USO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA SUPORTE À CERTIFICAÇÃO RACIAL DE OVINOS

Fábio Danilo Vieira[1], *Stanley Robson de Medeiros Oliveira*[1]
[1] Embrapa Informática Agropecuária
Caixa Postal 6041, CEP 13083-970 – Campinas – SP – Brasil
fabio.vieira@embrapa.br, stanley.oliveira@embrapa.br

RESUMO

Este artigo apresenta um conjunto de modelos baseados 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. O conjunto de dados utilizado é composto por 72 animais das raças citadas, sendo que cada animal possui 49.034 marcadores SNP. Dado que o número de marcadores é muito maior que o de observações (animais), foram aplicadas técnicas para a geração de modelos preditivos que incorporam métodos de seleção de atributos. Os resultados demostraram que os modelos preditivos foram capazes de selecionar os principais marcadores SNP para identificação das raças. 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, Classificação, Regressão penalizada.

ABSTRACT

This paper introduces a set of models based on data mining techniques to select the major markers SNP (Single Nucleotide Polymorphism) for the breeds of sheep Creole, Morada Nova and Santa Inês. The data used were obtained from the International Consortium of Sheep and consist of 72 animals of these breeds, in which each animal has 49,034 SNP markers. Considering that the number of markers is much larger than the observations (animals), techniques for generation of predictive models and feature selection methods were applied to the data. The results revealed that the models were able to select the main SNP markers for identification of such breeds. Through the intersection of the generated models a subset of 18 markers was identified with greater potential for identification of breeds.

KEYWORDS: Single nucleotide polymorphism, Feature selection, Predictive modeling, Penalized methods.

INTRODUÇÃO

O Brasil possui diversas raças de ovinos que se desenvolveram a partir de raças trazidas pelos colonizadores, as quais adquiriram características específicas de adaptação às condições ambientais, passando a ser conhecidas como locais ou localmente adaptadas. Algumas dessas raças encontram-se ameaçadas de extinção, devido a cruzamentos indiscriminados com animais de raças exóticas (MARIANTE et al., 2009). As raças locais constituem uma importante fonte de informações que pode levar à descoberta de genes envolvidos com determinadas características adaptativas como a resistência a doenças e parasitas.

Para evitar a perda deste importante material genético, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) decidiu incluir algumas raças locais em seus Bancos de Germoplasma, sendo que as raças que possuem maior destaque nacional são: Crioula, Morada Nova e Santa Inês. A constituição desses bancos é feita por meio de avaliação de características morfológicas e produtivas dos animais. Entretanto, essa avaliação está sujeita a falhas, pois alguns animais cruzados mantém características semelhantes àquelas dos animais locais.

Uma forma eficiente de auxiliar na solução para este tipo de problema é empregar tecnologias que fazem uso de marcadores moleculares SNP. Os marcadores SNP constituem uma variação que ocorre em um único nucleotídeo da cadeia de bases nitrogenadas (Adenina, Citosina, Timina e Guanina) do DNA, afetando ou não o fenótipo alvo entre indivíduos de uma espécie. Contudo, as tecnologias para geração destes dados são capazes de genotipar milhares de marcadores SNP por animal.

Assim, realizar a seleção dos marcadores mais informativos 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), surge como uma alternativa promissora, uma vez que essas técnicas são amplamente utilizadas na descoberta de padrões novos e úteis em grandes volumes de dados (HAN et al., 2011). Em particular, técnicas que combinam métodos de seleção de atributos e modelos preditivos capazes de lidar com problemas em que o número de atributos (marcadores) p é muito maior que o número de observações (animais) n, isto é, ($p >> n$).

Muitos trabalhos já foram conduzidos no desenvolvimento de modelos computacionais

e estatísticos para identificação de conjuntos de SNP relacionados com características fenotípicas interessantes em variados organismos, como em bovinos (MOKRY et al., 2013; SUEKAWA et al., 2010) e frangos (GONZÁLEZ-RECIO et al., 2010), porém, são raros os trabalhos abordando o desenvolvimento de modelos envolvendo dados de ovinos. O objetivo deste trabalho foi desenvolver modelos baseados em técnicas de mineração de dados para selecionar os marcadores SNP mais relevantes para as raças de ovinos Crioula, Morada Nova e Santa Inês.

MATERIAL E MÉTODOS

A metodologia utilizada é composta de quatro etapas principais, a saber: entendimento dos dados, preparação dos dados, modelagem e validação dos modelos.

Na etapa de entendimento dos dados, o conjunto de dados analisado foi obtido do Consórcio Internacional do Genoma Ovino (ISGC et al., 2010), sendo composto por 72 animais das raças estudadas (23 animais da raça Crioula, 22 da Morada Nova e 27 da Santa Inês) e 49.034 marcadores SNP por animal.

Na preparação dos dados, realizou-se uma verificação quanto à existência de amostras idênticas dentro do conjunto de dados e de marcadores SNP com valor único de genótipo em todas as raças. Porém, não constatou-se amostras idênticas. Existiam 384 marcadores com valor único em todas as raças, que foram removidos do conjunto de dados final.

