

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**MÉTODOS PARAMÉTRICOS E NÃO
PARAMÉTRICOS PARA A PREDIÇÃO DE VALORES
GENÉTICOS GENÔMICOS DE CARACTERÍSTICAS DE
IMPORTÂNCIA ECONÔMICA EM SUÍNOS**

**Letícia Borges Joaquim
Zootecnista**

**JABOTICABAL-SP
2019**

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**MÉTODOS PARAMÉTRICOS E NÃO
PARAMÉTRICOS PARA A PREDIÇÃO DE VALORES
GENÉTICOS GENÔMICOS DE CARACTERÍSTICAS DE
IMPORTÂNCIA ECONÔMICA EM SUÍNOS**

Letícia Borges Joaquim

Orientador: Prof. Dr. Danísio Prado Munari

Coorientadora: Dra. Priscila Arrigucci Bernardes

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

**JABOTICABAL-SP
2019**

J62m

Joaquim, Letícia Borges

Métodos paramétricos e não paramétricos para a predição de valores genéticos genômicos de características de importância econômica em suínos / Letícia Borges Joaquim. -- Jaboticabal, 2019
65 p. : il., tabs.

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

Orientador: Danísio Prado Munari

Coorientadora: Priscila Arrigucci Bernardes

1. Brnn. 2. Espessura de gordura. 3. Número de leitões nascidos vivos. 4. "Random forest". 5. SnnR. 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.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: MÉTODOS PARAMÉTRICOS E NÃO PARAMÉTRICOS PARA A PREDIÇÃO DE VALORES GENÉTICOS GENÔMICOS DE CARACTERÍSTICAS DE IMPORTÂNCIA ECONÔMICA EM SUÍNOS

AUTORA: LETICIA BORGES JOAQUIM

ORIENTADOR: DANISIO PRADO MUNARI

COORDINADORA: PRISCILA ARRIGUCCI BERNARDES

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

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



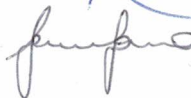
Pós-doutoranda TATIANE CRISTINA SELEGUIM CHUD Participação por videoconferência
UOGELPH / University of Guelph - Canadá



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



Pesquisadora Dra. LENIRA EL FARO ZADRA
Instituto de Zootecnia / Sertãozinho/SP



Pós-doutorando GUILHERME BATISTA DO NASCIMENTO
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal



Jaboticabal, 29 de novembro de 2019

DADOS CURRICULARES DO AUTOR

LETÍCIA BORGES JOAQUIM – nascida em dia 5 de maio de 1991, na cidade Tubarão, Santa Catarina, filha de Rildo Vieira Joaquim e Márcia Borges Joaquim. Iniciou o curso de Zootecnia em março de 2009 na Universidade Federal de Santa Catarina, Câmpus Florianópolis, e obteve o título de bacharel em Zootecnia em março de 2014. Em agosto de 2014, ingressou no curso de mestrado pelo Programa de Pós-graduação em Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias, UNESP, Câmpus Jaboticabal sob orientação do Prof. Dr. Danísio Prado Munari e coorientação da Dra. Mônica Corrêa Ledur e do Dr. Rodrigo Pelicioni Savegnago. Foi bolsista CAPES pelo período de 01 de agosto de 2014 a 29 de fevereiro de 2016. Em março de 2016, ingressou no curso de doutorado pelo Programa de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciência Agrárias e Veterinárias, UNESP, Câmpus de Jaboticabal sob orientação do Prof. Dr. Danísio Prado Munari e coorientação da Dra. Priscila Arrigucci Bernardes. Foi bolsista CAPES de março de 2016 a julho de 2019.

Maria, Maria
É um dom, uma certa magia
Uma força que nos alerta
Uma mulher que merece viver e amar
Como outra qualquer do planeta

Maria, Maria
É o som, é a cor, é o suor
É a dose mais forte e lenta
De uma gente que ri quando deve chorar
E não vive, apenas aguenta

Mas é preciso ter força
É preciso ter raça
É preciso ter gana sempre
Quem traz no corpo a marca
Maria, Maria
Mistura a dor e a alegria

Mas é preciso ter manha
É preciso ter graça
É preciso ter sonho sempre
Quem traz na pele essa marca
Possui a estranha mania
De ter fé na vida

Milton Nascimento / Fernando Brant

DEDICO

Aos meus pais, Márcia e Rildo, por me ensinar que o conhecimento ninguém pode nos roubar e por não medirem esforços para nos proporcionarem o melhor ensino.

Ao meu avô, Ary (*in memoriam*), por todas as tardes que ficou ao meu lado quando eu era uma criança assustada que chorava para não ficar na escola.

A minha irmã, Patrícia, por ser realmente minha irmã mais velha e estar sempre ao meu lado.

AGRADECIMENTOS

A toda minha família, por me incentivar e apoiar as minhas escolhas e compreenderem a minha ausência. Aos meus pais Rildo e Márcia por serem meus maiores modelos de honestidade, altruísmo e força e vontade. À minha irmã, Patrícia e meu cunhado, Anderson, por aguentarem todas as vezes que eu implico com eles e por sempre estarem presentes quando precisamos. Aos meus avós, Ary (*in memoriam*) e Ana, por toda inspiração e mimos, Pedro (*in memoriam*) e Ivonete (*in memoriam*), por todo cuidado que sempre ofereceram quando era preciso. Aos meus tios e primos pelos momentos alegres compartilhados.

Ao meu orientador, Prof. Dr. Danísio Prado Munari, por me aceitar e receber tão bem durante todo o período da pós-graduação. Pela confiança e paciência em mim depositada na execução desse trabalho. Obrigada por compreender e aceitar a minha teimosia.

A minha coorientadora doutora e agora professora Priscila, pelo exemplo de bondade e prestatividade. Por embarcar comigo nessa loucura de espera de dados e tornar possível a realização dessa tese em praticamente um mês. Obrigada também por toda amizade, ajuda e momentos de distração que você e o Elieder me proporcionaram. Contem com meu apoio nessa nova fase no melhor estado do país.

Ao Gilmar por tornar essa fase final muito mais leve e fácil. Obrigada por entender meus momentos difícil e estar presente me apoiando e incentivando. Principalmente, obrigada por me ensinar que não devemos adiar sonhos quando podemos realizarmos, e assim colocar na minha vida o mais novo membro de quatro patas da família. Ao Bart, que com um só olhar tem meus melhores sorrisos e me faz esquecer de todo o resto.

A todos os amigos de departamento de Ciências Exatas e da Pós-Graduação de Genética e Melhoramento Animal. Aqueles que já não estão mais tão presentes: Tatiane, Jorge, Ana Paula, Daniela Barreto, Thiago Bruno, Marcus, Guilherme Nascimento, Rodrigo e Jaqueline; e aqueles que tornaram esses últimos anos mais divertidos: Rafael, Fernando, Rebeka, Samla, Larissa e Leyde. Obrigada por me ensinarem o verdadeiro sentido da frase quem tem amigo tem tudo.

A Hayala e Naiara por dividirem o lar comigo entendendo minhas manias, me proporcionando boas risadas e compartilhando bons cafés, almoços e jantas. Com cada uma de vocês eu aprendi algo, obrigada!

Aos meus amigos de Santa Catarina, da faculdade, do colégio e da vida... por todos os momentos que já compartilhei com cada um. Ao Carlos Ignácio, Nacho ou simplesmente Max por essa amizade que nos faz chorar sempre que um liga ou manda mensagem para contar sobre uma nova conquista. Ao Bianchini por esse amigo que eu sei que posso ficar um mês ou um ano sem conversar e estará ok, por ser esse amigo que eu sei que atenderia um chamado de socorro ou encontro na França. Obrigada por ter ido me apresentar Paris ter me proporcionado momentos que eu jamais teria sonhado.

Aos meus amigos do “Crossfit” Bull Terrie: Priscila, Júlia, Leticia, Monique, Felipe, Albertina, Edgar, Nicolas... por tornarem nossas manhãs de exercícios muito mais divertidas e criarem momentos de lazer que não acabavam em assuntos da pós-graduação.

Aos meus colegas da Choice, em especial ao Sr. Yves por me ensinar muito sobre a dinâmica de uma empresa de melhoramento de suínos, e principalmente por confiar no meu trabalho.

Aos membros da banca do Exame Geral de Qualificação, Dr. Roberto Carneiro, Dr. Rodrigo Pelicioni Savegango, Dr. Henrique Nunes de Oliveira e Dra. Ana Fabrícia Braga Magalhães, e aos membros da banca defesa da tese, Dra. Lenira El Faro Zadra, Dr. Guilherme Batista do Nascimento, Dra. Tatiane Cristina Seleguim Chud e Dr. Roberto Carneiro, pelas valiosas sugestões que enriqueceram esse trabalho.

A todos que foram importantes em algum momento, mas minha memória não permitiu citar.

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

SUMÁRIO

	Página
RESUMO.....	iii
ABSTRACT	v
CAPÍTULO 1 – Considerações gerais	1
1 INTRODUÇÃO	1
2 REVISÃO BIBLIOGRÁFICA.....	3
2.1 Características de importância econômica na suinocultura	3
2.2 Seleção Genômica	7
2.3 Métodos não paramétricos de aprendizado de máquina.....	10
3 REFERÊNCIAS.....	12
CAPÍTULO 2 – Estudo de metodologias de seleção genômica em uma população de tipo Landrace.....	19
1 INTRODUÇÃO	20
2 MATERIAL E MÉTODOS	21
2.1 Descrição dos dados fenotípicos e genotípicos	21
2.2 Edição dos dados.....	22
2.3 Estimativas de parâmetros genéticos e predição do valor genético .	23
2.4 Predição dos valores genéticos genômicos	24
2.5 Acurácia dos valores genéticos genômicos	27
3 RESULTADOS E DISCUSSÃO.....	28
4 CONCLUSÕES	31
5 REFERÊNCIAS.....	31
CAPÍTULO 3 – Predição dos valores genéticos genômicos utilizando conjuntos de SNPs identificados com o método “Random Forest”	35
1 INTRODUÇÃO	36
2 MATERIAL E MÉTODOS	38

2.1 Dados fenotípicos e genotípicos	38
2.1 Cenários utilizados na seleção de SNPs.....	38
2.3 “Random Forest”	39
2.4 Importância dos marcadores	40
2.5 Seleção genômica utilizando grupo de marcadores.....	41
2.6 Acurácia dos valores genéticos genômicos	42
4 RESULTADOS E DISCUSSÃO.....	42
4 CONCLUSÕES	48
5 REFERÊNCIAS.....	48

MÉTODOS PARAMÉTRICOS E NÃO PARAMÉTRICOS PARA A PREDIÇÃO DE VALORES GENÉTICOS GENÔMICOS DE CARACTERÍSTICAS DE IMPORTÂNCIA ECONÔMICA EM SUÍNOS

RESUMO – A seleção genômica tem sido usada em vários programas de melhoramento de plantas e animais proporcionando ganhos na acurácia de seleção quando comparado a seleção tradicional. Contudo, é importante ressaltar que a superioridade da seleção genômica depende de vários fatores tais como a metodologia de predição dos valores genéticos genômicos. Além do desafio de encontrar a melhor metodologia para a aplicação da seleção genômica, o alto custo de implementação também representa uma dificuldade na sua ampla utilização nos programas de melhoramento animal, principalmente em suínos e aves. Os objetivos deste trabalho foram: (i) avaliar a capacidade de predição de quatro diferentes métodos de seleção genômica para características reprodutivas e produtivas; (ii) utilizar o método “Random Forest” para selecionar conjuntos de marcadores SNPs mais relevantes na explicação dos fenótipos para as características estudadas e verificar o impacto do uso desses conjuntos de SNPs na acurácia dos valores genômicos preditos para duas características de importância econômica na suinocultura. No estudo foi utilizado um arquivo de pedigree com 879.965 animais distribuídos em 13 gerações e um arquivo de fenótipos composto por 73.439 observações de espessura de gordura (EG) e 69.505 registros de número de leitões nascidos vivos por leitegada (NV). Além disso, informações de 969 animais de uma linhagem fêmea tipo Landrace genotipados com um painel customizado com 57.692 marcadores distribuídos por todo o genoma foram utilizadas. A seleção genômica foi realizada aplicando os métodos lineares GBLUP – “Genomic Estimated Breeding Value” e “Single-step” GBLUP e os não lineares brnn - “Bayesian Regularized Neural Network” e snnR – “Sparse Neural Networks”. Para verificar o impacto do uso de conjuntos de SNPs na habilidade de predição da seleção genômica foram avaliados 5 cenários: 1 - 3.000 SNPs mais importantes para EG; 2 - 3.000 SNPs mais importantes para NV; 3- 3.000 SNPs mais importantes para as duas características em conjunto; 4 - 3.000 SNPs selecionados aleatoriamente; e 5 - todos os SNPs. O método “Random Forest” foi utilizado para seleção dos SNPs dos três primeiros cenários e os valores genéticos genômicos foram preditos utilizando o método “Single-step” GBLUP. No estudo das metodologias de seleção genômica verificou-se que as acurácias de predições foram melhores para a característica EG em relação a característica NV. Também se observou que o método que apresentou a melhor predição dos valores genéticos genômicos na população estudada para ambas as características foi “single-step” GBLUP quando comparado ao método GBLUP e os métodos de redes neurais snnR e brnn. Na avaliação da predição dos valores genéticos genômicos utilizando conjuntos de SNPs identificados com o método de aprendizagem de máquina, demonstrou-se que esse método pode ser utilizado para selecionar eficientemente conjuntos de marcadores para formarem painéis de genotipagem de baixa densidade e assim reduzir os custos da implementação da seleção genômica. Isto porque, o conjunto de marcadores selecionados com o mesmo apresentou habilidade de predição semelhante quando comparado a habilidade de predição

com a informação de todos os SNPs na seleção genômica apenas para a característica número de leitões nascidos vivos.

