revista engenharia agrícola
Journal of the Brazilian Association of Agricultural Engineering
ISSN: 1809-4430 (on-line)

METODOLOGIA BASEADA EM TÉCNICAS DE MINERAÇÃO DE DADOS PARA SUPORTE À CERTIFICAÇÃO DE RAÇAS DE OVINOS

Doi:http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v35n6p1172-1186/2015

FÁBIO D. VIEIRA[1], STANLEY R. DE M. OLIVEIRA[2], SAMUEL R. PAIVA[3]

RESUMO: O objetivo deste trabalho foi desenvolver uma metodologia baseada em técnicas de mineração de dados para selecionar os principais marcadores SNP (*Single Nucleotide Polymorphism*) para as raças de ovinos: Crioula, Morada Nova e Santa Inês. Os dados utilizados foram obtidos do Consórcio Internacional de Ovinos e são compostos por 72 animais das raças citadas, e cada animal possui 49.034 marcadores SNP. Considerando que o número de atributos (marcadores) é muito maior que o de observações (animais), foram aplicadas as técnicas de predição LASSO (*Least Absolute Shrinkage and Selection Operator*), Random Forest e Boosting para a geração de modelos preditivos que incorporam métodos de seleção de atributos. Os resultados revelaram que os modelos preditivos selecionaram os principais marcadores SNP para identificação das raças estudadas. O modelo LASSO selecionou um total de 29 marcadores relevantes. A partir dos modelos Random Forest e Boosting, foram obtidos 27 e 20 marcadores importantes, respectivamente. Por meio da intersecção dos modelos gerados, identificou-se um subconjunto de 18 marcadores com maior potencial de identificação das raças.

PALAVRAS-CHAVE: polimorfismo de nucleotídeo único, seleção de atributos, modelos preditivos, regressão penalizada.

DATA MINING-BASED TECHNIQUE ON SHEEP BREED CERTIFICATION

ABSTRACT: This study aimed at developing a method based on data mining techniques to select key SNP markers (*Single Nucleotide Polymorphism*) for the sheep breeds Crioula, Morada Nova and Santa Inês. We gathered data from the International Sheep Consortium of 72 animals belonging to the aforementioned breeds; each animal has 49,034 SNP markers. Whereas the number of attributes (markers) is much greater than observations (animals), the LASSO (Least Absolute Shrinkage and Selection Operator), Random Forest and Boosting prediction methods were used to generate predictive models, incorporating selection methods and attributes. The results revealed that the predictive models selected the main SNP markers for sheep breed identification. The LASSO technique selected 29 relevant markers. Yet from Random Forest and Boosting selected 27 and 20 major markers, respectively. By intersecting the generated models, we could identify a subset of 18 markers with major potential for sheep breed identification.

KEYWORDS: single-nucleotide polymorphism, feature selection, predictive modeling, penalized regression.

INTRODUÇÃO

O Brasil possui diversas raças de ovinos que se desenvolveram a partir de raças trazidas pelos colonizadores e que adquiriram características específicas de adaptação às condições ambientais brasileiras. Essas raças passaram a ser conhecidas como locais ou localmente adaptadas. A maioria delas encontra-se ameaçada de extinção, principalmente devido a cruzamentos indiscriminados com

[1] Tecnólogo em Processamento de Dados, Mestre em Engenharia Agrícola, Analista da Embrapa Informática Agropecuária, Av. André Tosello, 209, Barão Geraldo, Campinas – SP, Fone: (19) 3211-5798, fabio.vieira@embrapa.br
[2] Cientista da Computação, Dr. em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Av. André Tosello, 209, Barão Geraldo, Campinas – SP, Professor colaborador da Feagri/Unicamp, stanley.oliveira@embrapa.br
[3] Biólogo, Dr. em Genética e Melhoramento, Pesquisador da Embrapa Labex EUA, Secretaria de Relações Internacionais, Parque Estação Biológica - PqEB, Brasília – DF, samuel.paiva@embrapa.br
Recebido pelo Conselho Editorial em: 02-10-2014
Aprovado pelo Conselho Editorial em: 04-9-2015

animais de raças exóticas (GOUVEIA, 2013). As raças locais constituem uma importante fonte de informações que pode levar à descoberta de genes envolvidos com características adaptativas, tais como resistência a doenças e parasitas (MARIANTE et al., 2009).

Para evitar a perda deste importante material genético, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) decidiu incluir as raças localmente adaptadas em seus Bancos de Germoplasma. Entre essas raças, as que possuem maior destaque nacional são as raças Crioula, Morada Nova e Santa Inês.

A seleção dos ovinos de uma determinada raça para compor esses bancos é realizada por meio de avaliação de características morfológicas e produtivas. Entretanto, essa avaliação está sujeita a falhas, pois alguns animais cruzados mantêm características semelhantes àquelas dos animais locais. Desta forma, identificar se os animais cadastrados nos bancos são ou não pertencentes a uma raça é uma tarefa que exige muita cautela (PAIVA, 2005).

Para auxiliar na busca de soluções para este tipo de problema, o emprego de tecnologias que fazem uso de marcadores moleculares SNP (*Single Nucleotide Polymorphism*, ou, em português, Nucleotídeo de Polimorfismo Único) se destacam-se entre as mais importantes. Os marcadores SNP constituem uma variação que ocorre em apenas um único nucleotídeo da cadeia de bases nitrogenadas (Adenina, Citosina, Timina e Guanina) do DNA, afetando ou não o fenótipo-alvo entre os membros de uma espécie em estudo. Contudo, as novas tecnologias para geração destes dados moleculares são capazes de genotipar milhares de SNPs para cada animal (PAIVA, 2005).

Desta forma, selecionar os marcadores mais relevantes para a identificação racial torna-se um problema desafiador. A aplicação de técnicas de mineração, etapa principal do processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD), constitui uma alternativa promissora para determinação dos marcadores mais relevantes, uma vez que essas técnicas são amplamente utilizadas na descoberta de padrões novos em grande volume de dados (HAN et al., 2011). Em particular, técnicas que combinam métodos de seleção de atributos e modelos preditivos são capazes de lidar com problemas em que o número de atributos p é muito maior que o número de observações n, isto é, $p>>n$. Entre essas técnicas estão: LASSO (*Least Absolute Shrinkage and Selection Operator*), Random Forest e Boosting.

