Universidade de Brasília
Departamento de Estatística

Análise de Ampla Associação do Genoma: Redução de Dimensionalidade via Seleçao por Torneios e Análises Indicativas

Cayan Atreio Portela Bárcena Saavedra

Monografia apresentada para obtenção
do título de Bacharel em Estatística.

Brasília
2015

Cayan Atreio Portela Bárcena Saavedra

Análise de Ampla Associação do Genoma: Redução de Dimensionalidade via Seleçao por Torneios e Análises Indicativas

Orientador:
Prof. Dr. **Eduardo Monteiro de Castro Gomes**

Coorientadora:
Pesquisadora Msc. **Joseane Padilha da Silva**

Monografia apresentada para obtenção
do título de Bacharel em Estatística.

Brasília
2015

DEDICATÓRIA

A Deus.

Aos meus amados pais,
Francisca Portela Correia *e* ***Juan Manuel Bárcena Saavedra****, pelo amor e apoio incondicional.*

Ao meu avô,
Alfredo Bárcena Ayala *(in memoriam), pela iluminação e inspiração. Nunca mais terei a chance de lhe agradecer, mas posso viver de maneira a honrá-lo.*

AGRADECIMENTOS

À minha mãe, **Francisca Portela**, pelo exemplo de vida. Minha maior motivação é você ter orgulho de mim.

Aos meus irmão, **Andrey Portela**, **Hanaa Portela** e **William Portela**, pelo companherismo, incentivo, apoio e bons momentos. Vocês são minha base.

Às minhas amadas sobrinhas, **Ana Luiza Olivera Portela** e **Júlia Oliveira Portela**, meus anjos terrenos. Vocês terão um futuro brilhante.

À minha cunhada, **Lucília Raposo Oliveira Portela**, pelos conselho, por sempre estar com nossa família e pelas duas lindas meninas que você nos trouxe.

À minha namorada, **Thais Torres**, por todo apoio e pela maravilhosa companhia em todos os momentos.

A todos os meus amigos da **105 norte** e aos amigos de graduação, em especial, **Cristiano**, **Frederico**, **Ian** e **Felipe**.

Ao meu orientador, Professor Dr. **Eduardo Monteiro de Castro Gomes**, pelos conselhos, ensinamentos, paciência e prestatividade.

À minha supervisora e coorientora **Joseane Padilha**, pelos ensinamentos ao longo do ano e ao Dr. **Luis Palhares**, pelos conselhos, ajuda e apoio durante o período de elaboração deste trabalho.

A todas as pessoas que contribuiram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

RESUMO

Análise de Ampla Associação do Genoma: Redução de Dimensionalidade via Seleçao por Torneios e Análises Indicativas

O presente trabalho tem como objetivo identificar possíveis SNPs responsáveis pela cor da pelagem de ovinos da raça Morada Nova. Dados de genotipagem de 61 animais, consistindo em 54.412 SNPs e característica fenotípica de interesse (cor da pelagem), foram processados visando identificar os SNPs causadores do fenômeno. Três tipos de abordagem foram realizados para explicar a relação proposta; (1) modelos logísticos simples, (2) seleção por torneios e (3) modelos logísticos multivariados, montados a partir de inidcativas das análises univariadas. Primeiramente, foram construídos modelos logísticos simples, sendo a variável resposta a cor da pelagem (1 para vermelho e 0 para não vermelho) e a variável preditora um único SNP do conjunto de todos os disponíveis, visando a identificação de regiões promissoras. SNPs influentes que se encontram próximos no decorrer do genoma, caracterizam regiões promissoras. Estes SNPs foram identificados, e a partir destes, modelos multivariados foram construídos. Além desta abordagem, uma seleção por torneios foi realizada para a construção de um modelo final. Devido a eventuais cancelamentos na matriz de delineamento, foram excluídos SNPs que apresentaram valor faltante, restando então, 1.522 SNPs. Mesmo após uma filtragem inicial, o número de SNPs restantes continuou elevado de modo a acarretar problemas para o uso de métodos estatísticos convencionais. Para contornar os problemas decorrentes da explosão de graus de liberdade, causada pela grande quantidade de SNPs, foi implementada uma seleção por torneios, que consiste em subdividir o conjunto de convariáveis iniciais em subconjuntos disjuntos, utilizando modelo de regressão logística multivariada em cada subconjunto e selecionando SNPs menos promissores, usando p-valor como critério. O procedimento é repetido até que se chegue a um número de SNPs desejado, previamente estabelecido, para uma modelagem final. A partir das duas abordagens foram montados modelos logisticos multivariados, visando obter um bom poder de predição. Tais modelos foram comparados pelo Critério de Informação de Akaike (AIC) e pelo poder de predição. Ao final, os cromossomos 9 e 13, bem como o modelo obtido através da seleção por torneios, obtiveram um poder de predição de 100%, acusando possíveis regiões responsáveis pela característica desejada e demonstrando que a seleção por torneios possui um bom poder para escolha de variáveis.

