, LOC104975310, IGF1R, LOC107131589, LOC101907342* | 3.23 |
| BovineHD2500000080 | 25 | 4.19 | *NPRL3, HBQ1, HBA1, HBM, HBZ, HBA, LUC7L, FAM234A, RGS11, PDIA2, ARHGDIG, RAB11FIP3, AXIN1, TMEM8A, DECR2, MRPL28, NME4, PIGQ, CAPN15, LOC104975822, RAB40C, NHLRC4, PRR35, WFIKKN1, STUB1, RHOT2, RHBDL1, LOC516108, METRN, FBXL16, WDR90, WDR24, LOC100139040, C25H16orf13, FAM173A, HAGHL, NARFL, CCDC78, MSLN* | 3.28 |
| BovineHD2700006032 | 27 | 21.24 | *MIR383, ENSBTAG00000029960* | 3.51 |

BTA = Bos Taurus Autosome; $f_{SNPj}$= SNP relative importance score

Initialing 54 Kb upstream from the *BovineHD1800011017* (BTA18) and neighboring the *NOB1* and *NFAT5*, it was identified the *WWP2* gene, which has been previously associated with gestation length in Holstein cattle using a representative database consisting of 27,214 bulls with ~3 million imputed sequence variants (Fang et al., 2019). Further, *WWP2* transcripts were reported to be highly expressed in the placentas of pregnant women exposed to an epigenetic compound (De Felice et al., 2015). Three genes flagged by the SNP *BovineHD1800018414* on BTA18 (63.61 Mb), namely *RPS9*, *CNOT3 and NLRP5*, may have an important role in the regulation of age at first calving. The *RPS9* gene encodes the ribosomal protein S9 that is a component of the 40S subunit. This gene is located at an intronic region of a putative *QTL* for calving traits (calving ease, calf size, stillbirth, birth index, body depth, stature) segregating in Holstein cattle at approximately 57 Mb (Mao et al., 2016). Furthermore, eight sequence variants of *RPS9* had the strongest associations with fertility traits ($p < 1 \times 10^{-10}$) in dairy cattle and, at the same time, exhibiting lesser expression in the corpus luteum of low fertility cows (Moore et al., 2016).

CCR4-NOT transcription complex subunit 3 (*CNOT3*) is a transcription activity regulator, this gene was flagged by SNPs validated for fertility (pregnancy within the first 42 days of mating) in two distinct dairy breeds (Pryce et al., 2010). In mammals, CNOT3 may have roles in embryonic viability, since a deficiency in this gene resulted in lethality at early embryonic stages in mice (Morita et al., 2011). Interestingly, interactions between *NANOS2* gene (Nanos C2HC-Type Zinc Finger 2) and the CCR4-NOT deadenylation complex (including *CNOT3*) perform an essential role in male germ cell development in mouse (Suzuki et al., 2012). Lastly, the *NLRP5*, also known as maternal antigen that embryo requires (MATER), integrates the subcortical maternal complex, an essential multiprotein complex for embryonic development and uniquely expressed in mammalian oocytes and early embryos (Bebbere et al., 2016).

Among the SNPs with the highest importance scores, four markers were located on BTA21, between 0.01 and 2.22 Mb (Table 1). This region harbors the *SNRPN*, *SNURF*, *MAGEL2, MKRN3* and *UBE3A* genes, which have well-known roles on epigenetic regulation of precocious puberty onset, reproductive hormones synthesis, oocytes development, and, pre and post-implantation of embryos in cattle and human (Suzuki Jr et al., 2009; O'Doherty et al., 2012; Abreu et al., 2013; Duittoz et al., 2016).

Using the single-step GBLUP (ssGBLUP) approach, the same set of genes was reported for early pregnancy in Nellore cattle, in a study that used partially the same dataset (Irano et al., 2016). Those authors noticed that a window comprising the genomic region between 8,725 and 3,028,689 bp on BTA21 (which flanks the SNPs with highest importance scores in the present study) was responsible for the largest genetic variance explained (1.31 %) for early pregnancy. Similarly, the *MAGEL2* gene was previously flagged using sliding windows with 1 Mb length that explained the highest proportion of genetic variance for scrotal circumference in Nellore cattle (Utsunomiya et al., 2014). Therefore, the empirical evidence provided so far points out that the SNPs identified in our study on BTA21 (between 0.01 and 2.22 Mb) are in linkage disequilibrium with a QTL presenting a major effect for fertility-related traits in Nellore cattle.