Diversos estudos já foram conduzidos na geração de metodologias computacionais e estatísticas para identificação de subconjuntos de atributos que possam estar relacionados com características fenotípicas interessantes em variados organismos. Dentre eles, MOKRY et al. (2013) utilizaram Random Forest para a identificação de marcadores SNP ligados à espessura de gordura de gados Canchim. LEWIS et al. (2011) aplicaram Análise de Componentes Principais (PCA) para a identificação de SNP relevantes na rastreabilidade de bovinos. SUEKAWA et al. (2010) e SASAZAKI et al. (2011) selecionaram marcadores SNP para identificação racial de gados japoneses e americanos por meio de análise de frequência alélica. Outros trabalhos abordaram seleção de atributos em suínos (CORDEIRO et al., 2012) e humanos (WU et al., 2012), porém são raros os trabalhos envolvendo dados de ovinos. De forma geral, observou-se que os trabalhos relacionados selecionaram um número menor que 100 marcadores SNP em seus resultados finais (MOKRY et al., 2013; SASAZAKI et al., 2011; SUEKAWA et al., 2010), número a que este trabalho buscou referenciar-se como limite para a seleção dos SNPs mais informativos. Além disso, considerou-se o possível desenvolvimento de um microarranjo de baixa densidade, que aloca múltiplos de 48 marcadores SNP em sua superfície (ROORKIWAL et al., 2013).

O objetivo deste trabalho foi desenvolver uma metodologia baseada em algoritmos de mineração de dados para selecionar os marcadores SNP mais relevantes para as raças Crioula, Morada Nova e Santa Inês. A metodologia desenvolvida será utilizada na certificação racial de animais já cadastrados nos bancos de germoplasma e de dados de novos animais a serem inclusos nestes bancos, assim como poderão ser utilizados por associações de criadores interessadas no controle de animais registrados em seus próprios bancos de dados.

MATERIAL E MÉTODOS

As atividades de pesquisa foram executadas nos laboratórios de Inteligência Computacional e Bioinformática Aplicada da Embrapa Informática Agropecuária. A metodologia utilizada é composta de quatro etapas principais, a saber: entendimento dos dados, preparação dos dados, aplicação dos algoritmos e validação dos resultados.

Na primeira etapa (entendimento dos dados), o conjunto de dados analisado foi obtido do Consórcio Internacional do Genoma Ovino (ISGC et al., 2010) por meio da Rede Genômica Animal, projeto da Embrapa. Este conjunto era composto por dados de 72 animais das raças estudadas (23 animais da raça Crioula, 22 da Morada Nova e 27 da Santa Inês), sendo que, para cada animal, estavam armazenados valores de genótipos de 49.034 marcadores SNP. Observou-se, então, que o conjunto de dados é uma matriz em que o número de marcadores (p) é muito maior que o número de instâncias (n), isto é, $p>>n$. Cada um desses marcadores SNP possui um valor de genótipo, que é composto por dois alelos, sendo que cada alelo pode conter uma Adenina (A) ou uma Timina (T) ou uma Citosina (C) ou uma Guanina (G). A Figura 1 ilustra o formato do conjunto de dados de ovinos em estudo:

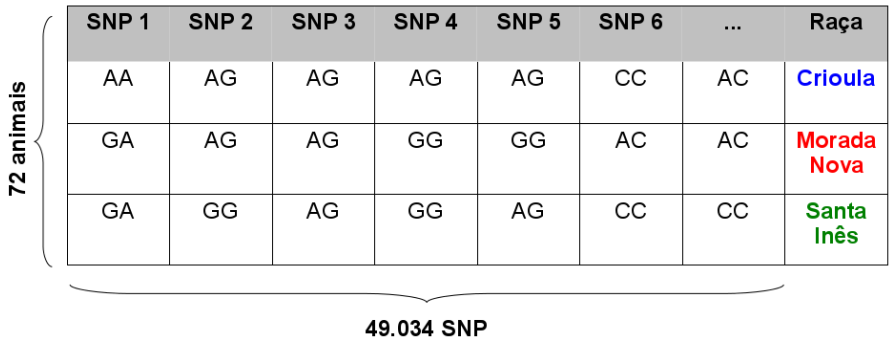

FIGURA 1. Formato do conjunto de dados de marcadores SNP das três raças em estudo. **Data set format composed by SNP markers of the three studied sheep breeds.**

Na etapa seguinte (preparação dos dados), foi realizada uma verificação quanto à existência de amostras idênticas dentro do conjunto de dados e de marcadores SNP que tivessem um valor único de genótipo para todas as raças. Após a verificação, constatou-se que não existiam amostras idênticas. Entretanto, existiam 384 marcadores SNP com valor único para todas as raças, os quais foram removidos do conjunto de dados final.

Na etapa de aplicação de algoritmos, foram utilizadas técnicas que combinam seleção de atributos e desenvolvimento de modelos preditivos para identificar os marcadores SNP mais relevantes para três raças de ovinos. Logo, devido ao elevado número de atributos (SNP) e ao baixo número de registros (animais), técnicas capazes de lidar com esta situação foram empregadas, a saber: LASSO (*Least Absolute Shrinkage and Selection Operator*), Random Forest e Boosting.