Palavras-chave: Modelos logísticos, Seleção por torneios, SNP.

1 INTRODUÇÃO

O agronegócio possui expressiva participação na economia brasileira, setor este que é responsável por aproximadamente um quarto do PIB nacional e possui estimativa de crescimento de 2,8% em 2015, segundo o Centro de Estudos Avançados em Economia Aplicada (Cepea, ESALQ/USP). Fatores climáticos favoráveis e diversidades regionais tornam o Brasil um país com vocação natural para o setor, com grande importância no cenário internacional.

Dentre os segmentos do agronegócio, os ovinos possuem importante participação por terem capacidade de fornecer diversos produtos tais como lã, couro, leite e sendo todos de excelente qualidade. A ovinocultura apresenta uma diversidade de raças, no qual se destaca a raça Morada-Nova pela adaptabilidade a regiões quentes e baixo custo de produção. Segundo Rede Morada Nova (2015),

> *a raça Morada Nova apresenta, ainda, elevado valor adaptativo às condições de produção do semi-árido nordestino, sendo capaz de apresentar elevadas taxas de fertilidade, mesmo sob condições pouco favoráveis. Portanto, a raça Morada Nova se constitui em importante material genético para o produtor rural do Nordeste brasileiro. (REDE MORADA NOVA, 2015).*

O projeto da Rede Morada Nova tem como objetivo principal promover ações de pesquisa e desenvolvimento de forma a melhor caracterizar a raça Morada Nova e seus produtos e fundar as bases para um amplo programa de conservação e melhoramento genético, dando valor de uso à raça e minimizando os riscos de descaracterização e desaparecimento (REDE MORADA NOVA, 2015).

Diversas instituições participam deste projeto, que se iniciou em 2008, tais como EMBRAPA, Associação Brasileira dos Criadores de Ovinos da Raça Morada Nova (ABMOVA), Federação de Agricultura e Pecuária do Estado do Ceará (FAEC), Universidade de Brasília (UnB), Universidade Estadual do Ceará (UECE), dentre outras (REDE MORADA NOVA, 2015).

Ovinos da raça Morada Nova são animais de médio porte, deslanados (possuem pelo curto ao invés de lã) e apresentam diversidade na cor da pelagem, podendo ser vermelhas (coloração predominante), brancas ou pretas. Sob o ponto de vista econômico é mais interessante que os animais apresentem pelagem do tipo vermelha, pela maior atratividade de comercialização deste produto. Através de técnicas como melhoramento genetico, pode-se aumentar a obtenção de pelagem vermelha na ovelhas. Em linhas gerais, o melhoramento genético pode ser visto como um conjunto de técnicas que são utilizadas para selecionar indivíduos (animais ou plantas) que apresentem genes que potencializem a presença de características desejáveis, tais como cor da pelagem, peso médio, altura média, peso médio dos frutos produzidos, resistência a ambientes secos e/ou úmidos, formato das folhas, etc.

Desta forma, estudos envolvendo melhoramento genético animal, em geral buscam aumentar a precisão em seleções de cruzamento baseado em características (fenótipos) de interesse econômico (FAN et. al. 2008). Assim, esses estudos se tornam de fundamental importância, aprimorando a competitividade no setor agropecuário, que é um dos mais importantes na economia brasileira.