Palavras-chave: Brnn, espessura de gordura, número de leitões nascidos vivos, “Random Forest”, snnR

PARAMETRIC AND NON-PARAMETRIC METHODS FOR THE PREDICTION OF GENETIC VALUES OF ECONOMICALLY IMPORTANT TRAITS IN PIGS

ABSTRACT - Genomic selection has been used in several plant and animal breeding programs providing selection accuracy gains when compared to traditional genetic analysis. However, it is important to emphasize the superiority of genomic selection depends on several factors, such as the methodology for predicting genomic values. In addition to the challenge of finding the best model for application of genomic selection, the high cost of implementation also makes difficult the use of genomic selection in breeding programs of all animal species. Therefore, the aims of this study were: (i) evaluate the prediction ability of four different genomic selection methods using reproductive and productive traits; (ii) use Random Forest analysis to select the most important markers for studied traits and to verify the impact of using these subsets of SNPs on the accuracy of predicted genomic values for two economically important traits in pig production. The pedigree files contained 879,965 individuals, which spanned up to 13 generations and phenotype file composed of 73,439 backfat thickness observations and 69,505 litter records. A total of 969 animals of a Landrace female line were genotyped using a custom Illumina chip consisting of 57,692 SNPs distributed throughout the genome. Genomic selection was performed by applying the linear methods GBLUP (Genomic Estimated Breeding Value) and Single-step GBLUP and the nonlinear methods brnn (Bayesian Regularized Neural Network) and snnR ("Sparse Neural Networks). The prediction capacity of each method was tested by Pearson correlation, mean square measure error and the regression coefficient. To verify the impact of the use of SNP sets on the prediction accuracy of genomic selection, 5 scenarios were evaluated: 1 - 3,000 most important SNPs for backfat thickness; 2 - 3,000 most important SNPs to born alive piglets; 3 - 3,000 most important SNPs for both traits together; 4 - 3,000 randomly selected SNPs; and 5 all SNPs. The Random Forest method was used to select SNPs from the first three scenarios and genomic selection was performed using the GBLUP single-step method. In the study of genomic selection methodologies, it was found that the prediction accuracy was better for the backfat thickness trait in relation to the number of piglets born alive. The results also show that the method that presented the best prediction of genomic genetic values in the population studied for the backfat thickness and number of piglets born alive was single-step GBLUP when compared to GBLUP and neural network methods snnR and brnn. In assessing the prediction of genomic genetic values using subsets of SNP selected with the machine learning method, it has been shown that the Random Forest method can be used to efficiently select subset marker to form low density genotyping panels and thus reduce the costs of implementing genomic selection. This is because the set of markers selected from using machine learning method presented similar prediction ability when compared the prediction ability with the information of all SNPs in the genomic selection for the trait number of born alive piglets.

Keywords: Brnn, backfat thickness, number of piglets born alive, Landrace, “Random Forest”, snnR

CAPÍTULO 1 – Considerações gerais

1 INTRODUÇÃO

Desde sua domesticação, há aproximadamente 4.900 AC, os suínos têm representado papel importante na produção de alimento para o consumo humano, sendo atualmente a fonte de proteína animal mais consumida no mundo (Zhang et al., 2018). No cenário mundial da produção de suínos, o Brasil destaca-se como quarto maior produtor, com produção crescente apesar de seu plantel de reprodutores se manter constante com aproximadamente 38,8 milhões de cabeças (USDA, 2018). Esse resultado denota a evolução na eficiência da produtividade brasileira, sendo consequência da profissionalização da cadeia produtiva e do melhoramento genético da espécie.

As primeiras atividades de melhoramento genético de suínos no Brasil datam do século passado. No entanto, a consolidação dos programas de melhoramento ocorreu somente nas décadas de 50 e 60, com o início das primeiras importações de animais dos Estados Unidos da América e Europa (Euclides Filho, 1999; Lopes et al., 2008). Nessa mesma época, Henderson (1975) desenvolveu a metodologia dos modelos mistos que permitiu a seleção de animais com base na estimação acurada de seus valores genéticos. Desde então, o modelo animal tem sido direcionado para a predição dos valores genéticos, permitindo a seleção para características quantitativas de importância econômica em plantas e animais (Robinson, 1991; Jonas e Koning, 2015). Na década seguinte, com o estabelecimento de grandes empresas de melhoramento genético no Brasil, iniciou-se a comercialização de reprodutores e matrizes (Euclides Filho, 1999; Lopes et al., 2008).

A implementação da genética molecular ampliou as estratégias para os estudos de características quantitativas e possibilitou melhorias em programas de melhoramento genético, como no caso do desenvolvimento de teste de DNA para identificar animais portadores da síndrome do estresse suíno, fazendo com que os suínos fossem a primeira espécie de produção a se beneficiar com a introdução da genética molecular (Samorè e Fontanesi, 2016). Posteriormente, Meuwissen et al.

(2001) foram os primeiros a propor estudos de seleção genômica (GS, “Genomic Selection”) para prever o valor genético de indivíduos candidatos à seleção utilizando a informação dos genótipos de milhares de marcadores do tipo polimorfismo de nucleotídeo único (SNP, “single-nucleotide polymorphism”).

Quando o conceito de GS foi definido, ferramentas para sua realização ainda não haviam sido desenvolvidas, mas com os avanços das tecnologias genômicas e dos métodos estatísticos para estimar o efeito dos marcadores tornou-se possível a implementação da seleção genômica (Jiang et al., 2013; Feuk et al., 2006). Isso permitiu a primeira implementação prática da seleção genômica em 2009 na avaliação do gado leiteiro nos Estados Unidos da América (VanRaden et al., 2009).

Embora a seleção genômica tenha se tornado importante para o melhoramento de bovinos de leite, principalmente pela substancial diminuição no intervalo de geração e redução dos custos com o teste de progênie (Schaeffer, 2006), em suínos os programas de melhoramento genético existentes já trabalham com um intervalo de geração muito curto. Logo, espera-se que o impacto da GS nesse parâmetro seja muito menor em comparação com os obtidos no gado leiteiro. Em contrapartida, a acurácia dos valores genéticos, geralmente, é baixa na espécie, e, portanto, a GS pode auxiliar na predição de valores genéticos mais acurados em comparação ao método tradicional baseado no pedigree (Tribout et al., 2013).

Diante desse cenário, os objetivos deste trabalho foram: (i) avaliar a habilidade de predição de quatro diferentes métodos de seleção genômica para características reprodutivas e produtivas; (ii) utilizar o método “Random Forest” para selecionar conjuntos de marcadores mais importantes na explicação dos fenótipos para características estudadas e verificar o impacto do uso desses conjuntos de SNPs na acurácia dos valores genéticos genômicos preditos para duas características de importância econômica na suinocultura.

2 REVISÃO BIBLIOGRÁFICA

2.1 Características de importância econômica na suinocultura

Muitos programas de melhoramento envolvem características cuja mensuração tem elevado custo e por esse motivo um número limitado de animais tem registro de desempenho em relação ao total de indivíduos envolvidos no sistema de produção. Desta forma, a seleção é realizada em um número reduzido de animais, os quais possuem registro de todas as características e o ganho genético disseminado para toda população. Essa seleção em um número limitado de animais, subsequente multiplicação na próxima geração e a produção de animais de abate em grande escala na geração final originam a estrutura piramidal do programa de produção de suínos e aves (Oldenbroek e Waaij, 2015).

Na suinocultura, normalmente a pirâmide possui todos os estratos de produção bem definidos, constituída pelos rebanhos núcleo, multiplicador e comercial. Na base da pirâmide tem-se o estrato comercial que é responsável pela produção dos animais de abate gerados a partir do cruzamento dos animais originários dos estratos multiplicadores ou núcleo. O estrato multiplicador recebe animais das granjas núcleo e promove o cruzamento entre esses, com o objetivo de obter a máxima complementariedade entre raças, enquanto que o estrato núcleo é responsável pela seleção intensiva sobre as características economicamente importantes. Neste estrato, os animais dividem-se em rebanhos de linha paterna e rebanhos de linha materna, em razão do objetivo genético que se almeja alcançar com cada linhagem (Fávero e Figueiredo, 2009; Da Costa, 2014).

As linhagens paternas são formadas normalmente por animais puros ou cruzados das raças Duroc, Large White, Pietrain, Hampshire e Landrace, selecionadas para alta produção de carne na carcaça. O cruzamento entre essas raças tem por objetivo obter maiores ganhos por heterose, de forma a alcançar um produto final com as características exigidas pelo mercado e reprodutores que atendam essas exigências. O critério de seleção utilizado, geralmente, nas linhas machos são: espessura de gordura, consumo alimentar, ganho médio diário de peso e conversão alimentar (Fávero e Figueiredo, 2009).

Nas linhagens maternas o objetivo é selecionar matrizes com capacidade de desmamar grande número de leitões por parto e por ano. Para isso, as raças utilizadas em maior escala são Landrace e Large White. Além dessas raças, algumas linhagens desenvolvidas também apresentam raças chinesas em sua composição, a exemplo da raça Meishan que é conhecida por sua alta prolificidade. A seleção praticada nos rebanhos núcleo da linha fêmea fundamenta-se, principalmente, nas características reprodutivas, como número de leitões nascidos vivos, desmamados e peso da leitegada. Além das características reprodutivas, os índices de seleção desenvolvidos para estas linhagens também consideram características produtivas, como crescimento e características da carcaça (Fávero e Figueiredo, 2009).

Uma das características mais utilizadas na avaliação da carcaça de suínos é a espessura de gordura, visto que essa apresenta fácil mensuração no animal vivo e possui associação genética favorável com a maioria das características de carcaça, além de apresentar importância econômica (Lopes et al., 2001). Por isso, essa característica recebeu grande ênfase nos programas de melhoramento a partir dos anos 90, quando o mercado consumidor passou a exigir carne mais magra e, desde então, a seleção aplicada na espécie modificou o suíno conhecido popularmente como tipo banha em animal com pouca gordura na carcaça, conhecido atualmente como suíno “light”. Números extraídos do banco de dados mantido pela Associação Brasileira de Criadores de Suínos (2005) exemplificam essa evolução na espessura de gordura dos animais avaliados no teste de seleção, mostrando que houve redução de aproximadamente 12 mm na espessura de gordura de 1985 até 2005.

Segundo Imboonta et al. (2007), outra característica economicamente importante na produção de suínos é o tamanho de leitegada, uma vez que a rentabilidade do sistema de produção suinícola é mensurada pelo número de leitões desmamados por porca ao ano. E essa, por sua vez, é fortemente influenciado por inúmeros fatores, tais como número de dias não produtiva, intervalo desmame-cio e número de leitões nascidos vivos por parto (Lucia Junior, 2000). Por isso, as empresas desenvolvem programas de melhoramento genético que permitam oferecer aos produtores matrizes com maior número de leitões

nascidos vivos ao parto, possibilitando, em alguns casos, uma produção com mais de 35 leitões por porca por ano. De acordo com o Relatório Anual do Desempenho de Produção de Suínos (AGRINESS, 2017), de 2008 a 2017 houve aumento na média de nascidos vivos brasileira de 2,04 leitões, a partir de uma produção que em 2008 apresentava média de 12,67 e em 2017 média de 14,71 leitões nascidos vivos por parto.

Como observado, a seleção tem promovido melhora em características como espessura de gordura e quantidade de leitões nascidos vivos. Para alcançar essa evolução simultânea em características produtivas e reprodutivas, que geralmente apresentam correlações genéticas negativas, é fundamental a obtenção de estimativas acuradas dos parâmetros genéticos, garantindo assim a predição acurada dos valores genéticos (Da Costa et al., 2001; Imboonta et al., 2007). Em consequência da variação dos parâmetros genéticos entre populações e ambientes, torna-se necessária a estimação dos mesmos na população em que a seleção será aplicada. Sendo assim, vários trabalhos sobre estimação de parâmetros genéticos são encontrados na literatura para características como número de leitões nascidos vivos por leitegada e espessura de gordura (Tabela 1).

Tabela 1: Médias observadas e estimativas de herdabilidade para número de leitões nascidos vivos por parto e para espessura de gordura da carcaça em diferentes raças de suínos.