Em problemas de regressão, LASSO é um método utilizado para reduzir os efeitos dos atributos que não contribuem para a identificação do atributo-meta (ou variável resposta), reduzindo seus coeficientes para zero e excluindo-os do modelo (TIBSHIRANI, 1997). O método é usado normalmente para estimar os parâmetros de $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ regressão no modelo da [eq. (1)]:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i = \mu + X_i\beta + e_i \quad (1)$$

em que,

y_i é uma variável resposta numérica. Para a aplicação em um procedimento de classificação com duas categorias, a variável resposta pode ser codificada em ± 1 e para o caso de mais de uma categoria pode ser utilizado o procedimento *One versus All* (OVA), tal que y_i representa a raça do i-ésimo animal $(i = 1,2,\ldots,n)$;

μ é o coeficiente denominado intercepto, cujo valor é comum a todos os registros;

x_{ij} é o valor do genótipo do marcador $j(j = 1,2,\ldots,p)$do animal i;

o coeficiente $\beta_j$representa o efeito do marcador j,

e_i é o erro residual (HASTIE et al., 2011).

Sendo $\hat{\beta} = (\widehat{\beta_1},\widehat{\beta_2},\ldots,\widehat{\beta_p})^T$, a estimativa LASSO $(\hat{\mu},\hat{\beta})$ para problemas de classificação é definida pela função de máxima verossimilhança penalizada descrita na [eq. (2)]:

$$l(\hat{\mu},\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \quad [y_i(\mu + \sum_{j=1}^{p} \quad x_{ij}\beta_j) - \log(1 + e^{\mu+\sum_{j=1}^{p} \quad x_{ij}\beta_j})] \qquad (2)$$

sujeito à restrição $\sum_{j=1}^{p} \quad |\beta_j| \leq t$ para $t \geq 0$,

em que,

t é um parâmetro de penalização e que deve ser determinado separadamente. Normalmente, os algoritmos de implementação do LASSO fornecem o valor ótimo para tal parâmetro, utilizando uma análise por validação cruzada de um intervalo de *n* possíveis valores.

Random Forest (em português, Floresta Aleatória) é uma técnica de classificação e regressão desenvolvida por BREIMAN (2001), que consiste num conjunto de árvores de decisão combinadas para solucionar problemas de classificação. Cada árvore de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de *m* atributos é utilizado para a escolha dos atributos mais informativos. No final, Random Forest gera uma lista dos atributos mais importantes no desenvolvimento da floresta, que são determinados pela importância acumulada do atributo nas divisões dos nós de cada árvore da floresta (JAMES et al., 2013). Os principais passos do algoritmo Random Forest podem ser vistos na Figura 2.

Dado um conjunto de dados $X = x_1, x_2, \ldots, x_j$ e $Y = y_1, y_2, \ldots y_k$.

Para $b = 1,2,3,\ldots B$, repita:

(a) Cria uma amostra *bootstrap* (X_b, Y_b) com *n* exemplos de (X, Y).

(b) Ajusta uma árvore de decisão f^b para o conjunto de treinamento (X_b, Y_b), utilizando *m* atributos para a escolha de cada nó.

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

FIGURA 2. Algoritmo básico da técnica Random Forest (BREIMAN, 2001). **Random Forest algorithm (BREIMAN, 2001).**

De forma geral, uma árvore de decisão é um modelo gráfico representado por nós e ramos, em que os nós intermediários, ou decisórios, representam os testes de atributos (variáveis independentes), enquanto os ramos representam os resultados desses testes. O nó localizado no topo

da árvore representa seu início e é denominado nó-raiz. Já o nó externo, que não possui um nó descendente, localizado na extremidade inferior, é denominado folha ou terminal, e representa o valor de predição do atributo-meta ou classe (HAN et al., 2011). Para evitar *overfitting* (em português, superajuste), foi utilizada a abordagem Random Forest que, em geral, lida melhor com o problema de superajuste nos modelos (MEGETO et al., 2014).

A ideia principal da técnica Boosting (em português, sua tradução seria algo como "melhorar a performance") é transformar múltiplos classificadores ruins em um único muito bom (FREUND & SCHAPIRE, 1999). Essa definição pode ser interpretada da seguinte maneira: um classificador será fraco se a probabilidade deste classificador ser construído, com base numa amostra *D*, tiver erro menor do que 50%. Ou seja, considerar um classificador fraco será ligeiramente melhor do que escolher aleatoriamente uma das classes com probabilidade de 50%. Os métodos desta abordagem funcionam aplicando-se, sequencialmente, um algoritmo de classificação a versões reponderadas do conjunto de dados de treinamento, dando maior peso aos registros classificados erroneamente no passo anterior. O algoritmo que mostra a execução básica de Boosting é descrito na Figura 3.

Para a aplicação das técnicas de modelagem, escolheu-se o *software* R (versão 3.0.1). O pacote instalado para o algoritmo LASSO foi o glmnet (FRIEDMAN et al., 2010), para Random Forest foi instalado o pacote randomForest (LIAW & WIENER, 2002) e, para Boosting, foi instalado o algoritmo gbm (RIDGEWAY, 2013). Além destes, instalou-se o pacote caret (KUHN, 2013), utilizado para a escolha dos melhores valores para alguns parâmetros de cada técnica aplicada.

Foram realizados vários experimentos, utilizando cada uma das técnicas, procurando obter modelos que fornecessem os melhores resultados em termos de acurácia e menor número de marcadores selecionados. Para tanto, antes da utilização do pacote caret, os principais parâmetros de cada uma das técnicas foram ajustados diversas vezes para atingir tal objetivo.

Dado um conjunto de dados de treinamento $X = x_1, x_2, \ldots, x_j$ e $Y = y_1, y_2, \ldots y_k$.

Define $\hat{f}(x) = 0$ e $residuos_i = y_i$ para todos os registros do treinamento.

Para $b = 1, 2, 3, \ldots B$, repita:

(a) Ajusta um modelo f^b para o conjunto de treinamento $(X, residuos)$.

(b) Atualiza $\hat{f}$ com o novo modelo:

$\hat{f}(x) = \hat{f}(x) + f^b(x)$.

(c) Atualiza os resíduos (erros na classificação):

$residuos_i = residuos_i - f^b(x_i)$.