Na etapa da modelagem, aplicou-se técnicas que combinam seleção de atributos e desenvolvimento de modelos preditivos. Devido ao elevado número de atributos e o baixo número de registros, técnicas capazes de lidar com esta situação foram empregadas, a saber: LASSO (*Least Absolute Shrinkage and Selection Operator*), Random Forest e Boosting.

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

Primeiramente, aplicou-se a técnica LASSO, e o único parâmetro avaliado foi o intervalo de possíveis valores para o coeficiente de penalização λ (lambda). Como padrão, o valor do intervalo é de 100 valores possíveis (JAMES et al, 2013), obtidos separadamente pelo

algoritmo LASSO, via validação cruzada. Após a aplicação de LASSO, utilizou-se Random Forest para a busca dos SNP mais relevantes associados às raças. Para Random Forest, avaliou-se os parâmetros relacionados ao número de árvores a serem construídas e ao número de atributos selecionados para determinar o *split* em cada nó das árvores. Da mesma forma, a técnica Boosting foi utilizada para produzir um modelo com os marcadores mais relevantes para as raças. Em relação à técnica Boosting, o único parâmetro testado foi o número de árvores a serem desenvolvidas para o modelo final.

Para avaliar o desempenho dos modelos, utilizou-se dois tipos de particionamento dos dados: validação cruzada e *bootstrap*, além da métrica acurácia e do coeficiente Kappa.

RESULTADOS E DISCUSSÃO

Na aplicação do algoritmo LASSO, após a obtenção do valor ótimo de λ, selecionou-se 29 marcadores relevantes, dos quais, cinco se destacaram para Crioula, 12 para Morada Nova e 12 para Santa Inês. Os cinco SNP que se destacaram para Crioula e suas respectivas informações estão descritas na Tabela 1.

Tabela 1: Frequências alélicas dos marcadores SNP selecionados pelo algoritmo LASSO para a raça Crioula

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 nas raças.*

De forma geral, os marcadores mostraram potencial de identificação da raça Crioula, com destaque para quatro SNP pertencentes ao cromossomo X. Observou-se que os SNP da raça Crioula possuem altas diferenças de frequências em relação às outras raças, provavelmente pelo fato desta possuir características mais distintas entre elas (PAIVA, 2005).

Da Tabela 2 observa-se uma frequência relativamente maior dos alelos dos animais Morada Nova na raça Santa Inês, o que pode ser explicado pelo fato dos animais Santa Inês serem originários do cruzamento entre Morada Nova e ovinos do nordeste brasileiro, por isso ovinos da Santa Inês preservem características da Morada Nova (FIGUEIREDO et al., 1990).

Tabela 2: Frequências alélicas dos marcadores SNP, selecionados por LASSO para a raça Morada Nova.

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 raças.*

Entre os marcadores selecionados para Santa Inês (Tabela 3), destaca-se o fato dos marcadores do cromossomo três estarem 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 destaques os marcadores OARX_53305527.1 e s20468.1.

Tabela 3: Frequências alélicas dos marcadores SNP, selecionados pelo algoritmo LASSO para a raça Santa Inês

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	94733688	[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 raças.*

A acurácia obtida utilizando o algoritmo LASSO foi de 100% na predição de novas raças, e o índice Kappa foi igual a 1.

Random Forest gerou uma listagem dos marcadores mais importantes para o modelo de identificação das raças. O melhor resultado obtido foi utilizando os parâmetros fornecidos pelo pacote caret, que resultou em 1.000 árvores e 313 marcadores para *split*. Selecionou-se, então, os 27 melhores SNP classificados, pois a partir desta posição os SNP contribuíam com menos que 2% para o modelo. Mokry et al. (2013) primeiramente selecionaram 1% dos SNP mais relevantes de cada cromossomo e, em seguida, 1% dos SNP mais importantes do subconjunto anterior, e, no final, selecionados 70 SNP pela técnica Random Forest.

A Tabela 4 mostra os marcadores predominantes na raça Crioula. Do conjunto de 13 marcadores identificados para Crioula, quatro também foram identificados por LASSO (OARX_121724022.1, OARX_29830880.1, OARX_78903642.1, s56924.1).

Tabela 4: Frequências alélicas dos marcadores SNP, selecionados por Random Forest para a raça Crioula.

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 raças.*

Tabela 5: Frequências alélicas dos marcadores SNP, selecionados por Random Forest para a raça Morada Nova.

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 Morada Nova do lado esquerdo. ** Frequência do alelo específico nas raças.*

Conforme Tabela 5, o algoritmo Random Forest indicou três marcadores importantes para a raça Morada Nova. Observa-se os SNP OAR1_194627962.1 e OAR2_54691204.1, com frequência acima de 50% na Morada Nova e praticamente ausente nas outras duas raças.