The 10 most significant gene ontology terms obtained with DAVID analysis evidenced biological processes related to hemoglobin functions such as oxygen transporter activity and oxygen binding (Table S1). Hemoglobin (Hb) is mainly found in erythrocyte cells, however, there is recent evidence of ovarian regulation of Hb synthesis through the ovulatory signal cascade, with high expression of Hb subunits in human granulosa and cumulus cell samples, suggesting a potential role of the hemoglobin complex in early embryo development (Brown et al., 2015). As reported in the DAVID analysis, the genes clustered into hemoglobin complex pathways were *HBZ, HBM, HBA, HBA1, HBQ1, CYB5B* (Figure 4). Other genes such as *DPPA3*, *TERF2, APOBEC1,* and *AICDA* participate in the regulation of epigenetic processes, *e.g*. DNA methylation (Figure 4), suggesting that the biologic processes involving the response variables are not merely additive.

Another interesting biological pathway involving the genes *LOC107131465*, *LOC107131469, LOC107131476,* and *LOC107131477* was associated with pheromone receptor activity (Figure 4). Pheromone activity influences sexual behavior and reproductive hormones secretion in different species. Although the role of pheromone in cattle reproduction is not fully understood, there are shreds of evidence that beef heifers attain puberty faster when exposed to the male presence (Oliveira et al., 2009; Fiol et al., 2010). Furthermore, Fiol and Ungerfeld (2016) reported that exposing anestrous heifers to androgenized steers promoted an increase in basal

levels of LH after 10 days of exposure. Therefore, the high frequency of favorable alleles involved in pheromone recognition is particularly interesting in extensive beef production systems, where males and females are raised together.