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

FIGURA 3. Algoritmo básico do algoritmo Boosting (JAMES et al., 2013). **Boosting algorithm (JAMES et al., 2013).**

LASSO foi a primeira técnica a ser aplicada, e o único parâmetro testado foi o intervalo de possíveis valores para o coeficiente de penalização *t*. O número-padrão deste intervalo é de 100 valores possíveis (JAMES et al., 2013; FRIEDMAN et al., 2010), obtidos separadamente pelo algoritmo LASSO, via validação cruzada, sobre os dados analisados. Após a aplicação da técnica LASSO, utilizou-se de Random Forest para a busca dos marcadores SNP mais relevantes, associados a cada uma das raças. Os parâmetros avaliados para Random Forest foram o número de árvores a serem construídas e o número de atributos selecionados para determinar o *split* (divisão) em cada nó das árvores. Com a construção desta floresta, foi possível determinar os marcadores mais importantes para o modelo (do atributo mais importante ao menos relevante). Assim como Random Forest, Boosting foi utilizado para fornecer um modelo com a listagem dos marcadores mais importantes na identificação das raças. O único parâmetro testado para Boosting foi o número de classificadores a serem desenvolvidos para o modelo final. Os classificadores construídos pela

técnica Boosting foram baseados em árvores de decisão, as quais foram construídas em distribuições reponderadas dos dados.

Após a obtenção dos modelos e dos conjuntos de marcadores mais importantes para identificação das raças, foi realizada uma análise da frequência alélica de cada um desses marcadores a fim de verificar o quanto um alelo estava presente em uma raça e ausente em outras duas. Por fim, foi selecionado um subconjunto menor de marcadores SNP com maior potencial de identificação das três raças pesquisadas.

Na última etapa, para avaliar o desempenho dos modelos, dividiu-se o conjunto de dados inicial em duas partes disjuntas, sendo que uma parte constitui o conjunto de treinamento e a outra o conjunto de teste. As técnicas utilizaram dois tipos de particionamento dos dados: validação cruzada e *bootstrap* (reamostragem). Na validação cruzada, os dados são particionados em *k* subconjuntos de tamanhos aproximadamente iguais, e o indutor é treinado e testado *k* vezes. Para cada uma das vezes, o indutor é testado com uma das partições e treinado com o restante. O *bootstrap* consiste em gerar conjuntos de treinamento e teste a partir de uma amostragem randômica dos dados, repetindo esse processo de classificação por várias vezes. A cada ciclo, as amostragens são selecionadas com reposição, isto é, um mesmo exemplo poderá aparecer mais de uma vez no mesmo subconjunto.

Os modelos foram analisados por meio dos valores da acurácia e do coeficiente Kappa. A acurácia, ou taxa de acerto, fornece a porcentagem de observações que foram classificadas corretamente pelo classificador, enquanto o Kappa (COHEN, 1960) mede o grau de concordância entre as classes preditas e observadas, deduzindo o número esperado de acertos (utilizando uma classificação ao acaso) do número real de acertos do classificador (WITTEN et al., 2011).

RESULTADOS E DISCUSSÃO

Na aplicação do algoritmo LASSO, para a obtenção do melhor valor de *t*, avaliaram-se intervalos de 100 e de 1.000 valores possíveis. Entretanto, o número de marcadores selecionados e a acurácia permaneceram inalterados, mantendo-se, então, os 100 valores fornecidos por caret. Com o valor ótimo de *t* (0,0035243), o algoritmo LASSO selecionou 29 marcadores relevantes, dos quais cinco se destacaram para a raça Crioula, 12 para Morada Nova e 12 para Santa Inês. Os cinco marcadores com destaque para Crioula e suas respectivas informações estão descritos na Tabela 1.

TABELA 1. Frequências alélicas dos marcadores SNP selecionados pelo algoritmo LASSO para a raça Crioula. **Allele frequencies of SNP markers selected by LASSO algorithm from Crioula breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
OARX_29830880.1	X	29830880	[A/G]	0.80	0	0.05
OARX_78903642.1	X	78903642	[A/G]	0.95	0.07	0.09
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR1_268303279_X.1	1	268303280	[G/A]	0.78	0.07	0.09

* Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.

De forma geral, todos os marcadores mostraram alto potencial de identificação da raça Crioula, destacando-se, entre outros, quatro marcadores (OARX_121724022.1, s56924.1, OARX_78903642.1 e OARX_29830880.1) pertencentes ao cromossomo X. Foi observado que todos os marcadores da raça Crioula possuem altas diferenças de frequências em relação às outras raças, o que se deve, provavelmente, ao fato de ela possuir as características físicas mais distintas das demais, como possuir tamanho diminuto e ser lanada (PAIVA, 2005).

TABELA 2. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Morada Nova. **Allele frequencies of SNP markers selected by LASSO algorithm from Morada Nova breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
s05480.1	X	52592630	[G/A]	0.93	0.15	0.22
OAR1_187375309_X.1	1	187375310	[A/G]	0.86	0.02	0.31
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
DU373896_534.1	3	139464759	[A/C]	0.82	0.35	0.15
s32131.1	4	22382506	[A/G]	0.98	0.32	0.42
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31
OAR6_39029427.1	6	39029427	[A/G]	0.84	0.17	0.11
OAR9_39924477.1	9	39924477	[A/C]	0.95	0.17	0.33
OAR10_33338187.1	10	33338187	[A/G]	0.90	0.22	0.28
OAR17_22334380.1	17	22334380	[G/A]	0.79	0.19	0.13
OAR17_8472049.1	17	8472049	[A/G]	0.95	0.22	0.37
OAR20_45964534.1	20	45964534	[G/A]	0.75	0	0.15

* Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.

Para a raça Morada Nova, LASSO identificou os 12 marcadores listados na Tabela 2.

Os destaques para a raça Morada Nova são dois SNPs (OAR1_187375309_X.1 e OAR1_194627962.1) no cromossomo um, e dois SNPs (OAR17_8472049.1 e OAR17_22334380.1) no cromossomo 17, além do total de seis marcadores com frequência acima de 90%. Foi observado ainda que há uma frequência relativamente maior dos alelos dos animais Morada Nova na raça Santa Inês. Isto talvez seja explicado pelo fato de os animais Santa Inês serem originários do cruzamento entre Morada Nova e outros ovinos sem raça definida do Nordeste brasileiro, fazendo com que muitos ovinos Santa Inês preservem características do Morada Nova (PAIVA, 2005).