Autores	Raça	μ_{EG}	h^2_{EG}	μ_{NV}	h^2_{NV}
Lopez et al. (2018)	Landrace	12,9 (2,6)	0,54 (0,01)	-	-
Lopez et al. (2018)	Yorkshire	13,9 (2,7)	0,45 (0,01)	-	-
Smetko et al. (2015)	Pietrain	7,06 (1,2)	0,28 (0,03)	-	-
Imboonta et al. (2007)	Landrace	13,9 (3,6)	0,61 (0,02)	-	-
Giné et al. (2004)	Large White	35,3 (5,0)	0,17	-	-
Da Costa et al. (2001)	Landrace	17,1 (0,6)	0,50	-	-
Ogawa et al. (2019)	Landrace	-	-	10,1 (2,7)	0,12 (0,01)
Ogawa et al. (2019)	Large White	-	-	10,1 (2,5)	0,10 (0,01)
Ye et al. (2018)	Large White	-	-	12,2 (3,5)	0,06 (0,007)
Lopez et al. (2017)	Landrace	-	-	10,6 (3,1)	0,09 (0,002)
Lopez et al. (2017)	Yorkshire	-	-	11,4 (3,2)	0,11 (0,001)
Tusell et al. (2013)	Landrace	-	-	12,1 3,25	0,21
Hanenbergh et al. (2001)	Landrace	-	-	10,2 (2,9)	0,09 (0,005)
Lee et al. (2015)	Berkshire	25,1 (5,2)	0,57	7,6 (2,6)	0,06 (0,03)
Kapell et al. (2009)	Large White	9,9 (1,8)	0,52	11,1 (3,2)	0,14
Holm et al. (2004)	Landrace	9,9 (1,4)	0,44 (0,03)	10,4 (2,8)	0,10 (0,01)
Skorupski et al. (1996)	Landrace	14,4 (2,1)	0,45 (0,03)	9,9 (2,5)	0,09 (0,03)
Skorupski et al. (1996)	Duroc	13,0 (2,2)	0,46 (0,04)	9,3 (2,8)	0,16 (0,05)

μ_{EG} : Média observada e desvio padrão entre parênteses da espessura de gordura (mm); h^2_{EG} : herdabilidade de espessura de gordura; μ_{NV} : Média observada e desvio padrão entre parênteses do número total de leitões nascidos vivos; h^2_{NV} : herdabilidade do número total de leitões nascidos vivos;

2.2 Seleção Genômica

A Seleção Genômica proposta por Meuwissen et al. (2001) é baseada na predição do valor genético genômico (GEBV, “Genomic Estimated Breeding Value”) de cada indivíduo, utilizando informações fenotípicas, de pedigree, haplótipos e os efeitos dos marcadores moleculares espalhados por todo o genoma (Jiang et al., 2018; Ballesta et al., 2019). O efeito dos marcadores é estimado por meio da regressão do fenótipo pelo genótipo dos animais da população de treinamento, ou seja, animais que possuem tanto informações fenotípicas como genotípicas. Posteriormente, essas estimativas são utilizadas para prever o GEBV dos indivíduos em que há registro de informações genotípicas, mas sem informações de fenótipo. Portanto, esses animais são selecionados a partir das informações de seus SNPs sem a necessidade de informações de fenótipo de todos os animais que fazem parte do programa de seleção (Samorè e Fontanesi, 2016).

Estudos avaliando a aplicação da seleção genômica (GS) em diferentes espécies de interesse econômico tem demonstrado ganhos na acurácia de seleção quando comparado a seleção tradicional com base no pedigree (Preisinger, 2012; Tribout et al., 2013). Segundo Resende et al. (2012), a superioridade da GS em relação a seleção baseada em fenótipos pode ser atribuída aos seguintes fatores: uso da matriz de parentesco real e própria de cada caráter, aumentando a acurácia; viabilização da seleção precoce, o que aumenta o ganho genético por tempo; possibilidade da avaliação repetida de cada alelo sem realização de testes de progênie, fato que aumenta a acurácia; e uso combinado de três tipos de informação (fenotípica, genotípica e de parentesco) para corrigir os dados e fazer a avaliação genômica, aumentando também a acurácia. Contudo, é importante ressaltar que a superioridade da seleção genômica depende de vários fatores, dentre os quais podem ser citados a densidade dos marcadores, número de animais genotipados, o grau de parentesco entre os animais da população de treinamento e a população de predição e a metodologia de predição dos valores genéticos genômicos (Akanno et al., 2014).

Vários métodos de predição de valores genéticos genômicos foram propostos, como, por exemplo, os métodos GBLUP, Bayes A e Bayes B

(Meuwissen et al., 2001) e outros métodos bayesianos (Gianola et al., 2009). A diferença entre essas abordagens está na suposição da distribuição *a priori* considerada para estimar os efeitos dos SNPs. O método GBLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos, o método Bayes A assume que poucos marcadores têm grandes efeitos e muitos marcadores apresentam pequenos efeitos e no método Bayes B muitos efeitos de marcadores são assumidos como zero (Garrick e Saatchi, 2011). Neste último método, como muitos efeitos de SNPs são assumidos como zero *a priori*, reduz-se a quantidade de marcadores devido a concentração nas partes em que existem QTLs (“Quantitative trait loci”). De acordo com Resende et al. (2008), o método ideal é aquele que reflete melhor a natureza biológica da característica em questão em termos de efeitos gênicos, assim, o melhor método pode variar de acordo com a característica estudada.

Aguilar et al. (2010) e Misztal et al. (2009) propuseram um método alternativo conhecido por “single-step” GBLUP (ssGBLUP). Esse método incorpora informações de SNPs nos modelos tradicionais levando em consideração as informações de fenótipo e pedigree disponíveis e incluindo também a performance e pedigree de animais não genotipados. Desta maneira, a estratégia de “single-step” permitiu criar um novo cenário com a possibilidade de uso de grande quantidade de fenótipos colhidos em animais genotipados ou não genotipados e usando também as informações do pedigree já registrados nas linhagens de seleção (Chistensen e Lund., 2010). Esse método, quando comparado às abordagens de múltiplos passos é vantajoso em algumas populações pois permite a utilização de toda a informação disponível (pedigree, fenótipos e genótipos) na análise de predição, permitindo assim obter predições de valores genéticos genômicos para todos os animais, genotipados ou não. Principalmente em populações com limitado número de animais genotipados, permitindo assim que a matriz de parentesco atue complementando a genômica.

Os métodos de redes neurais também têm sido propostos como alternativa estatística para predição dos valores genéticos genômicos. O método de redes neurais artificiais, como exemplo do método de regularização bayesiana (brnn, “Bayesian Regularized Neural Network”), é caracterizado como o aproximador

universal de funções complexas (Hornik et al., 1989), que são baseadas no uso de uma ampla gama de funções arbitrárias capazes de explorar intuitivamente relações lineares e não lineares entre uma variável resposta (fenótipo) e várias outras variáveis preditoras (os genótipos). Isso é possível usando algoritmos de treinamento, que atualizam “pesos” específicos (divididos em diferentes níveis chamados neurônios), para cada variável explicativa em diferentes níveis de aprendizado (chamados de camadas).

No método de redes neurais proposto por Wang et al. (2016) (snnR - “Sparse Neural Networks”) encontra-se uma estrutura esparsa ideal de uma rede neural para atenuar o problema de ajuste quando a matriz de relacionamento é utilizada como entrada na aplicação da seleção genômica. O método implementa uma rede do tipo “feed-forward” com muitas camadas ocultas para melhorar o poder da modelagem em prever características complexas mais do que quando é utilizada a arquitetura de uma única camada na rede neural. Segundo os autores, em comparação com o método brnn, essa metodologia melhora a eficiência computacional, bem como a habilidade de predição da seleção genômica no melhoramento animal porque o snnR implementa um rede neural “sparse” com muitas camadas ocultas.

Os progressos genéticos ocasionados pela seleção genômica ocorrem principalmente pelo aumento da acurácia de seleção e redução do intervalo de geração. Porém, sistemas atuais de melhoramento de suínos já são caracterizados por curtos intervalos de geração, menos do que 2 anos, e com rápida renovação das gerações devido ao fato da espécie apresentar baixa idade à puberdade (Tribout et al., 2011). Assim, a principal vantagem da aplicação da GS em programas de melhoramento de suínos é ocasionada pelo aumento da acurácia dos valores genéticos preditos sem aumentar o intervalo de gerações e sem reestruturação do programa de melhoramento. Esse aumento em acurácia é importante principalmente para as características que apresentam baixa herdabilidade, como as características reprodutivas (tamanho de leitegada), para características mensuradas em poucos indivíduos (conversão alimentar) e também para características de qualidade de carne, medidas *post-mortem* (Samorè e Fontanesi, 2016).

Exemplos da implementação da seleção genômica em suínos para diferentes características de produção e de prolificidade tem reportado melhorias na acurácia de predição em relação ao método de avaliação com base apenas no fenótipo e no pedigree, pelo método BLUP. Forni et al. (2011), estudando a característica número total de nascidos por leitegada em animais cruzados verificaram que a média da acurácia de predição variou entre 0,28 a 0,49 para o método ssGBLUP, enquanto que a média da acurácia para a metodologia BLUP foi de 0,22. Resultado semelhante foi encontrado para suínos das raças Landrace e Yorkshire para a característica tamanho da leitegada, em que a acurácia de predição também foi maior para o método ssGBLUP (0,21) em comparação com as metodologias GBLUP (0,17) e BLUP tradicional (0,09) (Guo et al., 2015). Para características produtivas, trabalhos tem demonstrado que acurácia de predição pode variar de 0,30 a 0,55 dependendo da metodologia aplicada, da raça da população estudada e da densidade dos marcadores. Grossi et al. (2018), trabalhando com painel de 60K, encontraram valores de acurácia para a característica espessura de gordura de 0,55, 0,55 e 0,51 em suínos das raças Duroc, Landrace e Yorkshire, respectivamente.

2.3 Métodos não paramétricos de aprendizado de máquina

Os avanços nas tecnologias genômicas criaram enormes desafios para pesquisadores como aquele que envolve um pequeno número de observações e um grande número de variáveis preditoras, comumente conhecido como desafio grande P pequeno N. Métodos estatísticos convencionais baseados em hipóteses univariadas e assumindo variáveis explicativas independentes, sofrem significativamente devido à falta de poder e precisão para lidar com a complexidade de múltiplas interações ou correlações entre preditores (Li et al., 2018). Vários métodos estatísticos foram desenvolvidos para melhorar a predição de conjuntos de dados nessa condição, incluindo modelos paramétricos como método de seleção de subconjunto criado por Breiman (1995) e LASSO de Tibshirani (1996). No entanto, todos esses modelos são exigentes em termos computacionais e ignoram dependência entre variáveis explicativas. A aplicação de várias

abordagens com os métodos de aprendizado de máquina (ML – “Machine Learning”) tem sido proposta como solução desses problemas. Essa abordagem não exige qualquer conhecimento prévio sobre modelos genéticos subjacentes (aditivo, dominante ou recessivo), sendo apropriada para a pré-triagem de importantes variáveis de predição. Além disso, esses podem detectar as interações entre SNP-SNP ou SNP-covariável (Li et al., 2018).

Os métodos ML tem sido aplicados em diferentes pesquisas científicas, a exemplo dos estudos de associação genética (Stephan et al., 2015), detecção de epistasia para identificação e tratamento de câncer (Shi e He, 2016) e predição de sítios de ligação de DNA de proteína a partir de sequências de aminoácidos (Wu et al., 2009). Entre as metodologias de ML, o método mais conhecido é o “Random Forest” (RF) (Li et al., 2018). Esse método possibilita analisar dados de alta dimensionalidade e pode capturar modelos não-lineares baseados em genótipos, produzindo interpretações simples.

O método “Random Forest” foi proposto por Breiman (2001) como extensão de seus trabalhos anteriores, tendo como base o método “bagging” para gerar conjuntos de classificadores. Conforme descrito por Breiman (1996), “bagging” de preditores é um método para a geração de múltiplas versões de um preditor com objetivo de usá-los para conseguir uma predição agregada. O RF consiste de um conjunto de árvores de decisão para propósitos de classificação ou análises de regressão de dados, utilizando estruturas (árvores) do tipo CART (“Classification and Regression Tree”), em que o conjunto de entrada é particionado recursivamente usando divisões binárias em cada nó interno da árvore até obter respostas homogêneas ou quase homogêneas nos nós terminais ou nas folhas da árvore (Breiman et al., 1984). Esse método não-paramétrico é robusto a “outliers”, ruídos e ao super ajustamento dos dados de treinamento, além de ser de fácil treinamento (Touw et al, 2012).

O algoritmo de RF começa pela seleção de uma amostra aleatória de inicialização dos dados. O subconjunto aleatório de variáveis é selecionado e procura até encontrar a variável ideal para ser o nó (folhas) de divisão. Este processo é repetido para cada nó até que uma árvore seja formada. Ao final da construção das árvores não há alteração dessas estruturas. Os dados que não

fazem parte da amostra de “bootstrap” são aplicados em cada árvore para derivar a taxa de erro denominada “out-of-bag” (OOB) e as medidas de importância das variáveis. Este processo é repetido até que uma floresta esteja completa com o número de árvores especificado inicialmente. Após treinar k árvores de decisão (classificadores), para agrupar as escolhas e tomar uma decisão final, o modelo considera a escolha majoritária sem ponderação.

3 REFERÊNCIAS

ABCS (2005) Associação Brasileira de Criadores de Suínos Disponível em: <<http://www.abcs.org.br>> Acesso em: 20 maio 2018.

AGRINESS (2017) Relatório anual de desempenho da produção de suínos. 10^o Edição. Disponível em: <<http://145jio1p2h9417125d2acd6p.wpengine.netdna-cdn.com/wp-content/uploads/2018/05/RelatorioED10.pdf>> Acesso em: 20 maio 2018.

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal Dairy Science** 93: 743–752.

Akanno EC, Schenkel FS, Sargolzaei M, Friendship RM, Robinson JA (2014) Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. **Journal of Animal Breeding and Genetics** 131: 367–378.

Ballesta P, Maldonado C, Pérez-Rodríguez P, Mora F (2019) SNP and haplotype-based genomic selection of quantitative traits in *Eucalyptus globulus*. **Plants** 8:31

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

Breiman, L (1995) Better subset regression using the nonnegative garrote. **Technometrics** 37: 373–384.

Breiman L (1996) Bagging Predictors. **Machine Learning** 24(2):123-140.

Breiman, L (2001). Random forests. **Machine Learning** 45(1): 5–32.

Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution** 42.