Na realidade, o melhoramento genético é uma prática exercida pelo homem desde os primórdios, quando sem qualquer conhecimento da existência dos genes, realizava cruzamentos seletivos (artificiais) de animais e plantas que apresentavam as características de interesse com o intuito de se manifestarem repetidas nas gerações seguintes. A elucidação da estrutura do DNA na década de 1950 permitiu que o processo de melhoramento genético se tornasse mais acurado. Segundo Brito et. al. (2014),

> *uma explosão de descobertas nas áreas de métodos de análise do DNA, de equipamentos sofisticados de análise de grande quantidade de amostras, de ferramentas estatísticas e de informática (bioinformática) propiciaram o surgimento da Genômica, ciência que trata do genoma completo dos diferentes organismos. O que está ocorrendo é um espantoso acúmulo de dados "genômicos"que está à disposição dos pesquisadores para serem interpretados e utilizados, o que deixa em aberto uma enorme trilha a ser percorrida nos próximos anos (BRITO et. al., 2014).*

O "espantoso acúmulo de dados genômicos"citado por Brito et. al. (2014) pode e deve ser tratado por diversas técnicas matemáticas, estatísticas e computacionais. De fato, é impensável o tratamento das informações genômicas do indivíduo sem apoio de tais técnicas. Afinal, em essência, o desvelamento das diversas partes do DNA, enquanto estrutura de dados que registra todas as características (fenótipo) do indivíduo, envolve o estudo exaustivo de diversas combinações de várias "partes"do DNA, na tentativa de descobrir quais regiões são responsáveis pelas diversas características do indivíduo. Ocorre uma então, uma explosão combinatorial de possibilidades, que só podem ser efetivamente avaliadas com uso das técnicas de natureza acima citadas.

DNA é a sigla usada para designar o ácido desoxirribonucleico (do inglês, *deoxyribonucleic acid*) e contém as informações genéticas que coordenam o funcionamento de todos os seres vivos e de alguns vírus. Em outras palavras, é o DNA responsável pelas características físicas (peso, altura, cor da pele, cor dos olhos, etc) e pelas características comportamentais (agressividade, resistência a locais de altas/baixas temperaturas, etc) dos seres vivos. As característica que podem ser vistas ou percebidas são denominadas características fenotípicas, ou fenótipos.

O DNA localiza-se no núcleo das células dos indivíduos e está disposto ao longo dos cromossomos. Na Figura 1 observa-se uma célula, seu núcleo com diversos cromossomos (em formato de "X") e um detalhe da estrutura do DNA.

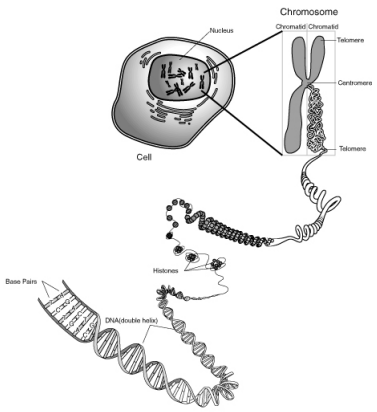

Figura 1 – Uma célula e seu núcleo realçado à presença de cromossomos.
Fonte: http://euquerobiologia.com.br/2014/04/o-que-e-dna.html

Sob o ponto de vista estrutural, o DNA é uma longa molécula formada por uma cadeia de diversos nucleotídeos. Um nucleotídeo, por sua vez, é uma estrutura química formada por (1) um grupo fosfato, (2) um açúcar ou pentose (no caso do DNA é a desoxirribose) e (3) uma base nitrogenada que pode ser a adenina (A), guanina (G), citosina (C) ou timina (T). Na Figura 2 observa-se um esquema representativo da estrutura do nucleotídeo.

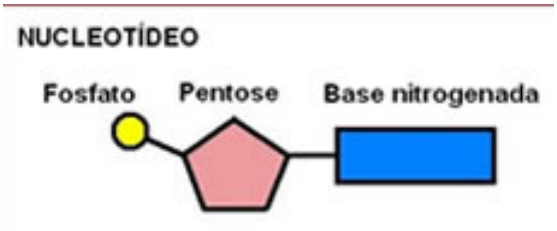

Figura 2 – Esquema da estrutura química de um nucleotídeo.
Fonte: http://www.brasilescola.com/biologia/nucleotideo.htm

O DNA pode então ser visto como uma longa sequência nucleotídeos, que se diferenciam entre si pela base nitrogenada (A, G, C ou T), ou seja, uma extensa fita de poliucleotídeos, como demonstra a Figura 3.