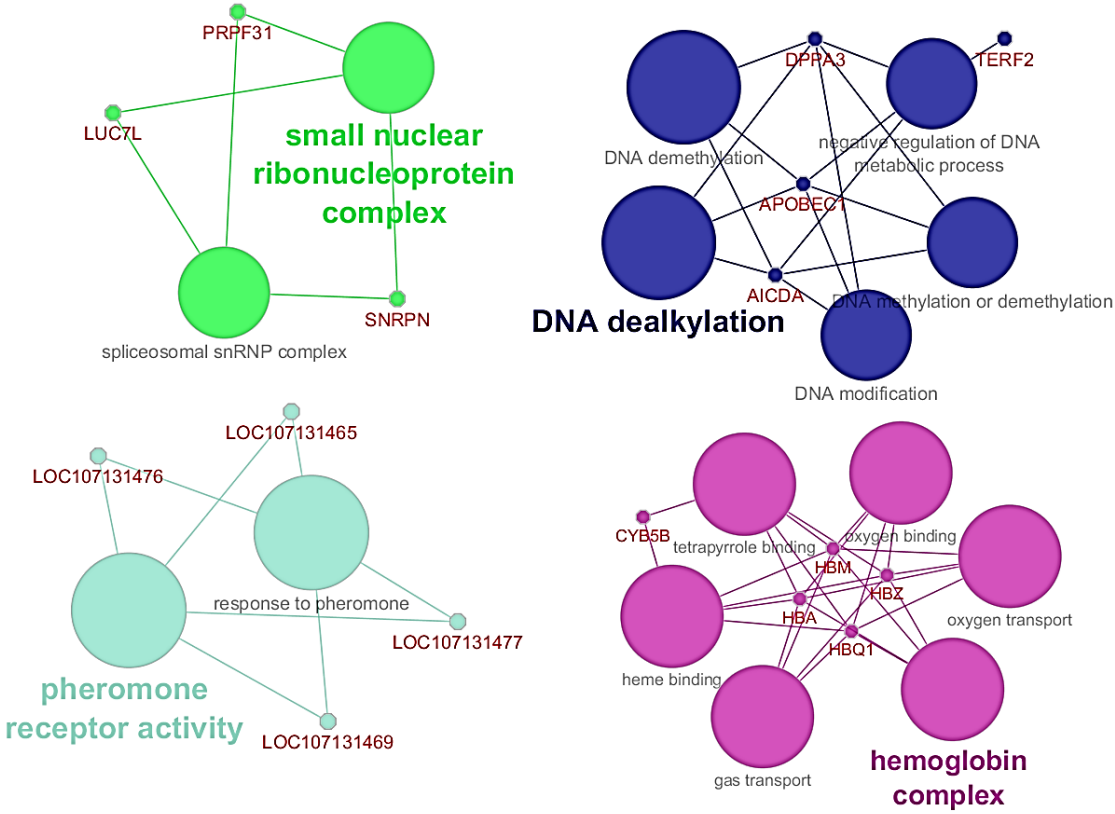


**Figure 4 –** Gene network for Age at First Calving (AFC) in Nellore cattle. *Different node colors represent the functional groups in which candidate genes are involved.*

In the present study, a non-parametric method was applied for ranking SNPs associated with AFC according to their predictive importance. In summary, an extensive search in the literature revealed that many annotated genes have well-known functions associated, among others, with embryo pre-implantation, embryonic development, male fertility, synthesis of reproductive hormones and, pheromone recognition. Some genomic regions identified on BTA3 and BTA5 in the present study were previously associated with weight gain from birth to weaning and visual scores at weaning in Nellore cattle (Terakado et al., 2017; Carreño et al., 2019), these traits are intimately related to heifers' body condition before puberty onset. In beef cattle, high body size delays the puberty onset, whereas animals with high weight-height ratios at

eleven months of age are expected to have low age at puberty (Pereira et al., 2017). Therefore, genes with important roles in the regulation of growth traits are expected to influence fertility as well.

Further, a QTL strongly associated with fertility-related traits in Nellore cattle, validated with different methods and in different populations (Utsunomiya et al., 2014; Irano et al., 2016) was also highlighted in the present study, which reinforces RF effectiveness for pre-screening candidate regions associated with complex traits. Nevertheless, some regions that were significantly associated with age at first calving in previous studies were not identified here, for instance on BTA14 (Mota et al., 2014). This may be due to data particularities such as sample size, the extent of LD, allelic frequency and herd management and to some extent to the used method.

Age at first calving reflects the heifer genetic potential in at least three different stages, the time to puberty onset, interval between puberty onset and the first conception and gestation length, it is a sex-limited trait, presenting low to moderated heritability estimates (Grossi et al., 2008; Mota et al., 2017; Schimidt et al., 2018) and with polygenic nature, which impose several limitations on gene mapping. Hence, results from GWAS using different methods are expected to provide complementary insights for clarifying the genetic mechanisms involved in the AFC expression. Most of the standard parametrical methods for genomic scans focus only on major effects whereas genomic variants with hidden patterns that may contribute to the phenotype expression remain few explored. The individual tree structure in RF enables to explore implicitly variables that present both marginal and interaction effects (Yao et al., 2013). This property is particularly appealing for GWAS since the biological mechanisms probably incorporate nonlinear processes such as pathways and gene-networks.

The random forest approach has been successfully applied for genome-wide scanning in livestock data. For instance, Mokry et al. (2013) applied a random forest approach to identify a subset of SNPs that explained approximately 50% of the deregressed breeding values for backfat thickness in Canchin beef cattle. Similarly, Yao et al. (2013) examining the most frequently occurring descendent pairs within RF, identified SNP with potential epistatic effects for residual feed intake in dairy cattle. Furthermore, it has been shown, recently, that RF is an efficient methodology to identify

an optimal subset of SNPs for genomic prediction of growth traits in beef cattle (Li et al., 2018).

## 4 Conclusions

To the best of our knowledge, this was the first attempt of applying a non-parametric approach for scanning potential loci affecting reproductive traits in Nellore cattle using high-density panels. The RF-based genome-wide scan and functional analysis highlighted candidate genes with key roles in fertility, including embryo pre-implantation and development, embryonic viability, male germinal cells maturation and pheromone recognition. Results reported in the present study are expected to enhance the knowledge of biological mechanisms regulating the expression of age at first calving in Nellore cattle.