TABELA 3. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Santa Inês. **Allele frequencies of SNP markers selected by LASSO algorithm from Santa Inês breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0	0.09
OAR2_145195113.1	2	145195113	[A/G]	0.74	0.04	0.38
OAR2_242658985.1	2	242658985	[A/G]	0.85	0.17	0.29
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_153703374.1	3	153703374	[A/G]	0.76	0.41	0.13
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
s16949.1	3	164901721	[G/A]	0.89	0.15	0.18
OAR5_93120389.1	5	93120389	[G/A]	0.89	0.19	0.38
OAR7_21409209.1	7	21409209	[G/A]	0.61	0.02	0.11
OAR7_94733688.1	7	94 733688	[G/A]	0.98	0.37	0.59
s11241.1	7	30741909	[C/A]	0.81	0.35	0.34
s59000.1	18	45393237	[A/G]	0.87	0.30	0.38

* Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.

Para a Santa Inês, foram selecionados os 12 marcadores (Tabela 3), sendo que três pertencem ao cromossomo dois (OAR2_145195113.1, OAR2_242658985.1 e s20468.1), três ao cromossomo três (OAR3_153703374.1, OAR3_165050963.1 e s16949.1) e três ao cromossomo sete (OAR7_21409209.1, OAR7_94733688.1 e s11241.1). Uma observação importante vem do fato de que os três marcadores do cromossomo três estão em posições muito próximas. De maneira geral, os marcadores para a raça Santa Inês têm altas diferenças de frequência alélica em relação às outras raças, tendo como destaque os marcadores OARX_53305527.1 e s20468.1.

A acurácia atingida com o conjunto de 29 marcadores SNP selecionados pelo algoritmo LASSO foi de 100% na predição de novas raças, e o índice Kappa foi igual a 1. O algoritmo LASSO teve ótimo desempenho, tanto em termos de acurácia quanto na questão computacional, conforme já demonstrado em AYERS & CORDELL (2010), cujos resultados também confirmaram uma boa performance de outras técnicas de regressão penalizada.

Random Forest gerou igualmente uma listagem dos marcadores mais importantes para a identificação das raças ovinas. Experimentaram-se classificadores combinando de 1.000 a 5.000 árvores, e conjuntos aleatórios de atributos variando de 20 a 49.033 atributos para *split* dos nós. Após esses experimentos, o melhor resultado obtido foi utilizando os parâmetros fornecidos pelo pacote caret, que resultou em 1.000 árvores e 313 marcadores para *split*. Selecionaram-se, então, os 27 melhores SNP classificados, considerando: a) estes marcadores estão sendo testados em microarrays que devem ter baixa densidade (múltiplos de 48). Portanto, quanto menor o número de marcadores SNP, menor será o custo da construção do chip; b) os marcadores foram ordenados de acordo com a queda da entropia quando utilizados em um *split* de uma árvore. Assim, a partir de um certo número de marcadores (por exemplo 0,3%), a queda da entropia foi irrelevante. Em MOKRY et al. (2013), utilizou-se de um critério de seleção diferente, no qual, primeiramente, selecionou-se 1% dos SNPs mais relevantes de cada cromossomo e, em seguida, foi selecionado 1% dos SNPs mais importantes do subconjunto anterior, sendo selecionados 70 marcadores SNP pela técnica Random Forest, utilizando tal critério.

Do conjunto total de 27 marcadores, nove coincidiram com aqueles selecionados pelo algoritmo LASSO. Agrupando-se os marcadores fornecidos por Random Forest, de acordo com a raça, foram geradas três tabelas para a análise da frequência do alelo específico de cada uma delas em relação às outras. A Tabela 4 mostra os marcadores predominantes na raça Crioula e as frequências dos alelos específicos desta raça em relação à Morada Nova e Santa Inês.

Do conjunto de 13 marcadores identificados por Random Forest para a raça Crioula, quatro também foram identificados por LASSO (OARX_121724022.1, OARX_29830880.1 e OARX_78903642.1, s56924.1). Os dois SNPs do cromossomo 25 estão em posições próximas e com frequência acima de 90% dentro da raça, surgindo como bons separadores raciais. De forma geral, os SNPs fornecidos por Random Forest mostraram-se importantes na identificação da raça Crioula.

TABELA 4. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Crioula. **Allele frequencies of SNP markers selected by Random Forest algorithm from Crioula breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
OARX_29830880.1	X	29830880	[A/G]	0.80	0	0.05
OARX_78903642.1	X	78903642	[A/G]	0.95	0.07	0.09
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR1_23724877.1	1	23724877	[G/A]	0.50	0	0.04
OAR2_212548956.1	2	212548956	[G/A]	0.80	0.04	0.18
OAR2_55853730.1	2	55853730	[A/C]	0.85	0	0.07
OAR11_18815864.1	11	18815864	[A/G]	0.93	0.34	0.22
s71482.1	14	41937578	[G/A]	0.91	0.18	0.50
OAR15_45152619.1	15	45152619	[G/A]	0.76	0.02	0.02
OAR16_39888776.1	16	39888776	[A/G]	0.89	0.11	0.15
s25195.1	25	7203123	[G/A]	0.93	0.02	0.30
s30024.1	25	7165805	[C/A]	0.91	0.02	0.28

* Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.

Na Tabela 5, os SNPs com predominância na raça Morada Nova são listados.

TABELA 5. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Morada Nova. **Allele frequencies of SNP markers selected by Random Forest algorithm from Nova Morada breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
OAR2_54691204.1	2	54691204	[G/A]	0.57	0.04	0
OAR18_65638912.1	18	65638912	[G/A]	1.00	0.56	0.41

* Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.