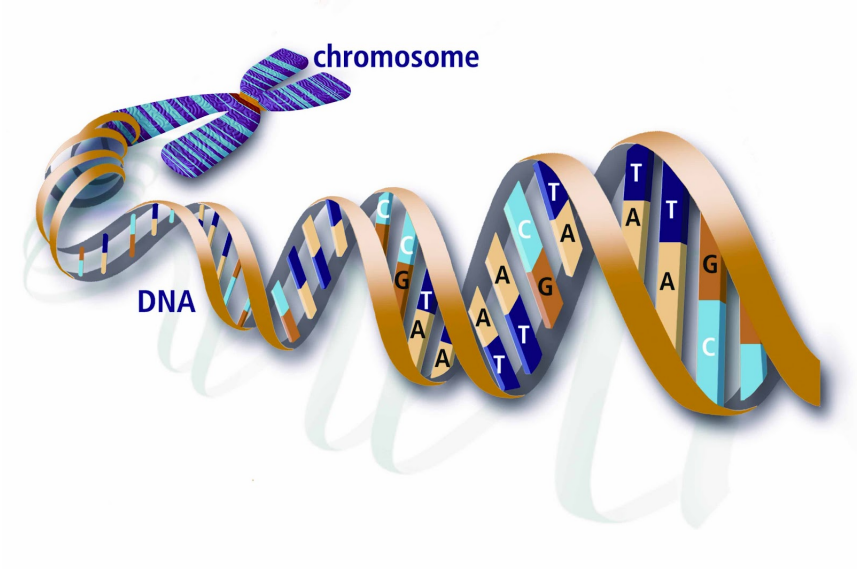

Figura 3 – O DNA visto como uma fita de sequências de nucleotídeos.
Fonte: http://bioelogia.blogspot.com.br/2012/07/genetica-parte-2.html

É esta longa fita de polinucleotídeos que está disposta ao longo dos cromossomos dos indivíduos. E existem partes desta fita que são responsáveis pelas características fenotípicas dos mesmos. Estas partes denominam-se genes. São portanto os genes que contêm o código das diversas características fenotípicas dos indivíduos. Na Figura 4 observa-se um cromossomo indicando alguns genes.

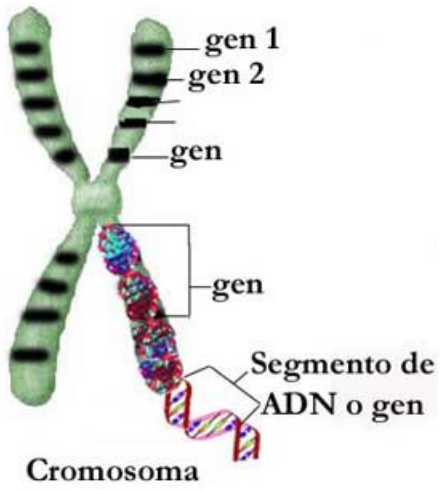

Figura 4 – O gen localizado ao longo de um cromossomo.
Fonte: http://definicion.de/gen/

Os genes são então, sequências de nucleotídeos que determinam características fenotípicas do indivíduos. A razão pela qual indivíduos de uma mesma população apresentam diferentes características fenotípicos é explicada pelas inúmeras diferenças entre os genes. As alterações na cadeia de DNA que causam essas diferenças são denominadas polimorfismos. Dentre os diversos tipos de polimorfismos, os mais abrangentes são os polimorfismos de base única, denominados SNPs (Nucleotide Polimorphism), que consistem na mudança de um único nucleotídeo na sequência de DNA, ou seja, a mudança de apenas uma "letra"no decorrer de uma sequência genômica. A mudança desta única base na sequência genômica causa modificação no funcionamento dos genes, por meio da alteração dos aminoácidos produzidos pelos mesmos, gerando assim, diferentes tipos de respostas fenotípicas (HEATON et al. 2002).