## 5 References

Abeygunawardena H, Dematawewa CMB (2004) Pre-pubertal and postpartum anestrus in tropical Zebu cattle. **Anim Reprod Sci.** 82:373–387.

Abreu AP, Dauber A, Macedo DB, Noel SD, Brito VN, Gill JC, Cukier P, Thompson IR, Navarro VM, Gagliardi PC, Rodrigues T, Kochi C, Longui CA, Beckers D de ZF, Montenegro LR, Mendonca BB, Carroll RS, Hirschhorn JN, Latronico AC, Kaiser UB (2013) Central precocious puberty caused by mutations in the imprinted gene MKRN3. **N. Engl. J. Med.** 368, 2467–2475.

Barthold JS, Pugarelli J, Madolyn LM, Ren J, Modupeore OA, Polson SW, Mateson A, Wang Y, Sol-Church K, McCahan SM, Akins Jr RE, Devoto M, Robbins AK (2016) Polygenic inheritance of cryptorchidism susceptibility in the LE/orl rat. Molecular **Human Reproduction** 22[1]:18-34.

Bebbere D, Masala L, Albertini DF, Ledda S (2016) The subcortical maternal complex: multiple functions for one biological structure? **J. Assist. Reprod. Genet.** 33:1431-1438.

Bindea G, Mlecnik B (2012) ClueGo, a Cytoscape plug-in to decifer biological networks v2.0.0. User's manual.

Brown HM, Anastasi MR, Frank LA, Kind KL, Richani D, Robker RL, Russel DL, Gilchrist RB, Thompson JG (2015) Hemoglobin: a Gas Transport Molecule That Is Hormonally Regulated in the Ovarian Follicle in Mice and Humans. **Biology of Reproduction** 92(1):1-10.

Carreño LOD, Pessoa MC, Espigolan R, Takada L, Bresolin T, Cavani L, Baldi F, Carvalheiro R, Albuquerque LG, Fonseca R (2019) Genome Association Study for Visual Scores in Nellore Cattle Measured at Weaning. **BMC Genomics** 20:150.

Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien AMP, Sölkner J, McEwan JC, Van Tassell CP, Sonstegard TS, Garcia JF (2014) Accuracy of genotype imputation in Nellore cattle. **Genetics Selection Evolution** 46:69.

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. **Genomics** 99:323-329.

Couture C, Saveanu A, Barlier A, Carel JC, Fassnacht M, Fluck CE, Houang M, Maes M, Phan-Hug F, Enjalbert A, Drouin J, Brue T, Vallette S (2012) Phenotypic homogeneity and genotypic variability in a large series of congenital isolated ACTH-deficiency patients with TPIT gene mutations. **J. Clin. Endocr. Metab.** 97: E486-E495.

De Felice B, Manfellotto F, Palumbo A, Troisi J, Zullo F, Di Carlo C, Sardo ADS, De Stefano N, Ferbo U, Guida M, Guida M (2015) **BMC Medical Genomics** 8:56.

Duittoz AH, Tillet Y, Bourhis DL, Schibler L (2016) The timing of puberty (oocyte quality and management. **Anim. Reprod**. 13:313-333.

Dunwell TL, Holland PWH (2017) A sister of *NANOG* regulates genes expressed in pre-implantation human development. **Open Biol.** 7:170027

Fang L, Jiang J, Li B, Zhou Y, Freebern E, Vanraden PM, Cole JB, Liu GE, Ma Li (2019) Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. **Comunications Biology** 2:100.

Fernandes Júnior GA, Rosa GJM, Valente BD, et al. (2016) Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution** 48:1-8.