Da Costa ARC, Lopes PS, Torres RA, Regazzi AJ, Silva MA, Euclydes RF, Pires AV (2001) Estimação de Parâmetros Genéticos em Características de Desempenho de Suínos das Raças Large White, Landrace e Duroc. **Revista Brasileira de Zootecnia** 30: 49-55.

Da Costa, ARC. Estruturas piramidais de melhoramento genético. In: Ferreira, A. H.; Carraro, B.; Dallanora, D.; Machado, G.; Machado, I. P.; Pinheiro, R.; Rohr, S. **Produção de suínos: teoria e prática**. Brasília: Qualitá, 2014. Cap. 2. p. 51-60.

Do DN, Janss LLG, Jensen J, Kadarmi HN (2015) SNP annotation- based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. **Journal of Animal Science** 93:2056–2063.

Euclides Filho, K. **Melhoramento genético animal no Brasil: fundamentos, história e importância**. Campo Grande: Embrapa Gado de Corte, 63p. 1999.

Fávero JA, De Figueiredo EAP, (2009) Evolução do melhoramento genético de suínos no Brasil. **Revista Ceres** 56: 420-427.

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. **Nature Review Genetics** 7: 85–97.

Forni S, Aguilar I, Mistral I (2011) Different genomic relationship matrices for single-step matrices analysis using phenotypic, pedigree and genomic information. **Genetics Selection Evolution** 43:1.

Garrick DJ, Saatchi M (2011) Opportunities and challenges for genomic selection of beef cattle. **Revista Brasileira de Zootecnia** 40: 310-316.

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. **Genetics** 136: 245–247.

Giné GAF, De Freitas RTF, De Oliveira AIG, Pereira IG, Gonçalves TM (2004) Estimativa de Parâmetros Genéticos para Características de Carcaça em um Rebanho de Suínos Large White. **Revista Brasileira de Zootecnia** 33: 337-343.

Grossi DA, Brito LF, Jafarikia M, Schenkel FS, Feng Z (2018) Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. **Animal** 12: 2235-2245.

Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G (2015) Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. **Journal of Animal Science** 93: 503–512.

Hanenbergh EHAT, Knol EF, Merks JWM (2001) Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. **Livestock Production Science** 69: 179–186.

Henderson CR (1975) Best Linear Unbiased estimation and prediction under a selection model. **Biometrics** 31: 423-439.

Holm B, Bakken M, Vangen O, Rekaya R (2004) Genetic analysis of litter size, parturition length, and birth assistance requirements in primiparous sows using a joint linear-threshold animal model. **Journal of Animal Science** 82: 2528–2533.

Hornik K., Stinchcombe M., White H (1989), Multilayer feedforward networks are universal approximators. **Neural Networks** 2: 359–366.

Imboonta N, Rydhmer L, Tumwasorn S (2007) Genetic parameters for reproduction and production traits of Landrace sows in Thailand. **Journal of Animal Science** 85: 53–59.

Jiang L, Jiang J, Yang J, Liu X, Wang H, Ding X, Liu J, Zhang Q (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. **BMC Genomics** 14: 131.

Jiang Y, Schmidt RH, Reif JC (2018) Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. **G3: GENES, GENOMES, GENETICS** 8(5): 1687-1699

Jonas E, Koning DJ (2015). Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. **Frontiers in Genetics** 6:49

Lee JH, Song KD, Lee HK, Cho KH, Park HC, Park KD (2015) Genetic Parameters of Reproductive and Meat Quality Traits in Korean Berkshire Pigs. **Asian-Australasian Journal of Animal** 28:1388-1393.

Li B, Zhang N, Wang Y-G, George AW, Reverter A and Li Y (2018) Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics** 9: 237.

Lopes BIM, Song C, Seo K (2018) Genetic parameters and trends for production traits and their relationship with litter traits in Landrace and Yorkshire pigs. **Animal Science Journal** 89: 1381-1388.

Lopez BI, Kim TH, Makumbe MT, Song CW, Seo KS (2017) Variance components estimation for farrowing traits of three purebred pigs in Korea. **Asian-Australasian Journal of Animal Sciences** 30: 1239–1244.

Lopes P.S, Freitas RTF, Ferreira A.S. **Melhoramento de suínos**. UFV. 39 p. (Caderno Didático, 37). 2001.

Lopes OS, Guimarães SEF, Carneiro PLS (2008) Integração da genética quantitativa e da genética molecular nos programas de melhoramento genético de suínos. In: SIMPÓSIO NACIONAL DE MELHORAMENTO ANIMAL, 7, São Carlos. **Anais...** São Carlos, SP: SBMA, p.1-20.

Lucia Junior T, Dial GD, Marsh WE (2000) Lifetime reproductive and financial performance of female swine. **Journal of the American Veterinary Medical Association** 216: 1802–1809.

Kapell DN, Ashworth CJ, Walling GA, Lawrence AB, Edwards SA, Roehe R (2009) Estimation of genetic associations between reproduction and production traits based on a sire and dam line with common ancestry. **Animal** 3: 1354-1362.

Misztal I, Legarra A, Aguilar I (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal Dairy Science** 92: 4648–4655.

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

Ogawa S, Konta A, Kimata M, Ishii K, Uemoto Y, Satoh M (2019) Estimation of genetic parameters for farrowing traits in purebred Landrace and Large White pigs. **Animal Science Journal** 90: 23-28.

Oldenbroek K, Waaij LVD, 2015. **Textbook Animal Breeding and Genetics for BSc students**. Centre for Genetic Resources The Netherlands and Animal Breeding and Genomics Centre, 2015. Groen Kennisnet: Diponível em: <https://wiki.groenkennisnet.nl/display/TAB/>

Preisinger R (2012) Genome-wide selection in poultry. **Animal Production Science** 52: 121–125.

Resende MDV, Silva FF, Lopes OS, Azevedo CF. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p.

Resende MDV, Lopes OS, Silva RL, Pires IE (2008) Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira** 56: 63-77.

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects: rejoinder. **Statistical Science** 6, 48–51.

Samorè AB, Fontanesi L (2016) Genomic selection in pigs: state of the art and perspectives. **Italian Journal of Animal Science** 15: 211-232.

Schaeffer R (2006) Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding Genetics** 123: 218-223.

Shi M, He J, (2016) SNRFCB: sub-network based randomforest classifier for predicting chemotherapy benefit on survival for cancer treatment. **Molecular BioSystems** 12: 1214–1223.

Skorupski MT, Garrick DJ, Blair HT (1996) Estimates of genetic parameters for production and reproduction traits in three breeds of pigs. **New Zealand Journal of Agricultural Research** 39: 387-395.

Smetko A, Škorput D, Luković Z, Mahnet Ž, Klišanić V, Špehar M (2015) Estimation of genetic parameters for production traits in pig breeds in Croatia. **Poljoprivreda/Agriculture** 21: 146-149.

Stephan J, Stegle O, Beyer A (2015) A random forest approach to capture genetic effects in the presence of population structure. **Nature Communications** 6:7432.

Tibshirani, R (1996) Regression shrinkage and selection via the lasso. **Journal of The Royal Statistical Society Series B-Statistical Methodology** 58, 267–288

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? **Briefings in bioinformatics** 14(3): 315-326.

Tribout T, Bidanel JP, Pochas F, Schwob S, Guillaume F, Larzul C (2011) La selection genomique: principe et perspectives pour l'amélioration des populations porcines. **Journées de la Recherche Porcine en France** 43:13–26.

Tribout T, Larzul C, Phocas F (2013) Economic aspects of implementing genomic evaluations in a pig sire line breeding scheme. **Genetics Selection Evolution** 45: 40-56.

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

UNITED STATES DEPARTMENT OF AGRICULTURE - USDA. Livestock and Poultry: Word Markets and trade. 2018. Disponível em: <https://downloads.usda.library.cornell.edu/usda-esmis/files/73666448x/mg74qq69r/j6731729p/livestock_poultry.pdf> Acesso em:16 janeiro 2018.

Ye J, Tan C, Hu X, Wang A, Wu Z (2018) Genetic parameters for reproductive traits at different parities in Large White pigs. **Journal of Animal Science** 96: 1215-1220.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. **Journal Dairy Science** 92: 16–24.

Wang Y, Mi X, Rosa GJM, Chen Z, Lin P, Wang S, Bao Z (2016) Technical note: an R package for fitting sparse neural networks with application in animal breeding. **Journal of Animal Science** 96(5): 2016-2026

Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, et al (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. **Bioinformatics** 25: 30–35.

Zhang J, Chai J, Luo Z, He H, Chen L, Liu X, Zhou Q (2018) Meat and nutritional quality comparison of purebred and crossbred pigs. **Animal Science Journal** 89: 202–210.

CAPÍTULO 2 – Estudo de metodologias de seleção genômica em uma população de suínos tipo Landrace

RESUMO – A acurácia da seleção genômica depende de diversos fatores, tais como a metodologia de predição dos valores genéticos genômicos utilizada. Logo, a escolha do método estatístico a ser aplicado deve ser considerado antes da implementação da seleção genômica em programas de melhoramento. Diante desse cenário, este estudo teve por objetivo avaliar a habilidade de predição de diferentes metodologias de seleção genômica aplicada a uma população de suínos de tipo Landrace para as características espessura de gordura e número de leitões nascidos vivos por leitegada. Os 969 animais (813 fêmeas e 156 machos) utilizados foram genotipados com um painel customizado da empresa Illumina contendo, aproximadamente, 57.692 SNPs. Os valores genéticos genômicos para as características espessura de gordura e número de leitões nascidos vivos foram preditos utilizando quatro métodos aplicados na seleção genômica: GBLUP - “Genomic Estimated Breeding Value”; “Single-step” GBLUP; brnn - “Bayesian Regularized Neural Network”; e snnR - “Sparse Neural Networks. As acurácias de predições foram melhores para a característica espessura de gordura em relação ao número de leitões nascidos vivos, com correlação média de 0.41 e 0.17 respectivamente. De maneira geral, os modelos lineares ssGBLUP e GBLUP foram os que apresentaram maiores acurácias, sendo que o ssGBLUP foi o que apresentou melhor habilidade de predição para ambas características. Os métodos de redes neurais brnn e snnR foram os que apresentaram o menor desempenho, sugerindo que outros estudos devem ser realizados para avaliar a habilidade de predição quando utilizadas diferentes estruturas de redes neurais.

Palavras-chave: brnn, espessura de gordura, GBLUP, número de leitões nascidos vivos, redes neurais, “single-step”, snnR

1 INTRODUÇÃO

A seleção genômica (GS) (Meuwissen et al., 2001) tem sido usada com sucesso em alguns programas de melhoramento de plantas e animais e em outros sua prática tem sido investigada (Henffner et al., 2011; Wang et al., 2013; Brito et al., 2017), sendo que na indústria de suínos esta apresenta grande potencial para aplicação (Guo et al., 2015; Knol et al., 2016). Isto porque a GS agrega informações para características que tem mensuração invasiva, que não são medidas em machos ou são tomadas tardiamente na vida do animal. Além disso, a GS permite a adição de informação, na avaliação genética, sobre a performance dos animais cruzados que não pode ser medida nas linhagens puras (Knol et al., 2016).

As pesquisas sobre aplicação da genômica na suinocultura começaram em 2009 com a disponibilização do painel “Illumina Porcine60K BeadChip” (San Diego, CA, EUA) com 62.163 marcadores do tipo SNP. Posteriormente, a GeneSeek (Lansing, MI, EUA) lançou o painel “Genomic Profiler for Porcine” (GGP-Porcine) que contém 70.231 SNPs. Atualmente há várias opções para genotipagem de suínos a exemplo do painel de alta densidade “Axiom Porcine Genotyping Array” (Santa Clara, CA, EUA) com 658.692 marcadores e a possibilidade da criação de um painel customizado. A customização apresenta a vantagem de possibilitar a inclusão de SNPs que podem ser lidos apenas pelo proprietário desta informação, e/ou marcadores patenteados associados às características de produção (Samorè e Fontanesi, 2016).

Assim como há grande disponibilidade de painéis de genotipagem, muitos modelos para aplicação da seleção genômica foram propostos. As diferenças entre as modelagens estatísticas nos estudos genômicos estão relacionadas com: tipo de relação entre as características-alvo e covariáveis explicativas (lineares ou não-lineares); tipo de informação genômica empregada (uso de toda informação ou um subconjunto reduzido de genótipos) e estratégias usadas para confrontar o problema da estimativa dos valores genéticos genômicos a partir de modelos que possuem grande quantidade de parâmetros com pequeno número de observações (Tussel et al., 2013).

Segundo Akanno et al. (2014), o desempenho da seleção genômica está associado a inúmeros fatores, sendo um destes a metodologia de predição dos valores genéticos genômicos utilizada. Portanto, a escolha do método estatístico a ser aplicado deve ser considerado antes da implementação da seleção genômica em programas de melhoramento, visando a aumentar a acurácia de predição. Diante desse cenário, o objetivo deste trabalho foi avaliar a habilidade de predição de diferentes metodologias de predição de valores genéticos genômicas aplicadas a uma população de suínos tipo Landrace para as características espessura de gordura e número de leitões nascidos vivos por leitegada.

2 MATERIAL E MÉTODOS

2.1 Descrição dos dados fenotípicos e genotípicos

Neste trabalho, utilizou-se dados de uma linhagem comercial tipo Landrace oriundos das colheitas rotineiras de oito granjas núcleo francesas, durante o período de 2003 a 2018. Características reprodutivas como número total de leitões nascidos vivos, nascidos mortos e desmamados foram tomadas na maternidade no momento do parto e desmame e digitalizadas no programa de gerenciamento das granjas. As características produtivas, como peso corporal, espessura de gordura e conformação de pernil foram mensuradas no teste de seleção realizado quando os animais estavam em média, com 157 dias de idade.