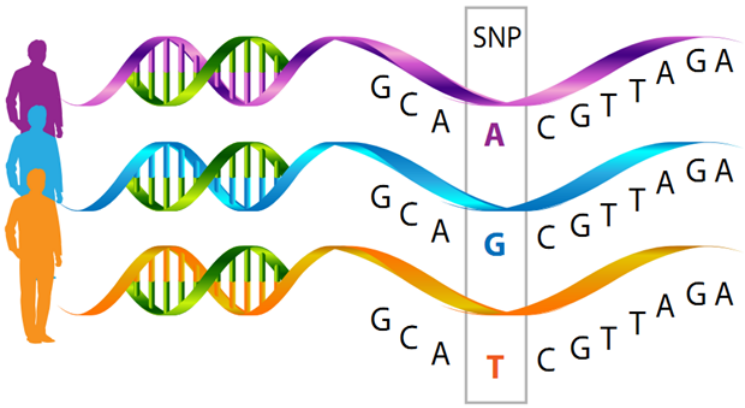

Figura 5 – Polimorfismos de apenas uma base nitrogenada caracterizam SNP.
Fonte: http://biogeniq.ca/en/snp/

Polimorfismos de base única (SNP), correspondem ao marcador molecular (fragmentos de sequências de DNA capazes de identificar diversos tipos de polimorfismo) mais abundante no genoma. Estudos com humanos e espécies animais, mostram que pode haver milhões de polimorfismos SNP no genoma de um indivíduo (Bovine Genome Sequencing and Analysis Consortium, 2009; Li et al., 2008). Além dos marcadores SNP serem abundantes, suas bases moleculares permitem que haja uma distribuição homogênea de SNPs pelo genoma (R. Bras. Zootec. vol.38 no.spe Viçosa, 2009). SNPs também são extremamente úteis na criação de mapas genéticos de alta densidade, o que não pode ser atingido diante de outro marcador molecular (JEHAN & LEKHANPAUL, 2006). Normalmente, os marcadores SNP são bi-alélicos, ou seja, geralmente são encontrados apenas dois níveis alélicos em uma certa região. Devido à sua abundância no genoma e pelo fato de serem estáveis, os SNPs são considerados ótimos marcadores para estudos genéticos e mais viáveis em estudos populacionais (JEHAN & LAKHANPAUL, 2006).

Em virtude dos recentes avanços tecnológicos, estudos genéticos envolvendo marcadores moleculares em programas de melhoramento genético obtiveram progressos significativos, uma vez que o custo do sequenciamento genético reduziu drasticamente levando ao aumento da quantidade de banco de dados deste tipo e facilitando o acesso aos mesmos, fato este, que foi um incentivo ao desenvolvimento de estudos de análise de ampla associação do genoma - GWAS (do inglês, *Genome Wide Association Studies*). Os

primeiros estudos feitos em humanos, por exemplo, envolvendo análise de ampla associação do genoma (GWAS), que visa identificar a relação entre genótipo e fenótipo, viabilizaram a identificação de regiões dos genes responsáveis por afetar doenças como câncer de próstata, transtorno bipolar, doença arterial, hipertensão, diabetes do tipo I e II (Yeager et al. 2007; WTCCC 2007) e características quantitativas, como a altura (VISSCHER, 2008).

Estudos GWAS (*Genome Wide Association Studies*), são estudos que buscam associação entre os polimorfismos do genoma e algum fenótipo de interessse. Porém, existem milhares de polimorfismos no DNA (sendo os SNPs, o tipo de polimorfismo mais abundante), o que torna o problema complexo, devido à sua dimensão. Sob o ponto de vista estatístico, estes estudos buscam detectar dentre as milhares de covariáveis possíveis (SNPs), quais realmente influenciam na resposta desejada, ou seja, no fenótipo. Mesmo com sua complexidade, estudos do tipo, que utilizam sequenciamentos do tipo SNP através de painéis de genotipagem de alta densidade, têm se mostrado uma poderosa ferramenta para identificar variações genéticas responsáveis por características fenotípicas de interesse (CORNELIS et al. 2010).

Análises GWAS apresentam em média 50 mil SNPs, o que torna o modelo estatístico de alta dimensão, uma vez que cada SNP é um candidato a interferir no fenótipo. Surge então, um problema referente à escassez de graus de liberdade para estimar o efeito dos marcadores. Uma alternativa para contornar tal situação, é o emprego de metodologias de regressão ridge (RR de Whittaker et al., 2000) ou assumir os marcadores como efeitos aleatórios, pois deste modo, tal ajuste não consome graus de liberdade, logo, os efeitos dos marcadores poderiam ser estimados simultaneamente (RESENDE, M. D. V. et al., 2012). Outra opção, proposta por Chen e Chen (2009), é a utilização de seleção por torneios para redução da dimensionalidade do modelo, excluindo covariáveis menos explicativas.