Fiol C, Quintans G, Ungerfeld R (2010) Response to biostimulation in peri-puberal beef heifers: Influence of male-female proximity and heifer's initial body weight. **Theriogenology** 74(4):569–575. doi:10.1016/j.theriogenology.2010.03.015

Fiol C, Ungerfeld R (2016) Positive effects of biostimulation on luteinizing hormone concentration and follicular development in anestrous beef heifers. **J. Anim. Sci.** 94:971–977. doi:10.2527/jas2015-9396

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: Methodological consideration & new findings. **BMC Genetics**11.

Grossi D, Frizzas O, Paz C, Bezerra L, Lôbo R, Oliveira J, Munari D (2008) Genetic associations between accumulated productivity, and reproductive and growth traits in Nelore cattle. **Livestock Science** 117(2), 139–146.

Crimmins S, Sutovsky M, Chung P-C, Huffman A, Wheeler C, Swing DA, Roth K, Wilson J, Sutovsky P, Wilson S (2009) **Dev. Biol.** 325(1):33-42.

Huang DW, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. **Nucleic Acids Research** 37:1–13.

Huang DW, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. **Nature Protocols**. 4:44-57

Irano N, Camargo GMF, Costa RB, Terakado APN, Magalhães AFB, Silva RMO, Dias MM, Bignardi AB, Baldi F, Carvalheiro R, Oliveira HN, Albuquerque LG (2016) Genome-Wide Association Study for Indicator Traits of Sexual Precocity in Nellore Cattle. **PLoS ONE** 11(8): e0159502. doi:10.1371/journal pone.0159502

James G, Witten D, Hastie T, Tibishirani R (2013) An Introduction to Statistical Learning: with applications in R. New York: SPRINGER, 426p.

Lamolet B, Pulichino A-M, Lamonerie T, Gauthier Y, Brue T, Enjalbert A, Drouin J (2001) A pituitary cell-restricted T box factor, Tpit, activates POMC transcription in cooperation with Pitx homeoproteins. **Cell** 104: 849-859.

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. **Nature Reviews** 16:321-331.

Li, B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genome Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics** 9.

Liaw A, Wiener M (2002) Classification and regression by randomForest. **R News** 2:18-22.

Malhado CHM, Malhado ACM, Filho RM, Carneiro PLS, Pala A, Carrillo JA (2013) Age at first calving of Nellore cattle in the semi-arid region of northeastern Brazil using linear, threshold, censored and penalty models. **Livest. Sci.** 154:28–33.

Mao X, Kadri NK, Thomasen JR, De Koning DJ, Sahana G, Guldbrandtsen B (2016) Fine mapping of a calving QTL on *Bos Taurus* autosome 18 in Holstein cattle. **Journal of Animal Breeding and Genetics.** 133:207-218.

Markholt S, Grondahl ML, Ernst EH, Andersen CY, Ernst E, Lykke-Hartmann K (2012) Global gene analysis of oocytes from early stages in human folliculogenesis shows high expression of novel genes in reproduction. **Molecular Human Reproduction** 18[2]:96-110.

Melo TP, Takada L, Baldi F, Oliveira HN, Dias MM, Neves HHR, Schenkel FS, Albuquerque LG, Carvalheiro C (2016) Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations. **BMC Genetics** 17:89.

Minozzi G, Nicolazzi EL, Stella A, Biffani S, Negrini R, Lazzari B, Ajmone-Marsan P, Williams JL (2013) Genome Wide Analysis of Fertility and Production Traits in Italian Holstein Cattle. **Plos One** 8:11.

Mokry FB, Higa RH, Mudadu MA, Lima AO, Meirelles SLC, Silva MVGB, Cardoso FF, De Oliveira MM, Urbinati I, Niciura SCM, Tullio RR, De Alencar MM, Regitano LCA (2013) Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**. 14:47.

Moore SG, Pryce JE, Hayes BJ, Chamberlain AJ, Kemper KE, Berry DP, McCabe M, Cornican P, Lonergan P, Fair T, Butler S (2016) **Biology of Reproduction** 94(1):1-11.

Morita M, Oike Y, Nagashima T, Kadomatsu, T, Tabata M, Suzuki T, et al. (2011) Obesity resistance and increased hepatic expression of catabolismrelated mRNAs in Cnot3+/- mice. **EMBO J**. 30:4678–4691. doi: 10.1038/emboj. 2011.320

Morota G, Boddhireddy P, Vukasinovic N, Gianola D, Denise S (2014) Kernel based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Front Genet**. 5:56.