No presente trabalho foram estudadas as características número de leitões nascidos vivos por leitegada (NV) e espessura de gordura (EG), sendo analisados, respectivamente, registros de 87.058 leitegadas e 73.630 suínos testados. A medida EG foi colhida transversalmente entre a décima e a décima primeira costela, 4 cm abaixo da linha média, utilizando os aparelhos ultrassônicos Aloka ou DRAMINSKI 4Vet mini.

Um total de 969 animais, nascidos entre 2010 e 2018, foi genotipado utilizando um painel customizado da empresa Illumina (San Diego, CA, EUA) com 57.692 marcadores distribuídos por todo o genoma. Do total dos animais

genotipados, 813 eram fêmeas e 156 eram machos, tendo sido os mesmos escolhidos para serem representativos da população do programa de melhoramento genético a qual essas granjas núcleo pertencem, ou seja, foram escolhidos os animais com maior número de progênie. Dos animais genotipados, após a edição dos dados fenotípicos, verificou-se que 909 apresentavam informações para a medida espessura de gordura e 789 fêmeas tinham registros para NV somando 2.707 informações de leitegada. O arquivo de pedigree continha 879.965 animais distribuídos em 13 gerações equivalentes.

2.2 Edição dos dados

Análises preliminares foram realizadas utilizando o programa estatístico R para verificar a distribuição dos dados e eliminar possíveis inconsistências no arquivo. As médias para as características EG e NV foram de 13,0 mm e 13,3 leitões, respectivamente (Tabela 2).

Tabela 2. Média e mediana dos valores fenotípicos, desvios-padrão (DP), valores mínimos (Mín), valores máximos (Máx) e coeficiente de variação (CV) das características estudadas.

Característica	Média	Mediana	DP	Mín	Máx	CV (%)
EG (mm)	13,00	12,80	2,78	4,00	26,00	21,38
NV	13,32	14,00	3,41	1,00	26,00	25,60

EG = Espessura de gordura; NV = número de leitões nascidos vivos por leitegada.

Para minimizar a influência dos efeitos ambientais sobre os parâmetros genéticos foram criados grupos de contemporâneos (GC). Para a característica espessura de gordura, os GC foram constituídos por animais da mesma granja, mesmo sexo e animais testados no mesmo lote, ou seja, aqueles nascidos na mesma semana. Para a característica número de leitões nascidos vivos por leitegada, os GC foram formados considerando granja e lote do parto, sendo consideradas leitegadas nascidas na mesma semana como o mesmo lote de parto. Para ambas as características foram excluídos todos os GC com menos de 20 registros. Além disso, para NV foram excluídas todas as observações de reprodutores (pais) que apresentaram leitegadas em um único grupo de

contemporâneos e também todas as informações referentes a nona ordem de parto ou acima.

Ao final da edição de dados foram eliminados 191 registros de espessura de gordura e 17.553 observações de leitegada. Portanto, o banco de dados analisado continha 73.439 registros de EG, distribuídos em 979 grupos de contemporâneos e, 69.505 registros de NV divididos em 2.431 GC.

O controle de qualidade dos genótipos foi realizado utilizando o pacote “snpStats” (Clayton, 2013) no programa R (R Core Team, 2017) para filtrar: (i) SNPs que falharam em mais que 10% das amostras; (ii) SNPs em cromossomos não autossômicos ou sem posição definida; (iii) amostras que falharam em mais que 10% dos marcadores; (iv) SNPs com desvios significativos ($p < 0,00001$) do equilíbrio de Hardy-Weinberg; e (v) SNPs com frequência mínima (MAF - “Minor allele frequency”) menor que 1%. Depois do controle de qualidade, nenhuma amostra foi excluída e o número de SNPs remanescentes foi de 37.773.

2.3 Estimativas de parâmetros genéticos e predição do valor genético

As estimativas dos componentes de (co)variância, assim como das herdabilidades e os valores genéticos predito foram calculados utilizando o modelo animal unicaracterística e bicaracterística, aplicando o método de máxima verossimilhança restrita usando o programa AIREMLF90 (Misztal, 2012). O modelo empregado incluiu os efeitos fixos de grupo de contemporâneos, idade ao teste como covariável linear e quadrática para a característica EG, a ordem de parto como covariável linear para a característica NV e os efeitos genético aditivo, de ambiente permanente animal e residual como aleatórios. O efeito de ambiente permanente animal foi considerado apenas para a característica NV. E para as duas características as pressuposições para a análise de variância, ou seja normalidade do resíduo e a homogeneidade da vari Assim, o modelo estatístico proposto para a análise de NV pode ser assim apresentado:

$$y = Xb + Za + Wp + e$$

em que y é o vetor da variável dependente, X é a matriz de incidência dos efeitos fixos, b é o vetor de efeitos fixos, Z é a matriz de incidência dos efeitos aleatórios

genéticos diretos, a é o vetor de efeitos aleatórios para o efeito genético aditivo, W é a matriz de incidência do efeito aleatório de ambiente permanente animal, p é o vetor do efeito aleatório de ambiente permanente animal e e é o vetor de efeitos residuais. O mesmo modelo foi utilizado para a análise de EG com exclusão do efeito aleatório de ambiente permanente animal.

Após a análise, as estimativas de variâncias genéticas para as características EG e NV observadas foram de 3,54 e 1,20, respectivamente. As herdabilidades estimadas para as duas características analisadas foram de $0,54 \pm 0,011$ para EG e $0,11 \pm 0,008$ para NV.

Os valores genéticos preditos foram desregredidos (dEBV – “deregressing estimated breeding value”) utilizando o método proposto por Garrick et al. (2009), para serem usados como pseudo-fenótipo na predição dos valores genéticos genômicos (GEBV – “genomic estimated breeding value”) quando utilizados as metodologias GBLUP, “Bayesian Regularized Neural Network” e “Sparse Neural Networks”. O cálculo dos dEBVs foi realizado para evitar o viés nas análises de seleção genômica que poderia ser causado devido à alta variabilidade do número de progênie por animais genotipados e também ao fato de que nem todos os animais genotipados possuíam fenótipos. Enquanto que para o método “single-step” GBLUP, os fenótipos ajustados para o efeito fixo utilizados na predição dos GEBV, isto porque no mesmo são utilizadas as informações de animais genotipados e não genotipados.

2.4 Predição dos valores genéticos genômicos

Os valores genéticos genômicos para as características EG e NV foram preditos utilizando quatro métodos aplicados na seleção genômica: (i) GBLUP, que assume o modelo infinitesimal com infinito locos de pequenos efeitos; (ii) “single-step” GBLUP (ssGBLUP), proposto por Misztal et al. (2009), que integra os registros de pedigree, fenótipo e genótipo e baseia-se no modelo infinitesimal; e os métodos não lineares de redes neurais (iii) “Bayesian Regularized Neural Network” (brnn), descrito por Rodriguez et al. (2013) e (iv) o método desenvolvido por Wang et al. (2016) “Sparse Neural Networks” (snnR).

i) GBLUP

No GBLUP, a matriz de parentesco tradicional foi substituída por uma matriz de parentesco genômica (G) baseada nos SNPs dos animais. A matriz G foi calculada a partir dos dados genotípicos conforme descrito por VanRaden (2008):

$$G = \frac{(W - P)(W - P)'}{2 \sum_{j=1}^p p_j(1 - p_j)}$$

em que W é a matriz dos genótipos dos p SNPs codificados como -1, 0 e 1 para os n indivíduos. P é matriz $n \times p$ com duas vezes o alelo de menor frequência p_j , para $j = 1, 2, \dots, p$ replicado n vezes dentro de cada coluna. O modelo GBLUP é dado por:

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

em que y é o vetor dos dEBVs para cada característica, $X\beta$ é vetor de efeitos fixos e Zu é o vetor do efeito aleatório, em que u é o vetor que contém GEBV de todos os animais genotipados, sendo $u \sim N(0, G\sigma_u^2)$, em que σ_u^2 é a variância genética explicada pelos marcadores. Esse modelo foi implementado utilizando a função “kin.blup” do pacote “rrBLUP” (Endelman, 2011) no programa estatístico R.

ii) ssGBLUP

O modelo ssGBLUP combinou todas as observações fenotípicas, genotípicas e de parentesco em uma única análise. O ssGBLUP foi implementado substituindo a matriz A utilizada no BLUP pela matriz H , que inclui a informação genômica. A inversa da matriz H foi obtida de acordo com Aguilar et al. (2010) e Legarra et al. (2009):

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

em que, H^{-1} é a inversa da matriz de parentesco incorporada pela inversa da matriz de parentesco genômica (G^{-1}) e pela inversa do numerador da matriz de parentesco dos animais genotipados (A_{22}^{-1}). A matriz G foi construída de acordo com VanRaden (2008) e as análises foram realizadas com os programas PreGS, BlupF90 e Predict da família BLUPF90 (MISZTAL et al., 2012).

iii) brnn

O modelo de redes neurais brnn utilizado na análise ajustou uma rede de camadas como descrito por MacKay (1992) e Foresse e Hagan (1997), sendo que o mesmo é dado por:

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

em que, i é igual a 1, ..., n , s é o número de neurônios, w_k é o peso do k -th neurônio, b_k é o viés para o k -th neurônio, $\beta_j^{[k]}$ é peso da j -th entrada da rede, g_k é a função de ativação:

$$g_k(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1}$$

O algoritmo de Gauss-Newton foi implementado na análise para atualizar os pesos e aplicada a função objetivo a ser minimizada:

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

em que D é o arquivo de dado utilizado na análise (GEBV), w são os pesos e M é a estrutura da rede utilizada. A regularização bayesiana produz “shrinkage” das estimativas de parâmetros para reduzir sua variância, e, assim, a função objetivo se torna:

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

em que $E_w(w|M)$ é a soma dos quadrados dos pesos da rede e α e β são os parâmetros de regularização positivos.

As análises para as características EG e NV foram realizadas no programa R (R Core Team, 2017) utilizando o pacote “brnn” (Rodriguez e Gianola, 2018). Nestas foi usado o modelo descrito anteriormente considerando 3 camadas com 15, 12 e 10 neurônios para a primeira, segunda e terceira camada, respectivamente.

iv) snnR

O modelo de redes neural snnR, assim como brnn foi aplicado utilizando o pacote “snnR” desenvolvida por Wang et al. (2016). Os mesmos parâmetros aplicados no modelo brnn também foram usados para esse modelo, ou seja, 3

camadas de redes com 15, 12 e 10 neurônios em cada camada. A diferença entre os dois modelos de redes neurais está no algoritmo utilizado para minimizar o aumento do erro dos mínimos quadrados.

2.5 Acurácia dos valores genéticos genômicos

O método de validação cruzada foi usado para testar a desempenho destes modelos dentro de cada uma das características estudadas. Portanto, o conjunto de dados foi dividido aleatoriamente em 4 partes iguais. A partir dessa divisão os efeitos de marcador SNP foram estimados com base no conjunto de 3 partes dos dados (75%), denominado de população de treinamento, e a habilidade de predição foi medida no restante dos animais (25%), denominada de população teste. Esse ciclo foi repetido quatro vezes rotacionando as 4 partes em treinamento e testes, com cada parte sendo usada para testar apenas uma vez e para treinar três vezes.

A capacidade de predição de cada modelo para cada conjunto de dados foi avaliada, utilizando a população teste, por meio da correlação de Pearson entre o valor genético genômico estimado (GEBV) e o valor genético estimado desregredido (dEBV), ou apenas o fenótipo corrigido para efeitos fixos, no caso da avaliação do modelo ssGBLUP. Também foi utilizada a medida denominada de quadrado médio do erro de predição (MSE - "Mean Squared Error") para verificar a qualidade do ajuste do modelo aplicado, sendo que, quanto menor o valor de MSE melhor é o ajuste do modelo.

Além disso, para avaliar o viés das predições genômicas, foi utilizado o seguinte modelo de regressão:

$$y = b_0 + b_1x + e$$

Em que y são dEBVs (ou fenótipo ajustado para ssGBLUP), b_0 é o intercepto, b_1 é o coeficiente de regressão linear indicando viés de predições (coeficiente de regressão igual a 1 indica que estimativas não são inflacionadas ou deflacionadas e que as predições apresentam escala similar às das variáveis dependentes), x são os GEBVs e e o resíduo.

3 RESULTADOS E DISCUSSÃO

As estimativas de correlação de Pearson entre os dEBVs (ou fenótipo corrigido para efeitos fixos) e os GEBVs variaram de 0,30 a 0,49 para a característica EG e de 0,11 a 0,20 para NV. As estimativas do quadrado médio do erro de predição variaram de 2,88 a 5,0 e 9,59 a 38,12 para EG e NV, respectivamente.

Tabela 3. Acurácia de predição medida pela correlação de Pearson (r) entre o dEBV (ou fenótipo corrigido) e o GEBV, quadrado médio do erro de predição (MSE) e coeficiente de regressão (b_1) do dEBV (ou fenótipo corrigido) sobre o GEBV, de acordo com os métodos testados.

Características		Métodos			
		ssGBLUP	GBLUP	brnn	snnR
R	EG	0,48	0,49	0,30	0,38
	NV	0,20	0,20	0,18	0,11
MSE	EG	2,88	3,65	3,45	5,00
	NV	13,07	15,02	9,59	38,12
b_1	EG	0,93	1,08	0,04	0,37
	NV	0,57	1,16	0,05	0,08

EG: espessura de gordura; NV: número de leitões nascidos vivos por leitegada.