O algoritmo Random Forest indicou três marcadores importantes para a raça Morada Nova. Como destaque, observam-se os marcadores OAR1_194627962.1, indicado também pelo algoritmo LASSO, e OAR2_54691204.1, com frequência acima de 50% na Morada Nova e praticamente ausente nas outras duas raças. O marcador OAR18_65638912.1 destaca-se com frequência de 100% na raça Morada Nova, apesar de sua frequência em outras duas raças ter ficado entre 40% e 60%.

Na Tabela 6, podem-se observar os SNPs com alta frequência na raça Santa Inês, em que 11 marcadores foram selecionados com altas frequências alélicas. Destes, quatro foram identificados pelo algoritmo LASSO (OARX_53305527.1, s20468.1, OAR3_165050963.1 e s16949.1). Um dado interessante é que cinco marcadores são originados do cromossomo três (OAR3_164788310.1, OAR3_165050963.1, OAR3_195698523.1, s16949.1 e s69653.1). O marcador s61697.1 também se destaca com alta frequência na raça Santa Inês e com frequências abaixo de 7% na raça Crioula e Morada Nova.

TABELA 6. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Random Forest para a raça Santa Inês. **Allele frequencies of SNP markers selected by Random Forest algorithm from Santa Inês breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0	0.09
s61697.1	-	-	[C/A]	0.68	0.06	0.04
OAR1_175474366.1	1	175474366	[G/A]	0.55	0.24	0
s03528.1	1	28583773	[A/G]	0.92	0.43	0.23
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_164788310.1	3	164788310	[G/A]	0.89	0.22	0.18
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
OAR3_195698523.1	3	195698523	[A/G]	0.66	0.15	0.04
s16949.1	3	164901721	[G/A]	0.89	0.15	0.18
s69653.1	3	164951744	[G/A]	0.90	0.08	0.36
OAR9_76802154.1	9	76802154	[A/G]	0.96	0.32	0.50

* Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.

Para treinamento e teste, foram desenvolvidas e combinadas 1.000 árvores utilizando as amostras *bootstrap*. O comitê de classificadores que formaram a floresta obteve uma acurácia de 99%, e Kappa de, 0,98.

Na aplicação da técnica Boosting, o único parâmetro testado foi o número de classificadores (neste caso, árvores de decisão) a serem construídos. Avaliaram-se classificadores desenvolvidos com totais entre 1.000 e 10.000 árvores, sendo que o melhor resultado, em termos de acurácia e Kappa, ocorreu com 1.000 árvores, número fornecido pelo pacote caret. Sselecionou-se os 20 melhores marcadores, pois, assim como em Random Forest, os SNPs a partir desta posição pouco contribuíam para a redução da entropia na construção dos *splits* das árvores. Entre os 20 marcadores ordenados por Boosting, seis estavam presentes nos resultados dos algoritmos LASSO e Random Forest, dois estavam somente em LASSO e sete somente em Random Forest. Com isto, Boosting selecionou apenas cinco marcadores diferentes das técnicas anteriores. Na Tabela 7, estão descritos os SNPs predominantes na raça Crioula.

TABELA 7. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Crioula. **Allele frequencies of SNP markers selected by Boosting algorithm from Crioula breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Crioula	Morada Nova	Santa Inês
OARX_121724022.1	X	121724022	[C/A]	0.98	0.02	0.05
s56924.1	X	53358543	[A/G]	0.98	0.13	0.15
OAR2_55853730.1	2	55853730	[A/C]	0.85	0	0.07
OAR4_51441757.1	4	51441757	[A/G]	0.91	0.25	0.16
OAR6_110447914.1	6	110447914	[G/A]	0.67	0.04	0.02
OAR15_45152619.1	15	45152619	[G/A]	0.76	0.02	0.02
s30024.1	25	7165805	[C/A]	0.91	0.02	0.28

* Alelo específico para a raça Crioula do lado esquerdo. ** Frequência do alelo específico na população Crioula e nas raças Morada Nova e Santa Inês.

Na lista de marcadores importantes para a raça Crioula, dois deles (OARX_121724022.1 e s56924.1) foram indicados pelos algoritmos anteriores, e outros dois (OAR2_55853730.1 e OAR15_45152619.1) foram selecionados por Random Forest, demonstrando o alto potencial destes marcadores. Os marcadores indicados apenas pelo algoritmo Boosting (OAR4_51441757.1, OAR6_110447914.1 e s30024.1) também mostraram ser potenciais discriminantes de raças.

TABELA 8. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Morada Nova. **Allele frequencies of SNP markers selected by Boosting algorithm from Nova Morada breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Morada Nova	Crioula	Santa Inês
OAR1_194627962.1	1	194627962	[G/A]	0.73	0	0.02
s32131.1	4	22382506	[A/G]	0.98	0.32	0.42
s06182.1	5	30787155	[A/G]	0.93	0.15	0.31
s10365.1	10	21720029	[G/A]	0.45	0	0

* Alelo específico para a raça Morada Nova do lado esquerdo. ** Frequência do alelo específico na população Morada Nova e nas raças Crioula e Santa Inês.

A Tabela 8 traz uma listagem dos marcadores com predominância na raça Morada Nova.

O algoritmo Boosting separou cinco marcadores com maior frequência em Morada Nova, sendo um deles (OAR1_194627962.1) presente nos dois algoritmos anteriores e dois (s32131.1, s06182.1) no algorimo LASSO. O marcador OAR1_194627962.1 possui frequência de apenas 2% na Santa Inês e ausente na Crioula, resultado que o confirma como um bom discriminante de raças. Os marcadores s32131.1 e s06182.1 surgem com frequência acima de 90% nos animais Morada Nova, o que também demonstra o bom potencial destes SNPs.