Para Santa Inês, 11 marcadores foram selecionados com altas frequências alélicas (Tabela 6). Destes, quatro estavam presentes no modelo LASSO (OARX_53305527.1, s20468.1, OAR3_165050963.1 e s16949.1), reforçando o potencial dos mesmos.

Tabela 6: Frequências alélicas dos marcadores SNP, selecionados por Random Forest para a raça Santa Inês.

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 raças.*

Para treinamento e teste, foram desenvolvidas e combinadas 1.000 árvores utilizando amostras *bootstrap*. O modelo Random Forest obteve uma acurácia de 99% e Kappa de 0,98.

Tabela 7: Frequências alélicas dos marcadores SNP, selecionados por Boosting para a raça Crioula.

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 raças.*

Na aplicação de Boosting, o parâmetro testado foi o número de árvores de decisão a serem construídas. Selecionou-se, então, os 20 melhores marcadores, pois os SNP a partir desta posição pouco contribuíam (menos que 1%) para o modelo. Entre os 20 marcadores ordenados

por Boosting, seis estavam presentes nos modelos LASSO e Random Forest, dois estavam somente em LASSO e sete somente no modelo Random Forest.

Na lista de marcadores importantes para a raça Crioula, dois deles (OARX_121724022.1 e s56924.1) foram indicados nos dois modelos anteriores, demonstrando o alto potencial destes marcadores (Tabela 7).

Como pode ser visto na Tabela 8, o algoritmo Boosting separou cinco marcadores com maior frequência em Morada Nova, sendo um deles (OAR1_194627962.1) presente nos dois modelos anteriores, resultado que o confirma como um bom discriminante de raças.

Tabela 8: Frequências alélicas dos marcadores SNP, selecionados por Boosting para a raça Morada Nova.

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 raças.*

Na Tabela 9, dentre os marcadores fornecidos pelo modelo 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, demonstrando o potencial desses SNP.

Tabela 9: Frequências alélicas dos marcadores SNP, selecionados por Boosting para a raça Santa Inês.

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 raças.*

A acurácia e o Kappa obtidos pelo modelo, com a combinação dos classificadores ajustados, foi de 100% e 1, respectivamente. Observando esses resultados, pode-se pensar em indícios de *overfitting*, porém os parâmetros ajustados para os algoritmo LASSO e Boosting foram obtidos pelo caret de forma a evitar um super-ajuste do modelo. Este bom desempenho

também foi obtido em (GONZÁLEZ-RECIO, et al., 2010), em que o algoritmo L2-Boosting foi utilizado em dois conjuntos de SNP, obtendo alta precisão nas predições.

Considerando-se apenas marcadores selecionados por dois e três modelos, um total de 18 SNP (Tabela 10) demonstra ter grande potencial na identificação das raças estudadas. Esse número de marcadores é próximo aos resultados de trabalhos relacionados à identificação racial em bovinos, como em Suekawa et al. (2010).

Tabela 10: Marcadores SNP selecionados pelos modelos e suas raças predominantes.

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.*

CONCLUSÕES

Os modelos obtidos com aplicação das três técnicas escolhidas revelaram resultados promissores para a seleção dos marcadores SNP mais informativos das raças estudadas. Visto que o conjunto de dados utilizado possui um elevado número de atributos, as técnicas aplicadas reduziram o número de SNP para menos de 0,2%. Na intersecção dos modelos, foram encontrados 18 SNP 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, assim como poderão ser utilizados por diversos segmentos ligados à ovinocultura.

REFERÊNCIAS BIBLIOGRÁFICAS

FIGUEIREDO, E.A.P.; SHELTON, M.; BARBIERI, M.E. Avaliable genetic resources: the origin and classification of the world's sheep. In: **Hair Sheep Production in Tropical and Subtropical Regions**, Davis, USA, p. 25-36, 1990.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1-22, 2010.

GONZÁLEZ-RECIO O.; WEIGEL K.A.; GIANOLA D.; NAYA H.; ROSA, G. J.M. L2-Boosting Algorithm Applied to High-Dimensional Problems in Genomic Selection. **Genetics Research**, v. 92, n. 03, p. 227–237, 2010.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, ed. 3, San Francisco, CA, USA, 2011.

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. **Anim. Genet.**, n.41, p.449–453, 2010.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: With Applications in R. Ed. Springer, London, 429 p., 2013.

KUHN, M. caret: Classification and Regression Training. R package version 5.16-24, 2013.

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

MARIANTE, A.S.; ALBUQUERQUE, M.S. M.; EGITO, A.A.; MCMANUS, C.; LOPES, M.A.; PAIVA, S.R. Present status of the conservation of livestock genetic resources in Brazil. **Livestock Sci**., v.120, n.3, p.204-212, 2009.

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**. Tese (Doutorado)- Universidade Federal de Viçosa, Viçosa, 2005. 108p.

RIDGEWAY, G. gbm: Generalized Boosted Regression Models. **R package version 2.1**, 2013.

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**, v. 85, n. 2, p. 285-8, jun. 2010.