As acurácias de predição foram melhores para a característica EG em relação a medida NV. Esse resultado corrobora com o descrito na literatura, uma vez que o aumento da estimativa de herdabilidade melhora a acurácia dos GEBVs (Calus, 2010). Apesar das baixas acurácias do GEBV para NV, vários autores têm argumentado que a seleção genômica é indicada para aumentar a taxa de ganho genético para características com estimativas de herdabilidade baixas se um número suficiente de indivíduos estiver disponível para treinar o modelo (Ibañez-Escriche e Gonzalez-Recio, 2011). Além disso, o uso de informações genômicas permite obter predições muito mais cedo na vida dos animais quando comparado ao uso de informações fenotípicas. Portanto, predições genômicas permitem realizar seleção dentro de leitegada, enquanto que com a seleção tradicional isso não é possível já que todos os irmãos apresentam o mesmo valor genético estimado (EBV = a média dos pais) (Tusell et al., 2013; Grossi et al., 2018).

De modo geral, os modelos lineares (ssGBLUP e GBLUP) aplicados no estudo de predição genômica dessa população foram os que apresentaram melhor

habilidade de predição quando comparadas as estimativas de correlação, MSE e coeficiente de regressão tanto na seleção para EG como para a seleção baseada no NV. Ainda que as estimativas de correlação entre os métodos lineares testados tenham sido praticamente iguais, o método ssGBLUP apresentou menores MSE para EG e NV, indicando que o modelo demonstrou melhor ajuste nas predições para população estudada. Portanto, pode-se dizer que o método ssGBLUP foi o que apresentou melhor habilidade de predição para as características EG e NV na população estudada.

Alguns trabalhos encontrados na literatura mostraram resultados semelhantes aos deste estudo, como Guo et al. (2015), que utilizando informações de número de leitões nascidos totais de 3.445 animais genotipados da raça Landrace e 3.372 animias genotipados da raça Yorkshire, também relataram superioridade do método ssGBLUP quando comparado ao método BLUP e ao método GBLUP. Esses resultados devem estar relacionados ao fato de o método ssGBLUP utilizar todas as informações disponíveis (animais genotipados e não genotipados) para predizer os GEBVs. Com esse método, os fenótipos dos animais não genotipados também fornecem informação por meio da matriz de parentesco, que combina tanto informações baseadas no pedigree quanto nos genótipos. Além disso, a predição genômica deve ser mais acurada quando as observações do pedigree são incluídas no modelo, pois nem toda a variância genética aditiva pode ser estimada pelos SNPs, uma vez que a habilidade de estimação pode depender da densidade do painel e de outros fatores, como estrutura genética, herdabilidade e métodos estatísticos (Zhang et al., 2019).

Os métodos de redes neurais foram os que apresentaram o menor desempenho tanto para a correlação entre o dEBV e GEBV, quanto para o quadrado médio do erro e o coeficiente de regressão. Contudo, trabalhos realizados com outras espécies de animais têm demonstrado superioridade dos métodos não lineares de redes neurais, a exemplo do brnn, em comparação aos métodos tradicionais lineares (Gianola et al., 2011; Okurt et al., 2011). Tusell et al. (2013), avaliando tamanho da leitegada para duas linhagens puras de suínos Landrace e uma linhagem oriunda dos cruzamentos entre as mesmas, também encontraram menores estimativas de acurácia para o método brnn quando comparado aos

métodos lineares. Porém, quando estimados os valores genéticos genômicos da linhagem cruzada, o método de redes neurais foi que apresentou melhor acurácia. Os autores sugerem que a melhor habilidade de predição encontrada para a população cruzada pode ser ligada aos efeitos genéticos não aditivos, como o efeito de dominância esperado em função do cruzamento. Os autores ainda acrescentam que os resultados encontrados na literatura indicam que modelos de redes neurais, os quais são mais flexíveis a sistemas não lineares, podem predizer fenótipos mais complexos de forma mais acurada.

O método *snnR* apresentou pior resultado em acurácia para ambas as características quando comparado ao outro método de redes neurais *brnn*, ao contrário do descrito na literatura (Wang et al., 2016). A correlação de Pearson entre o *dEBV* e o *GEBV* relatada no presente trabalho com o método *snnR* para a característica *NV* foi a menor, e o método também foi o que apresentou maiores *MSE* para as duas características estudadas. Wang et al. (2016), utilizando um banco de dados com informações de bovinos leiteiros da raça Jersey, demonstraram que o método *snnR* apresentou melhores correlações entre o valor observado e o predito para a característica produção de leite, em todas as arquiteturas de redes neurais testadas, quando comparado ao método *brnn*. Além disso, o modelo *snnR* com três camadas e cinco neurônios em cada camada para cada um dos componentes aditivo e de dominância, foi que apresentou estimativas mais precisas, inclusive quando comparado com o *ssGBLUP*. A estrutura de rede neural utilizada é um fator que pode ter contribuído para as diferenças entre os resultados do presente estudo e os encontrados por Wang et al. (2016). Talvez a escolha de utilizar mais neurônios por camadas tenha favorecido o método *brnn*, em detrimento do *snnR* no presente trabalho. Portanto, os resultados sugerem que novos estudos devem ser conduzidos com os métodos de redes neurais para avaliar a habilidade de predição quando utilizadas diferentes estruturas de redes.

4 CONCLUSÕES

Os melhores resultados de predição dos valores genéticos genômicos para a população estudada foram obtidos com o método “single-step” GBLUP, tanto para a característica espessura de gordura quanto para o número de leitões nascidos vivos por leitegada.

5 REFERÊNCIAS

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal Dairy Science** 93: 743–752.

Akanno EC, Schenkel FS, Sargolzaei M, Friendship RM, Robinson JA (2014) Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. **Journal of Animal Breeding and Genetics** 131: 367–378.

Brito LF, Clarke SM, McEwan JC, Miller SP, Pickering NK, Bain WE, Dodds KG, Sargolzaei M and Schenkel FS (2017) Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. **BMC Genetics** 18:7.

Calus MPL (2010) Genomic breeding value prediction: methods and procedures. **Animal** 4(2): 157-164.

Clayton D (2013). snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.12.0.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome** 4:250-255.

Foresee FD, Hagan MT (1997) "Gauss-Newton approximation to Bayesian regularization" **Proceedings of the 1997 International Joint Conference on Neural Networks**.

Garrick DJ, Taylor JF and Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution** 41:55.

Gianola D, Hayrettin O, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics** 12: 87.

Grossi DA, Brito LF, Jafarikia M, Schenkel FS, Feng Z (2018) Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. **Animal** 12: 2235-2245.

Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G (2015) Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. **Journal of Animal Science** 93: 503–512.

Heffner EL, Jannink J, Sorrells ME (2011) Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. **The Plant Genome** 4:65-75.

Ibañez-Escriche N, Gonzalez-Recio O (2011). Review. Promises, pitfalls and challenges of genomic selection in breeding programs. **Spanish Journal of Agricultural Research** 9: 404–413.

Knol EF, Nielsen B, Knap PW (2016) Genomic selection in commercial pig breeding. **Animal Frontiers** 6:15–22.

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science** 92: 4656–4663.

Mackay DJC (1992) Bayesian interpolation. **Neural Computation** 4(3): 415-447.

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

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2009) “BLUPF90 and related programs (BGF90)” **Proceedings of the 7th World Congress on Genetics Applied to Livestock Production**. Montpellier: INRA, 21–22p.

Misztal, I. BLUPF90 - a flexible mixed model program in Fortran 90. Manual. University of Georgia. 2012

Okut H, Gianola D, Rosa GJM, Weigel KA (2011) Prediction of body mass index in mice using dense molecular markers and a regularized neural network. **Genetics Research** 93: 189–201.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Rodriguez PR, Weigel KA, Gianola D, Rosa GJM (2013) Technical Note: An R package for fitting Bayesian regularized neural networks with applications in animal breeding. **Journal of Animal Science** 91:3522–3531

Rodriguez PR, Gianola D (2018). brnn: Bayesian Regularization for Feed-Forward Neural Networks. R package version 0.7. Disponível em: <<https://CRAN.R-project.org/package=brnn>>

Samorè AB, Fontanesi L (2016) Genomic selection in pigs: state of the art and perspectives. **Italian Journal of Animal Science** 15: 211-232.

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

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

Wang C, Habier D, Peiris BL, Wolc A, Kranis A, Watson KA, Avendano S, Garrick DJ, Fernando RL, Lamont SJ and Dekkers JCM (2013). Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. **Poultry Science** 92:1712–1723

Wang Y, Mi X, Rosa GJM, Chen Z, Lin P, Wang S, Bao Z (2016) Technical note: an R package for fitting sparse neural networks with application in animal breeding. **Journal of Animal Science** 96(5): 2016-2026

Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. **Frontiers in Genetics** 10: 189

CAPÍTULO 3 – Predição dos valores genéticos genômicos utilizando conjuntos de SNPs identificados com o método “Random Forest”.

RESUMO – Alguns trabalhos têm demonstrado que uma grande quantidade de SNPs presentes nos painéis de alta densidade usados na seleção genômica apresentam pouco ou nenhum efeito sobre os fenótipos. Assim, torna-se de interesse o estudo sobre o uso de conjunto de SNPs com relevância biológica na aplicação da seleção genômica para assim reduzir o custo de genotipagem com o uso de painéis de SNPs customizados de baixa densidade. Logo, o objetivo deste trabalho foi utilizar o método “Random Forest” para selecionar conjuntos de marcadores mais importantes na explicação dos fenótipos para características espessura de gordura (EG) e número de leitões nascidos vivos por leitegada (NV) e verificar o impacto do uso desses conjuntos de SNPs na acurácia dos valores genômicos preditos. Os 969 animais (813 fêmeas e 156 machos) utilizados foram genotipados com um painel customizado da empresa Illumina com 57.692 marcadores e após o controle de qualidade restaram 37.773 SNPs. Análises de “Random Forest” foram realizadas para determinar a importância dos marcadores para as características estudadas e a partir dessas análises foram selecionados os 3.000 SNPs mais importantes para EG (cenário 1), 3.000 SNPs mais importantes para NV (cenário 2) e os 3.000 SNPs mais importantes para as duas características em conjunto (cenário 3). Foram formados também um cenário com 3.000 SNPs selecionados aleatoriamente (cenário 4) e um quinto cenário com todos os SNPs (37.773 SNPs). Depois de determinados os cenários, a habilidade de predição dos valores genéticos genômicos foi avaliada com o método de “Single-step” GBLUP. O cenário 2 foi que apresentou melhor habilidade de predição na seleção genômica para NV. Portanto, os resultados do estudo indicam que os métodos de aprendizagem de máquina podem ser utilizados para selecionar eficientemente conjuntos de marcadores para formarem painéis de genotipagem de baixa densidade e assim reduzir os custos da implementação da seleção genômica.

Palavras-chave: Aprendizado de máquina, espessura de gordura, Landrace, número de leitões nascidos vivos, “single-step” GBLUP

1 INTRODUÇÃO

O desenvolvimento de estratégias de genotipagem como polimorfismo de nucleotídeo único (SNP, "single nucleotide polymorphisms") e genotipagem por sequenciamento (GBS, "genotyping-by-sequencing") tornaram a aplicação da seleção genômica (GS, "Genomic Selection") possível, oferecendo assim novos procedimentos para melhorar a eficiência dos programas de melhoramento genético de animais e plantas (Tusell et al., 2013). Crossa et al. (2010), por exemplo, estudando um conjunto de dados com quatro variedades de trigos, observaram que a inclusão de marcadores superou significativamente (de 8 a 36%) um modelo de predição baseado em pedigree. No melhoramento animal, a GS contribui para aumentar a eficiência de seleção nos programas de melhoramento de bovinos leiteiros. Segundo Hayes et al. (2009), o aumento da confiabilidade do valor genético genômico em relação ao valor genético para bezerros sem registros de filhas variou de 2 a 20%.

No entanto, para algumas espécies outras questões fazem com que os benefícios da seleção genômica não sejam completamente aproveitados. As principais razões que tem dificultado aplicação da GS nos programas de melhoramento de diferentes espécies são: (i) baixa qualidade dos genomas de referências; (ii) falta de programas de melhoramento genético com aplicabilidade para a seleção genômica; e (iii) alto custo de implementação associado com a necessidade de genotipagem de um grande número de animais para a população de referência (Xiang, 2015; Li et al., 2018).

Outro ponto que tem sido levantado, associado ao alto custo de genotipagem de um grande número de animais, está relacionado com o fato que uma grande quantidade de SNPs presentes nos painéis de alta densidade usados na GS apresentam pouco ou nenhum efeito sobre os fenótipos (MacLeod et al., 2016). E isso tem levantado a questão sobre usar conjuntos de SNPs que apresentem impacto relevante na função biológica de características de interesse e assim, reduzir o custo de genotipagem com o uso de painéis de SNPs customizados de baixa densidade.

Muitos estudos tem sido realizados para avaliar a aplicação de métodos de aprendizado de máquina para seleção de SNPs com a finalidade de reduzir os painéis de SNPs de alta densidade em painéis de SNPs com um conjunto menor de marcadores, e assim utilizar esses painéis de menor densidade nos estudos de associação ampla do genoma (Liang e Kelemen, 2008; Walters et al., 2012) e da GS (Long et al., 2011; Bermingham et al., 2015). Entre as metodologias de aprendizado de máquina, um dos métodos que tem apresentado melhores resultados em estudos genômicos é o “Random Forest” (RF) (Li et al., 2018).