TABELA 9. Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo Boosting para a raça Santa Inês. **Allele frequencies of SNP markers selected by Boosting algorithm from Santa Inês breed.**

SNP	Cromossomo	Posição	Alelos*	Frequência alélica**		
				Santa Inês	Crioula	Morada Nova
OARX_53305527.1	X	53305527	[A/G]	0.72	0	0.09
s61697.1	-	-	[C/A]	0.68	0.06	0.04
s03528.1	1	28583773	[A/G]	0.92	0.43	0.23
s20468.1	2	56248983	[A/G]	0.76	0.15	0
OAR3_164788310.1	3	164788310	[G/A]	0.89	0.22	0.18
OAR3_165050963.1	3	165050963	[A/G]	0.80	0.02	0.07
s39114.1	3	232410568	[A/G]	0.59	0.08	0.07
s69653.1	3	164951744	[G/A]	0.90	0.08	0.36
OAR9_40217510.1	9	40217510	[C/A]	0.54	0.08	0.02

* Alelo específico para a raça Santa Inês do lado esquerdo. ** Frequência do alelo específico na população Santa Inês e nas raças Crioula e Morada Nova.

Na Tabela 9, dentre os marcadores fornecidos pelo algoritmo Boosting para a raça Santa Inês, destacam-se três deles (OARX_53305527.1, s20468.1, OAR3_165050963.1) também selecionados pelas técnicas LASSO e Random Forest. Além disso, dois SNPs (s39114.1, OAR9_40217510.1) foram selecionados exclusivamente por Boosting. De forma geral, a maioria dos marcadores selecionados para a raça Santa Inês apresenta alta frequência de alelo, o que atesta o potencial do algoritmo na identificação da raça Santa Inês. Prova de tal afirmação está na seleção dos três SNPs também indicados pelos dois métodos anteriores.

Para realização de treinamento e teste, o algoritmo Boosting foi executado por meio de validação cruzada em 10 subconjuntos de dados, sendo que o resulado final foi obtido por meio da média dos 10 subconjuntos. A acurácia e o Kappa obtidos pelo algoritmo, com a combinação dos classificadores ajustados, foram de 100% e 1, respectivamente. Observando-se esses resultados, pode-se acreditar que há indícios de superajuste, porém os parâmetros ajustados para a execução do algoritmo foram obtidos pelo caret de forma a evitar um superajuste do modelo.

Com a seleção dos principais marcadores para a identificação das raças, foi realizada uma análise daqueles SNPs que foram identificados na intersecção dos resultados de dois e de três técnicas. A intersecção dos resultados envolvendo a raça Crioula mostra que os marcadores OARX_121724022.1 e o s56924.1 foram selecionados pelos três algoritmos, demonstrando alta relevância na identificação da raça Crioula. O marcador OARX_ 121724022.1, em especial, possui a frequência de 98%, ou seja, demonstra ser um SNP com alto potencial de identificação da raça.

A intersecção dos resultados relativa à raça Morada Nova exibe o marcador OAR1_194627962.1 com frequência de 73% para a raça Morada Nova e frequências praticamente nulas nas outras raças, o que caracteriza esse SNP como bom discriminante da raça. Os algoritmos LASSO e Boosting selecionaram os SNPs s32131.1 e s06182.1, os quais possuem frequências acima de 90% na raça Morada Nova, colocando-os também como altamente relevantes para a raça.

Em relação à raça Santa Inês, a intersecção destaca a presença de três marcadores (OARX_53305527.1, s20468.1 e OAR3_165050963.1) que apresentam frequências acima de 70% em ovinos Santa Inês e abaixo de 10% em outras raças, confirmando alta capacidade na discriminação racial. Entre os marcadores obtidos por Random Forest e Boosting, destaca-se o s61697.1, com frequência de 68%, posicionando-o como um potencial identificador da raça.

TABELA 10. Marcadores SNP selecionados pelos modelos e suas raças predominantes. **Selected SNP markers and their respective predominant breeds**.

SNP	Nº Modelos	Cromossomo	Posição	Alelos*	Raça Predominante
OARX_121724022.1	3	X	121724022	[C/A]	Crioula
s56924.1	3	X	53358543	[A/G]	Crioula
OAR1_194627962.1	3	1	194627962	[G/A]	Morada Nova
OARX_53305527.1	3	X	53305527	[A/G]	Santa Inês
s20468.1	3	2	56248983	[A/G]	Santa Inês
OAR3_165050963.1	3	3	165050963	[A/G]	Santa Inês
OARX_29830880.1	2	X	29830880	[A/G]	Crioula
OARX_78903642.1	2	X	78903642	[A/G]	Crioula
OAR2_55853730.1	2	2	55853730	[A/C]	Crioula
OAR15_45152619.1	2	15	45152619	[G/A]	Crioula
s30024.1	2	25	7165805	[C/A]	Crioula
s32131.1	2	4	22382506	[A/G]	Morada Nova
s06182.1	2	5	30787155	[A/G]	Morada Nova
s61697.1	2	-	-	[C/A]	Santa Inês
s03528.1	2	1	28583773	[A/G]	Santa Inês
OAR3_164788310.1	2	3	164788310	[G/A]	Santa Inês
s69653.1	2	3	164951744	[G/A]	Santa Inês
s16949.1	2	3	164901721	[G/A]	Santa Inês

* Alelo específico para a raça predominante do lado esquerdo.

A Tabela 10 apresenta os 18 marcadores selecionados pela intersecção dos resultados de dois e de três algoritmos. A seleção dos 18 SNPs foi influenciada pela confirmação de mais de um algoritmo, tornando esses marcadores com maior potencial. Esse número de marcadores é próximo aos resultados de trabalhos relacionados à identificação racial em bovinos, como em SUEKAWA et al. (2010), onde foram encontrados cinco marcadores por meio de análise de frequência alélica capaz de distinguir gados japoneses e americanos. Por sua vez, SASAZAKI et al. (2011) desenvolveram um modelo no qual foram selecionados 11 SNPs importantes para gados provenientes de rebanhos dos Estados Unidos.