Mota RR, Guimarães SEF, Fortes MRS, Hayes B, Silva FF, Verardo LL, Kelly MJ, Campos CF, Guimarães JD, Wenceslau RR, Penitente-Filho JM, Garcia JF, Moore S (2017) Genome-wide association study and annotating candidate gene networks affecting age at first calving in Nellore cattle. **J Anim Breed Genet** 1-9.

Nascimento AV, Romero ARS, Utsunomiya YT, Utsunomiya ATH, Cardoso DF, Neves HHR, Carvalheiro R, Garcia JF, Grisolia AB (2018) Genome-wide association study using haplotype alleles for the evaluation of reproductive traits in Nelore cattle. **Plos One** 13(8): e0201876.

Noda T, Sakase M, Fukushima M, Harayama H (2013) Novel approach for the detection of the vestiges of testicular mRNA splicing errors in mature spermatozoa of Japanese Black bulls. **PLoS ONE** 8:e57296.

O'Doherty AM, O'Shea LC, Fair T (2012) Bovine DNA methylation imprints are established in an oocyte size specific manner, which are coordinated with the expression of the DNMT3 family proteins. **Biol Reprod**. 86. doi: 10.1095/biolreprod.111.094946.

Oliveira CMG, Oliveira Filho BD, Gambarini ML, Viu MAO, Lopes DT, Sousa APF (2009). Effect of biostimulation and nutritional supplementation on pubertal age and pregnancy rates of Nelore heifers (Bos indicus) in a tropical environment. **Anim. Reprod. Sci.** 113(1–4):38–43. doi:10.1016/j.anireprosci.2008.08.006.

Payer B, Saitou M, Barton SC, Thresher R, Dixon JP, Zahn D, Colledge WH, Carlton MB, Nakano T, Surani MA (2003) Stella is a maternal effect gene required for normal early development in mice. **Curr Biol** 13:2110 – 2117.

Pereira GR, Barcellos JOJ, Sessim AG, Tarouco JU, Feijó FD, Neto JB, Prates ER, Canozzi MEA (2017) Relationship of post-weaning growth and age at puberty in crossbred beef heifers. **R. Bras. Zootec.** 46(5):413-420.

Pryce JE, Bolormaa S, Chamberlain AJ, Bowman PJ, Savin K, Goddard ME, Hayes BJ (2010) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. **J. Dairy Sci.** 93:3331-3345.

R Development Core Team (2011) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. R Core Team.

Rezende FM, Diestsch GO, Peñagaricano F (2018) Genetic dissecation of bull fertility in US Jersey dairy cattle. **Animal Genetics** 49:393-402.

Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. **BMC Genomics** 15:1-12.

Sartori R, Bastos MR, Baruselli PS, Gimenes LU, Ereno RL, Barros CM (2010) Physiological differences and implications to reproductive management of Bos taurus and Bos indicus cattle in a tropical environment. **Soc Reprod Fertil** 67:357-375.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. **Genome Res.** 13:2498-2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658.

Small CL, Shima JE, Uzumcu M, Skinner MK, Griswold MD (2005) Profiling gene expression during the differentiation and development of the murine embryonic gonad. **Biol Reprod** 72:492– 501.

Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AM, Mills JL, Brody LC, Stambolian D, Bailey-Wilson J (2016) r2VIM: A new variable selection method for random forests in genome-wide association studies. **BioData Mining** 9[7]:1-15.

Schmid M, Bennewitz J (2017) Invited review: Genome-wide association analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs. **Arch. Anim. Breed.** 60:335–346.

Schmidt PI, Campos GS, Lôbo RB, Souza FRP, Brauner CC, Boligon AA. Genetic analysis of age at first calving, accumulated productivity, stayability and mature weight of Nellore females (2018) **Theriogenology** 108(1):81-87.

Suzuki Jr J, Therrien J, Filion F, Lefebvre R, Goff AK, Smith LC (2009) In vitro culture and somatic cell nuclear transfer affect imprinting of SNRPN gene in pre- and post-

implantation stages of development in cattle. **BMC Developmental Biology** 9:9.