O método RF é composto por um conjunto de muitas estruturas (árvores), em que as mesmas são construídas a partir de amostras “bootstrap” não paramétricas dos dados originais. No RF, as árvores crescem de forma independente e as predições são calculadas para reduzir a variância do erro. Em vez de encontrar a melhor regra de divisão em um nó da árvore usando todas as variáveis de predição, o RF seleciona em cada nó de cada árvore, um subconjunto aleatório de variáveis que é usado como candidatos para encontrar a melhor regra de divisão para o nó. A intenção disso é desvincular as árvores de modo que a média sobre o conjunto da floresta terá uma menor variância. Esse método também pode ser usado para selecionar e classificar variáveis por meio de diferentes medidas de importância, o que o torna uma ferramenta importante para análise de dados genômicos (Breiman, 2001; Chen e Ishwaran, 2012; Touw et al., 2012).

Diante deste cenário, o objetivo proposto para este capítulo foi utilizar o método RF para selecionar conjuntos de marcadores mais importantes na explicação dos fenótipos para a espessura de gordura e o número de leitões nascidos vivos por leitegada e verificar o impacto do uso desses conjuntos de SNPs na acurácia dos valores genômicos preditos para essas duas características de importância econômica na suinocultura.

2 MATERIAL E MÉTODOS

2.1 Dados fenotípicos e genotípicos

Um total de 969 animais de uma linhagem tipo Landrace pertencentes a granjas núcleo francesas foram genotipados utilizando um painel customizado com 57.692 marcadores distribuídos por todo o genoma. A população genotipada foi constituída de 813 fêmeas e 156 machos nascidos entre 2010 e 2018. O controle de qualidade dos genótipos foi realizado para excluir SNPs e amostras com “call rate” menor que 90%, SNPs em cromossomos não autossômicos ou sem posição definida, SNPs com desvios significativos ($p < 0,00001$) do equilíbrio de Hardy-Weinberg e SNPs com MAF menor que 1%. Após controle de qualidade, um total de 37.773 SNPs foram utilizados nas análises.

O arquivo de pedigree continha 879.965 animais distribuídos em 13 gerações. O arquivo de fenótipos era composto por 73.439 observações de espessura de gordura (EG) e 69.505 registros de número de leitões nascidos vivos por leitegada (NV), oriundos de 979 e 2.431 grupos de contemporâneos, respectivamente. A EG foi colhida transversalmente entre a décima e a décima primeira costela a 4 cm abaixo da linha média utilizando os aparelhos ultrassônicos Aloka ou DRAMINSKI 4Vet mini. Essa medida foi tomada no momento do teste de seleção, aos 157 dias de idade, aproximadamente.

2.1 Cenários utilizados na seleção de SNPs

No presente trabalho, 70% dos animais mais velhos da população genotipada (678 animais) foram escolhidos para formarem a população utilizada nas análises de aprendizado de máquina. Foram definidos 5 cenários de acordo com dois critérios. Para a formação dos três primeiros cenários, o número de fenótipos utilizados como resposta da análise variou de acordo com a quantidade de registros existentes para cada característica.

O cenário 1 foi realizado para selecionar os SNPs com maior importância para a característica EG. Desta forma, as informações de espessura de gordura

dos animais genotipados foram utilizadas como resposta e os dados dos marcadores foram considerados os preditores para a realização da análise. Neste cenário, a população era composta pelos 618 animais mais velhos que apresentavam observações para a característica EG.

Na seleção dos SNPs mais importantes para a característica NV, que representava o cenário 2, o registro de NV foi considerado a resposta e a observação de ordem de parto juntamente com as informações dos SNPs entraram como preditores nas análises de RF. Para este cenário a população era composta de 547 fêmeas com 2.272 fenótipos.

A seleção dos marcadores mais importantes para as características espessura de gordura e número de leitões nascidos vivos em conjunto formaram o cenário 3. Neste cenário, foi considerado como fenótipo de entrada para as análises de RF o componente principal um (PCA1) da análise de componentes principais entre os fenótipos de EG e NV, que representava 67% da variabilidade. O cenário 3 era composto por 2.244 informações de fenótipos de 543 animais.

Além dos três cenários descritos anteriormente, para fins de comparação, outros dois cenários também foram criados. O cenário 4 foi montado selecionando 3.000 marcadores aleatoriamente, enquanto que o cenário 5 foi composto de todos os SNPs que passaram pelo controle de qualidade, ou seja, 37.773 SNPs.

2.3 “Random Forest”

Nas análises de RF, dois parâmetros determinam os resultados da mesma, sendo estes o número de árvores na floresta, denominado de *Ntree* e o número de marcadores presentes em cada conjunto de variáveis usados para formar a árvore, conhecido por *mtry*. Com o intuito de determinar os melhores valores para esses parâmetros na população estudada, foi avaliado o impacto de variar os valores de *Ntree* e *mtry* sobre a média dos valores do quadrado médio do erro (MSE) de todos os SNPs.

Portanto, dez valores de *Ntree* foram variados entre 500 e 5000 árvores com intervalo de 500 árvores. Para o parâmetro *mtry* foram testados quatro valores: (i) 1; (ii) raiz quadrada de M; (iii) duas vezes a raiz quadrada de M; e (iv) M vezes 0,1,

em que M é igual a número de marcadores (37.773), conforme proposto por Li et al. (2018). Os valores dos parâmetros foram determinados quando o valor médio de MSE de todos os SNPs foi menor. Então, esses parâmetros foram utilizados para as análises subsequentes (Tabela 4). Essas análises foram realizadas utilizando o pacote “randomForest” (Liaw e Wiener, 2002), aplicada no programa computacional R (R Core Team, 2017).

Tabela 4. Valores dos parâmetros utilizados nas análises de “Random Forest” que determinaram o menor MSE para os três primeiros cenários estudados.

Cenário	<i>Ntree</i>	<i>Mtry</i>
1	3.000	0,1 x M
2	4.500	1
3	3.000	1

Cenário 1: 3.000 SNPs selecionados para EG; Cenário 2: 3.000 SNPs selecionados para NV; Cenário 3: 3.000 SNPs selecionados para EG e NV. EG = espessura de gordura; NV = número de leitões nascidos vivos por leitegada.

2.4 Importância dos marcadores

Uma vez determinados os valores dos parâmetros *Ntree* e *mtry* (Tabela 4) novas análises de RF foram realizadas para determinar a importância dos marcadores para cada cenário (montados com apenas 70% dos animais mais velhos da população), em que a importância dos marcadores é determinada pelo valor do %IncMSE que indica o aumento do quadrado médio do erro quando um marcador é permutado aleatoriamente, desta maneira os marcadores mais importantes são aqueles que apresentam maior valores para essa medida. Baseado nesses valores de importância, todos os SNPs foram ranqueados dos mais importantes até o que apresentou o menor valor de importância. Posteriormente, os 3.000 SNPs mais importantes foram identificados para cada cenário, ou seja, foram formados grupos com 3.000 SNPs mais importantes para a espessura de gordura, número de leitões nascidos vivos e a combinação das duas características. A escolha do número de SNPs selecionados foi determinada de acordo com o padrão de montagem das plataformas de genotipagem de diferentes laboratórios.

2.5 Seleção genômica utilizando grupo de marcadores

Depois de eleitos os 3.000 SNPs de maior importância para cada cenário, foi avaliada a habilidade de predição dos valores genéticos genômicos desses SNPs para as características EG e NV por leitegada com o método de “Single-step” GBLUP (ssGBLUP). A predição dos valores genéticos genômicos foi realizada para os cinco cenários desenvolvidos.

Na aplicação do método ssGBLUP foi utilizada como população de treinamento os 70% animais mais velhos, os quais formaram a população para as análises RF e, o remanescente dos animais mais jovens (291) foram utilizados como população de teste. O modelo ssGBLUP combinou todas as observações fenotípicas, genotípicas e de parentesco em uma única análise. O seguinte modelo estatístico bicaracterística foi considerado para as características estudadas:

$$y = Xb + Za + Wp + e$$

em que y é o vetor de observações fenotípicas; X é a matriz de incidência dos efeitos fixos; b é o vetor de efeitos fixos, incluindo o grupo de contemporâneo, idade ao teste como covariável linear e quadrática para a característica EG e ordem de parto como covariável para a característica NV; Z é a matriz de incidência dos efeitos aleatórios genéticos diretos; a é o vetor de efeitos aleatórios para o efeito genético aditivo; W é a matriz de incidência do efeito aleatório de ambiente permanente animal; p é o vetor do efeito aleatório de ambiente permanente animal e e é o vetor de efeitos residuais. O efeito de ambiente permanente animal foi considerado apenas para a característica número de leitões nascidos vivos por leitegada.

O ssGBLUP foi implementado substituindo a matriz A utilizada no BLUP pela matriz H , que inclui a informação genômica. A inversa da matriz H foi determinada de acordo com Aguilar et al. (2010) e Legarra et al. (2009):

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

em que, H^{-1} é a inversa da matriz de parentesco incorporada pela inversa da matriz de parentesco genômica (G^{-1}) e pela inversa do numerador da matriz de parentesco dos animais genotipados (A_{22}^{-1}). A matriz G foi construída de acordo com

VanRaden (2008) e as análises foram processadas nos programas da família BLUPF90 (Miszta et al., 2012).

2.6 Habilidade de predição dos valores genéticos genômicos

Para testar o desempenho de cada cenário utilizado na seleção genômica para cada uma das características estudadas, a capacidade de predição dos mesmos foi avaliada por meio da correlação de Pearson (entre o fenótipo ajustado e o GEBV), do quadrado médio do erro e do coeficiente de regressão. Para determinar o coeficiente de regressão foi utilizado o seguinte modelo de regressão:

$$y = b_0 + b_1x + e$$

em que y são os fenótipos ajustados, b_0 o intercepto, b_1 o coeficiente de regressão linear indicando viés de predições, x são os GEBVs e e o resíduo.

4 RESULTADOS E DISCUSSÃO

O diagrama de Venn (Figura 1) construído com os conjuntos de SNPs selecionados pela metodologia “Random Forest” para cada cenário considerando as características EG, NV por leitegada, as duas características em conjunto e 3.000 marcadores selecionados aleatoriamente, cenários 1, 2, 3 e 4 respectivamente, revelou que entre esses quatro cenários há apenas dois SNPs em comum. Além disso, a combinação dos cenários aos pares demonstrou que os dois cenários com mais marcadores importantes em comum foram os cenários 2 e 3, apresentando 314 SNPs iguais.

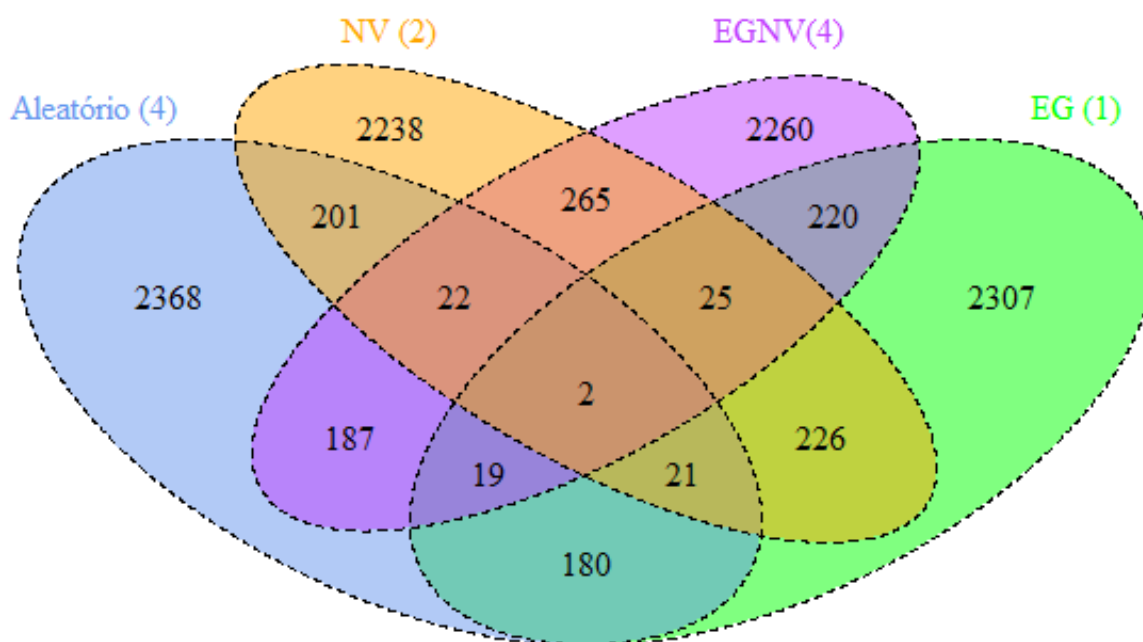


Figura 1. Diagrama de Venn demonstrando os SNPs em comum entre os 3.000 SNPs mais importantes de cada cenário estudado. Aleatório: 3.000 SNPs selecionados aleatoriamente; NV: 3.000 SNPs mais importantes para característica NV; EGNV: 3.000 SNPs mais importantes para as características NV e EG; EG: 3.000 SNPs mais importantes para características EG.

Quando analisadas as posições no genoma dos SNPs de maior importância determinados pelo método RF para as características EG, NV e para as duas características em conjunto (Figura 2), foi observado que os marcadores mais importantes para a característica NV estão dispersos por todo o genoma, enquanto que para a característica EG pode-se notar SNPs com importância consideravelmente maior nos cromossomos 1, 12 e 13. Quando verificada a importância dos SNPs ao longo do genoma considerando EG e NV em conjunto, nota-se que assim como para NV, os marcadores com maior importância estão relativamente espaçados por todo o genoma.