1.1 Justificativa

A melhor compreensão da relação entre genótipo e fenótipo através de análises estatísticas fornece apoio fundamental ao melhoramento genético, contribuindo para fatores de natureza econômica e substanciais à vida. Do ponto de vista econômico, o esclarecimento dessa relação pode significar otimização na produção e ganhos em uma importante área do setor econômico brasileiro, o agronegócio. Neste sentido, a demanda para o desenvolvimento de estudos de análise de ampla associação do genoma com apoio de profissionais de áreas quantitativas vem crescendo continuamente, em virtude da importância para os aspectos citados.

Por exemplo, a análise estatística para a identificação de regiões genômicas influentes na cor da pelagem de ovinos da raça Morada Nova, possibilita um ganho em seleções de cruzamento e melhoramento genético, campos que possuem influência econômica direta na ovinocultura e consequentemente na agropecuária brasileira de modo geral.

Com base nos argumentos descritos, o trabalho busca o uso de ferramentas quantitativas que viabilizem este tipo de estudo, mantendo um bom nível de explicação.

1.2 Objetivos

1.2.1 Objetivo geral

O presente trabalho tem como objetivo geral o desenvolvimento de uma análise de associação do genoma em larga escala (GWAS), explorando ferramentas estatísticas e computacionais capazes de abordar o problema proposto. As técnicas são utilizadas visando a detecção de polimorfismos de base única (SNP) no DNA que estejam associados à pelagem vermelha em ovinos da raça Morada Nova, para seleção genômica e melhoramento genético.

1.2.2 Objetivo específico

Especificamente, objetiva-se identificar e localizar SNPs no decorrer do genoma que de fato estejam associados à característica de interesse (pelagem vermelha) e predição de características fenotípicas de acordo com a sequência genômica.

Para tanto, é necessário buscar alternativas que driblem as dificuldades encontradas em modelos de alta dimensão. Neste trabalho, ferramentas estatísticas são combinadas com métodos computacionais para a redução da dimensionalidade do modelo, visando um modelo reduzido com SNPs mais promissores e com melhor poder de predição.

2 REVISÃO DE LITERATURA

2.1 Modelos de alta dimensão

Aplicações de estudos genéticos quantitativos envolvendo milhares de marcadores moleculares, caracterizam-se por natureza em um cenário de modelos de alta dimensão ($p >> n$). Devido a esta alta dimensão, técnicas estatísticas mais convencionais, como regressão linear, não se aplicam, devido à escassez de graus de liberdade para a estimação dos parâmetros. Outo desafio referente a este cenário, trata-se da alta correlação entre as covariáveis, que manifesta-se apesar de serem estocasticamente independentes. Fan & LV (2008) demonstram através de simulações, que a correlação aumenta quando P se torna maior, mesmo que as covariáveis tenham sido geradas independentemente, fazendo com que a correlação espúria também seja um problema. Fan & Fan (2008), mostram que a predição feita por um modelo sem prévia seleção, incluindo todas as covariáveis, pode ser tão ruim quanto classificação ao acaso, devido ao acumulo de perturbação, fato este, que torna desafiador a seleção de apenas algumas covariáveis dentre as milhares possíveis, para a montagem de um modelo de predição. Bickel (2008), aponta como os principais objetivos de uma modelagem em alta dimensão serem (a) construir um modelo com maior poder preditivo possível para futuras observações e (b) possuir o maior fator de explicação entre variáveis preditoras e variável resposta, por motivos científicos.

Como uma alternativa ao desafio proporcionado por tais modelos, Chen e Chen (2009), propuseram uma redução de dimensionalidade através de seleção de variáveis realizadas por torneios, utilizando como critério um modelo de verossimilhança penalizada. Alves (2014) utilizou da técnica proposta por Chen e Chen (2009), utilizando um fator penalizador, caracterizando a regressão LASSO (Least Absolute Shrinkage and Selection Operator) e usando também a regressão LASSO com todos os marcadores em um único modelo múltiplo, sem seleção por torneios, concluindo em seus resultados, que os métodos não são muito diferente em simulações. Porém, a autora constata que os torneios com regressões múltiplas possuem uma vantagem computacional pela paralelização direta da análise, demonstrando ser um processo computacional totalmente viável e consideravelmente rápido.