CONCLUSÕES

A avaliação dos modelos com aplicação das três técnicas escolhidas revelou resultados promissores para a seleção dos marcadores SNP mais informativos, que identificam as raças estudadas. Em particular, os modelos gerados pelas técnicas LASSO e Boosting obtiveram resultados melhores, em termos de acurácia e Kappa, em comparação com o modelo Random Forest. Considerando que o conjunto de dados utilizado possui elevado número de atributos, as técnicas utilizadas reduziram o número de SNP para menos de 0,2%. Na intersecção dos marcadores que compõem os modelos, foram encontrados 18 SNPs com maior potencial de identificação das raças, indicando que, realmente, os marcadores selecionados possuem alta correlação com a raça associada. Os modelos desenvolvidos podem ser utilizados na certificação racial de animais já depositados em bancos de germoplasma e de novos animais a serem inclusos nestes bancos, assim como poderão ser utilizados por diversos segmentos ligados à ovinocultura, como por exemplo, associações de criadores interessadas em certificar seus animais, e pelo MAPA

(Ministério da Agricultura, Pecuária e Abastecimento), no controle de animais registrados que apresentam alelos de outras raças, possibilitando a reclassificação desses animais. Adicionalmente, a metodologia proposta poderá ser estendida para toda e qualquer espécie animal de produção.

REFERÊNCIAS

AYERS, K. L.; CORDELL, H. J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. **Genetic epidemiology**, New York, v.34, n.8, p.879-91, 2010.

BREIMAN, L. Random forests. **Machine Learning**, Boston, v.45, n.1, p.5-32, 2001.

COHEN, J.A. A coefficient of agreement of nominal scales. **Educational and Psychological Measurement**, Durhan, v.20, p.37-46, 1960.

CORDEIRO, A. F. S.; NÄÄS, I. A.; OLIVEIRA, S. R. M.; VIOLARO, F.; ALMEIDA, A. C. M. Efficiency of distinct data mining algorithms for classifying stress level in piglets from their vocalization. **Engenharia Agrícola**, Jaboticabal , v.32, n.2, p.208-216, mar./abr. 2012.

FREUND, Y.; SCHAPIRE, R. A short introduction to boosting. **Journal of Japanese Society for Artificial Intelligence**, Amsterdam, v.14, n.5, p.771-780, 1999.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, Los Angeles, v.33, n.1, p.1-22, 2010.

GOUVEIA, J. J. S. **A utilização da genômica de populações na análise das principais raças de ovinos brasileiras**. 2013, 98f. Tese (Doutorado) – Universidade Federal do Ceará, Fortaleza, 2013.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 3rd ed., 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. London: Ed. Springer, 2011. 745 p.

ISGC - THE INTERNATIONAL SHEEP GENOMICS CONSORTIUM; ARCHIBALD, A.L.; COCKETT, N.E.; DALRYMPLE, B.P.; FARAUT, T.; KIJAS, J.W.; MADDOX, J.F.; MCEWAN, J.C.; HUTTON ODDY, V.; RAADSMA, H.W.; WADE, C.; WANG, J.; WANG, W.; XUN, X. The sheep genome reference sequence: a work in progress. **Animal Genetics**, Oxford, n.41, p.449–453, 2010.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**: with applications in R. London: Ed. Springer, 2013. 429 p.

KUHN, M. **Caret**: classification and regression training. R package version 5.16-24. 2013.

LIAW, A.; WIENER, M. Classification and regression by Random Forest. **R News**, v.2, n.3, p.18-22, 2002.

LEWIS, J.; ABAS, Z.; DADOUSIS, C.; LYKIDIS, D.; PASCHOU, P.; DRINEAS, P. Tracing cattle breeds with principal components analysis ancestry informative SNPs. **PloS one**, San Francisco, v.6, n.4, p. E18007, 2011.

MARIANTE, A. S.; ALBUQUERQUE, M. S. M.; EGITO, A. A.; MCMANUS, C.; LOPES, M. A.;

MEGETO, G. A. S.; OLIVEIRA, S. R. M.; PONTE, E. D.; MEIRA, C. A. A. Árvore de decisão para classificação de ocorrências de ferrugem asiática em lavouras comerciais com base em variáveis meteorológicas. **Engenharia Agrícola**, Jaboticabal, v.34, n.3, p.590-599, maio/jun. 2014.

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. A.; LIMA, A. O.; MEIRELLES, S. L. C.; SILVA, M. V. G. B.; CARDOSO, F. F.; OLIVEIRA, M. M. O.; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M.; REGITANO, L. C. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**, London, v.14, n.47, 2013.

PAIVA, S. R. **Caracterização da diversidade genética de ovinos no Brasil com quatro técnicas moleculares**. 2005. 108f. Tese (Doutorado) - Universidade Federal de Viçosa, Viçosa, 2005.

RIDGEWAY, G. GBM: generalized goosted gegression models. **R package version 2.1**. 2013.

ROORKIWAL, M; SAWARGAONKAR, S. L.; CHITIKINENI, A.; THUDI, M.; SAXENA, R. K.; UPADHYAYA, H. D.; VALES, M. I.; RIERA-LIZARAZU, O.; VARSHNEY, R. K. Single nucleotide polymorphism genotyping for breeding and genetics applications in chickpea and pigeonpea using the BeadXpress platform.. **The Plant Genome**, Madison, v.6, n.2, 2013. 10p.

SASAZAKI, S.; HOSOKAWA, D.; ISHIHARA, R.; AIHARA, H.; OYAMA, K.; MANNEN, H. Development of discrimination markers between Japanese domestic and imported beef. **Animal Science Journal**, Oxford, v.82, n.1, p.67-72, 2011.

SUEKAWA, Y.; AIHARA, H.; ARAKI, M.; HOSOKAWA, D.; MANNEN, H.; SASAZAKI, S. Development of breed identification markers based on a bovine 50K SNP array. **Meat Science**, Kidlington, v.85, n.2, p.285-288, jun. 2010.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Statistics in Medicine**, Chichester, v.16, p.385-395, 1997.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publishers, 2011.

WU, Q.; YE, Y.; LIU, Y.; NG, M. K. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. **IEEE Transactions on Nanobioscience**, Piscataway, v.11, p.216–227, 2012.