Suzuki A, Saba R, Miyoshi K, Morita Y, Saga Y (2012) Interaction between NANOS2 and the CCR4-NOT Deadenylation Complex Is Essential for Male Germ Cell Development in Mouse. **PLoS ONE** 7(3): e33558.

Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, Konig IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. **Genetic Epidemiology** 33(S1): S51-S57.

Teixeira DBA, Fernandes Júnior GA, Silva DBS, Costa RB, Takada L, Gordo DGM, Bresolin T, Carvalheiro R, Baldi F, Albuquerque LG (2017) Genomic analysis of stayability in Nellore cattle. **Plos One** 12[6]:e0179076.

Terakado APN, Loligon AA, Baldi F, Silva JAIV, Albuquerque LG (2015). Genetic associations between scrotal circumference and female reproductive traits in Nelore cattle. **Journal of Animal Science** 93[6]:2706-2713.

Terakado APN, Costa RB, Camargo GMF, Irano N, Bresolin T, Takada L, Carvalho CVD, Oliveira HN, Carvalheiro R, Baldi F, Albuquerque LG (2017) Genome-wide association study for growth traits in Nelore cattle. **Animal** 12(7):1358-1362.

Utsunomiya YT, Carmo AS, Neves HHR, Carvalheiro R, Matos MC, Zavarez LB, Ito PKRK, O'Brien AMP, Sölkner J, Porto-Neto LR, Schenkel FS, McEwan J, Cole JB, Silva MVGB, Van Tassel CP, Sonstegard TS, Garcia JF (2014) Genome-Wide Mapping of Loci Explaining Variance in Scrotal Circumference in Nellore Cattle. **Plos One** 9(2): e88561

Whiston R (2017) **Genetic variation in bulls divergent for fertility**. 191 p. Thesis (Doctor of Philosophy) – Scholl of Biochemistry and Immunology, Trinity College Dublin.

Xu H, Feng Y, Jia Z, Yang J, Lu X, Li J, Xia M, Wu C, Zhang Y, Chen J (2017) AXIN1 protects against testicular germ cell tumors via the PI3K/ AKT/mTOR signaling pathway. **Oncol Lett** 14:981–986.

Yao C, Spurlock DM, Armentano LE, Page Jr CD, Vandehaar MJ, Bickhart DM (2013) Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. **Journal of Dairy Science** 96:6716–6729.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassel CP, Sonstegard TD, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL (2009) A whole-genome assembly of the domestic cow, Bos taurus. **Genome biology**, 10(4):R42. doi: 10.1186/gb-2009-10-4-r42.

## SUPPLEMENTARY FILE

**Table S1 –** Most significant gene ontology terms identified in DAVID analysis involving annotated genes associated with age at first calving in Nellore cattle.

| Category | Term | Count | *p*-value[a] | FDR |
|---|---|---|---|---|
| INTERPRO | Hemoglobin, alpha | 5 | 2,3E-6 | 1,1E-5 |
| GOTERM_CC_DIRECT | Hemoglobin complex | 5 | 6,1E-5 | 5,8E-4 |
| UP_KEYWORDS | Oxygen transport | 5 | 1,6E-4 | 1,8E-3 |
| GOTERM_MF_DIRECT | Oxygen transporter activity | 5 | 1,7E-4 | 1,6E-3 |
| GOTERM_MF_DIRECT | Oxygen binding | 5 | 4,0E-4 | 3,8E-3 |
| INTERPRO | Globin-like | 5 | 6,1E-4 | 2,9E-3 |
| INTERPRO | Globin, structural domain | 5 | 6,1E-4 | 2,9E-3 |
| INTERPRO | Globin | 5 | 4,5E-4 | 2,1E-3 |
| INTERPRO | Hemoglobin, pi | 3 | 3,4E-2 | 1,6E-1 |
| UP_KEYWORDS | Heme | 6 | 5,9E-2 | 6,6E-1 |

[a]Adjusted by Bonferroni correction; FDR = False Discovery Rate