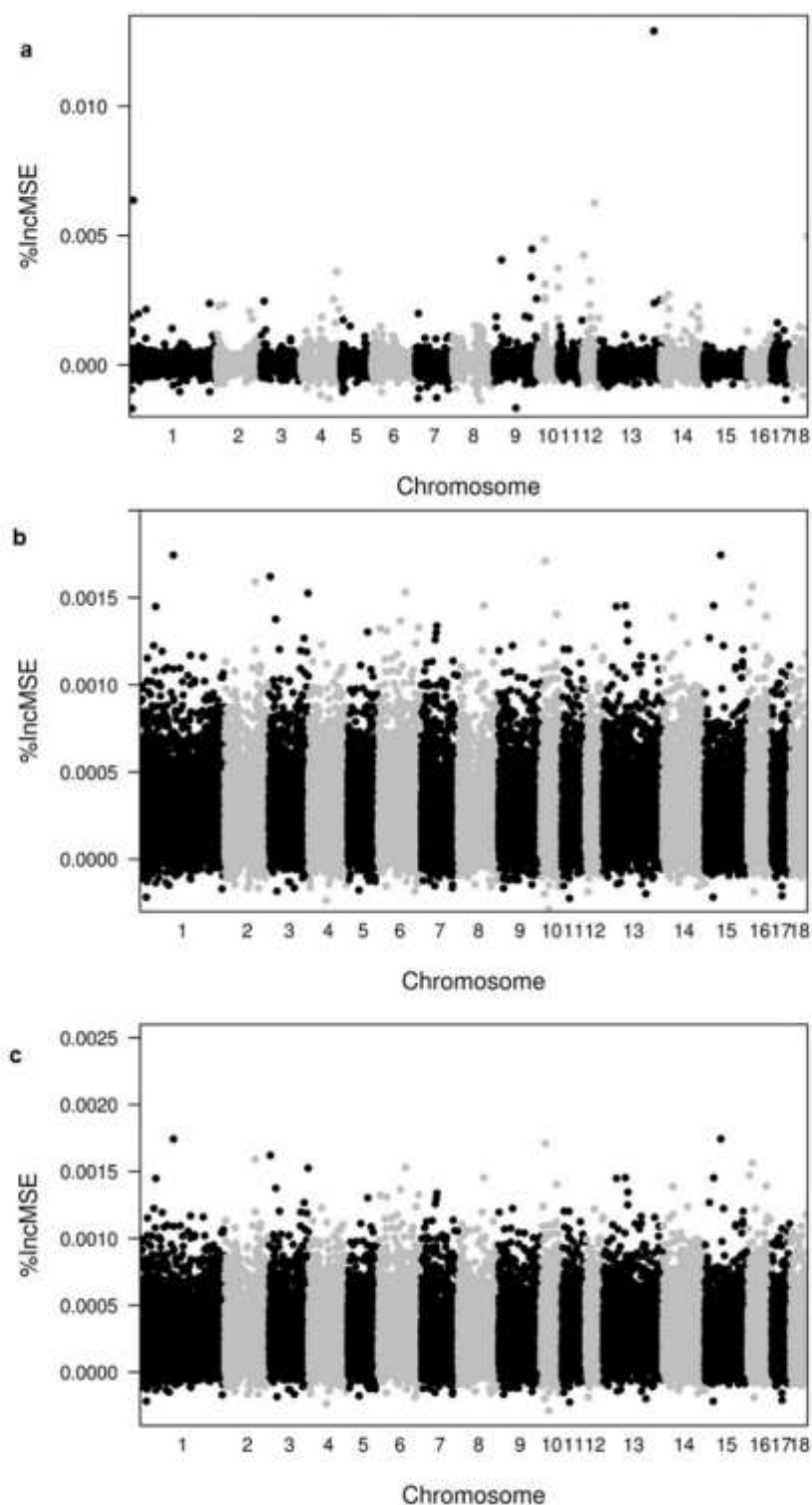


Figura 2: Valores de importância dos SNPs (%IncMSE) ao longo do genoma suíno. (a) Valores de importância determinado pelo método “Random Forest” para a característica espessura de gordura (cenário 1); (b) valores de importância para número de leitões nascidos vivos por leitegada (cenário 2); e (c) valores de importância determinados quando considerado as duas características em conjunto (cenário 3).

Em um trabalho de associação genômica ampla aplicado à característica EG em fêmeas do cruzamento Landrace x Large White, ZHU et al. (2014) encontraram que os quatro SNPs mais significantes, ou seja os marcadores com os menores P valor, para a característica estavam localizados no cromossomo 1, sendo que o SNP mais significativo foi localizado cerca de 528 kb próximo do gene MC4R. Segundo os autores, o MC4R é um receptor acoplado à proteína G que tem sido relacionado com a mediação do efeito da leptina na ingestão de alimentos e no balanço energético e também já foi anteriormente, na literatura, associado à deposição de gordura em suínos. O resultado indica que o método de RF pode ser uma alternativa adequada para a seleção de conjuntos de SNPs para características de interesse econômico, uma vez que um dos três marcadores de maior importância para a característica na população estudada também estava localizado no cromossomo 1. Além disso, em um estudo de seleção genômica ampla de bovinos da raça Canchim, Mokry et al. (2013), utilizando o método RF, identificaram um conjunto de SNPs correlacionados com aproximadamente 50% da variação do dEBV para a característica EG. Sendo assim, os autores também concluíram que é possível incluir esses SNPs no desenvolvimento de um painel de SNP de baixa densidade para implementação futura do programa de seleção genômica da raça Canchim.

Um fator importante que a ser considerado na interpretação dos resultados deste capítulo, é o fato de que a população estudada foi genotipada com um painel customizado que foi desenvolvido para a seleção de uma linhagem materna e, portanto, um painel com maior concentração de SNPs para características reprodutivas já que esse é o objetivo principal de seleção dessas linhagens. Desta forma, justifica-se que no cenário em que a importância dos SNPs foi determinada considerando as informações de fenótipo para NV (Figura 2) não houve variação muito grande no valor da importância dos marcadores. Isso pode ser explicado pelo fato de que o painel pode ter sido montado com uma grande quantidade de marcadores relacionados com características de prolificidade. Além disso, como na seleção das linhagens maternas também são consideradas características de produção, na seleção dos SNPs para comporem o painel de genotipagem também podem ter sido incluídos SNPs associados com estas características, a exemplo da

EG, justificando desta forma porque na análise de RF foram encontrados SNPs com relativa importância para a característica EG.

Avaliando os resultados das diferentes medidas utilizadas para medir a habilidade de predição dos valores genéticos genômicos para característica EG observou-se que o cenário criado com a informação de todos os SNPs (Cenário 5) foi o que apresentou a melhor habilidade de predição, uma vez que as estimativas de acurácia de predição medidas por meio da correlação, do quadrado médio do erro e do coeficiente de regressão foram melhores para este cenário quando comparado aos demais (Tabela 5). Porém, quando avaliado os resultados para a característica NV, verifica-se que a correlação apresentou o mesmo resultado para o cenário com todos os SNPs e para o cenário com 3.000 marcadores selecionados para NV (cenário 2). Na avaliação do quadrado médio do erro e do coeficiente de regressão pode-se dizer que para a característica NV os melhores resultados de predição foram obtidos com o cenário em que foram utilizados apenas as informações dos 3.000 SNPs mais importantes para NV por leitegada.

Tabela 5. Acurácia média de predição medida pela correlação de Pearson (r) entre o fenótipo corrigido e o GEBV, pelo quadrado médio do erro de predição (MSE) e coeficiente de regressão (b_1) do fenótipo ajustado sobre o GEBV para cada cenário montado.

Característica	Cenário	r (y^* , GEBV)	MSE	b_1
EG	1	0,43	2,90	0,87
	2	0,42	2,93	0,90
	3	0,43	2,91	0,91
	4	0,41	2,97	0,87
	5	0,47	2,80	0,97
NV	1	0,27	13,25	0,98
	2	0,28	13,17	1,07
	3	0,26	13,37	1,01
	4	0,26	13,32	1,01
	5	0,28	13,22	1,10

Cenário 1: 3.000 SNPs selecionados para EG; Cenário 2: 3.000 SNPs selecionados para NV; Cenário 3: 3.000 SNPs selecionados para EG e NV; Cenário 4: 3.000 SNPs selecionados aleatoriamente; Cenário 5: todos os SNPs (37.773).

De acordo com Li et al. (2018), na predição dos valores genéticos genômicos, a inclusão de maiores quantidades de informação de SNPs no modelo

não necessariamente significa melhor acurácia. Isto porque, uma grande quantidade de marcadores pode provocar mais erros de predição do que um pequeno número de SNPs que capturariam os efeitos principais de SNPs individuais, das interações SNP-SNP e dos relacionamentos não lineares entre SNPs. Esse fato corrobora com o observado para a seleção genômica da característica NV no presente trabalho, uma vez que o uso de um conjunto de SNPs na seleção genômica proporcionou melhores resultados de predição, avaliando a medida do quadrado médio do erro e coeficiente de regressão, quando comparado com a seleção com base na informação de todos os marcadores. Os autores também destacam que uma das vantagens do método “Random Forest” é o fato de que na listagem da importância dos marcadores, os valores negativos de importância indicam os SNPs não informativos, devendo, portanto, serem excluídos da predição genômica para reduzir os erros de predição. Essa vantagem pode ser utilizada para interpretar os resultados da seleção genômica para característica NV, visto que na seleção de apenas 3.000 SNPs com importância para a característica esses marcadores com valores negativos de importância foram removidos da análise, e, assim, se obteve uma pequena melhora na acurácia da predição genômica quando o valor de MSE foi usado para avaliar.

Uma vez que para a característica NV a acurácia da seleção genômica com um conjunto de SNPs foi semelhante à acurácia quando realizada a seleção com todos os marcadores, sugere-se que os métodos de aprendizagem de máquina podem ser utilizados para selecionar, eficientemente, conjuntos de marcadores para formarem painéis de genotipagem de baixa densidade e assim, reduzir os custos da implementação da seleção genômica, como já relatado na literatura (Liang e Kelemen, 2008; Long et al., 2011; Walters et al., 2012; Bermingham et al., 2015). Conforme sugerido por Li et al. (2018), uma das formas de aplicar os métodos de aprendizado de máquina em programas de melhoramento seria formar uma grande população de referência, inicialmente genotipando parte da população com painéis de alta densidade e posteriormente selecionar conjuntos de SNPs para genotipar o resto da população para realização da seleção genômica, reduzindo, desta forma, os custos com genotipagem e viabilizando a implantação ampla da seleção genômica nos programas de melhoramento animal.

Além disso, sugere-se que novos estudos sejam realizados para avaliar a habilidade de predição de conjuntos de SNPs quando mais características são utilizadas na seleção genômica, uma vez que no presente foram consideradas apenas duas características. Na prática, os animais são selecionados considerando as informações de uma variedade de características. Para seleção de SNPs mais importantes para as características utilizadas nos programas de melhoramento de suínos indica-se realizar análise de componentes principais com as mesmas para utilização do PCA1 como fenótipo de entrada das análises de RF, visto que essa metodologia aplicada no presente estudo demonstrou ser eficiente na determinação dos SNPs com importância para as características EG e NV em conjunto.

5 CONCLUSÕES

O método de aprendizagem de máquina “Random Forest” demonstrou ser eficiente na indicação da importância dos marcadores para características espessura de gordura e número de leitões nascidos vivos por leitegada com o objetivo de montar um conjunto de SNPs a serem utilizados na seleção genômica. O conjunto de marcadores selecionados a partir do “Random Forest” apresentou habilidade de predição semelhante quando comparado a habilidade de predição da seleção genômica com a informação de todos os SNPs para a característica número de leitões nascidos vivos. Portanto, essa metodologia pode ser utilizada para selecionar marcadores com objetivo de predição do valor genético genômico.

6 REFERÊNCIAS

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal Dairy Science** 93: 743–752.

Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al (2015) Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports* 5: 1–12.

Breiman L (2001). Random forests. **Machine Learning** 45(1): 5–32.

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

Crossa J, de los Campos G, Pérez-Rodríguez P, Gianola D, Burgueño J, Araus. JL, et al (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics** 186:713–724.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science** 92(3):1313.

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science** 92: 4656–4663.

Liang Y e Kelemen A (2008) Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. **Statistics Surveys** 2: 43–60.

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

Li B, Zhang N, Wang Y-G, George AW, Reverter A and Li Y (2018) Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics** 9: 237.

Long N, Gianola D, Rosa GJM, Weigel KA (2011). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **Journal of Animal Breeding and Genetics** 128: 247–257.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile MM, Kemper KE, Chamberlain AJ, et al. (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. **BMC Genomics** 17:144.

Misztal, I. BLUPF90 - a flexible mixed model program in Fortran 90. Manual. University of Georgia. 2012

Mokry FB, Higa RH, Mudadu MA, de Lima AO, et al. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics** 14:47.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT (2012) Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? **Briefings Bioinformatics** 14: 315-326

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

Xiang J (2015) Recent major advances of biotechnology and sustainable aquaculture in China. *Curr. Journal of Biotechnology* 4: 296–310.

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

Walters R, Laurin C, and Lubke GH (2012) An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. **Bioinformatics** 28: 2615–2623.

Zhu D, Liu X, Max R, Zhang Z, Zhao S, Fan B (2014) Genome-wide association study of the backfat thickness trait in two pig populations. **Frontiers of Agricultural Science and Engineering** 1(2): 91-95.