2.2 Seleção de variáveis

2.2.1 Seleção por torneios

Chen e Chen (2009) proporam uma alternativa para modelos de alta dimensão, denominada seleção por torneios. Esta alternativa consiste em subdividir o conjunto de covariáveis em grupos aleatórios disjuntos, sendo que, em cada grupo é aplicado um modelo de verissimilhança penalizada e um determinado número de covariáveis é selecionado, passando à próxima fase. O processo é repetido sobre as covariáveis selecionadas até que se consiga reduzi-las a um nível desejado, previamente estabelecido.

Ao adaptar o método de seleção por torneios para o uso da regressão logística, neste estudo, é aplicado um modelo logístico convencional (sem fator penalizador) multivariado em cada subconjunto disjunto e em cada passo, descarta-se uma covariável utilizando como critério o p-valor calculado a partir de seu efeito marginal. O tamanho dos subconjuntos é tal que se tenha graus liberdade suficiente para estimar os parâmetros dos modelos multivariados.

Seja p o número de covariáveis no modelo. Se denotarmos C como o conjunto de inteiros de 1 a p e o dividirmos em k subconjuntos disjuntos, obtemos:

$$C = \bigcup_{i=1}^{k} C_i = C_1 \cup C_2 \cup \cdots \cup C_k$$

Logo, em cada subconjunto $C_1 \cdots C_k$ é aplicado um modelo logístico e o algoritmo prossegue de acordo com os seguintes passos:

1. Estipula-se o número de covariáveis a se obter ao final do torneio.
2. É aplicado um modelo logístico em cada subconjunto.
3. O modelo é analisado e retira-se a covariável com maior p-valor.
4. Aplica-se, novamente, um modelo no mesmo subconjunto disjunto, que possui 1 covariável a menos em relação ao passo anterior.
5. O procedimento é repetido até que se obtenha o número de covariáveis desejado.

Após reduzir o numero inicial de covariáveis p ($>> n$) para j ($< n$), então métodos convencionais de seleção de modelo podem ser utilizados para a escolha do modelo final.

2.3 Regressão logística

A regressão logística se distingue de técnicas de regressão mais conhecidas, como por exemplo a regressão linear, principalmente pelo tipo de variável resposta que é abordada.

Segundo Oliveira (2011), a regressão logística caracteriza-se geralmente, no caso em que a natureza da variavel resposta (Y) é do tipo binária, assumindo apenas valores 0 ou 1. Com isso, a probabilidade de ocorrência de um evento pode ser estimada diretamente e a variavel Y assumirá distribuição Bernoulli. Logo, temos que,

$$Y \sim Bernoulli(\theta).$$

Os valores assumidos por Y, 1 ou 0, representam a probabilidade de sucesso e fracasso, respectivamente, de modo que essas probabilidades podem ser expressas por

$$P(Y_i = 1) = \theta \quad \text{e} \quad P(Y_i = 0) = 1 - \theta,$$

com $E[Y_i] = \theta$ e $Var[Y_i] = \theta(1 - \theta)$.

Assumindo $\mathbf{X} = (1, x_1, x_2, ..., x_p)$, como um vetor coluna de variáveis observáveis, com a primeira coluna igual a 1 (referente ao intercepto) e $\beta = (\beta_0, \beta_1, ..., \beta_p)$ como um vetor de parâmetros estimáveis, podemos abordar o modelo através da concepção de uma análise de regressão, ao admitir que $P(Y_i = 1)$ esteja relacionado com o vetor $\mathbf{X}$ de variáveis explicativas. Todavia, θ assume valores no intervalo fechado [0,1], de modo que, seria incorreto admitir uma relação linear do tipo $\theta(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, pois tal função, poderia assumir valores no intervalo ($-\infty, +\infty$).

Para isso, utiliza-se a transformação logito, com o intuito de linearizar a função e evitar o problema de restrição, já que valores referentes a probabilidade, estão contidos no

intervalo fechado [0,1]. Logo, resulta-se na seguinte função:

$$logito[\theta(x)] = ln\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Tal função apresenta características desejáveis, como ser linear em seus parâmetros e poder variar de $-\infty$ a $+\infty$. Ao exponencializar a equação acima, para isolar $\theta(x)$, pode-se obter a probabilidade de sucesso condicionada ao vetor coluna $\mathbf{X}$ de variáveis observadas, de modo que:

$$P(Y = 1|X = x) = E(Y|x) = \theta(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}$$

2.4 Estimativa dos parâmetros pelo método de máxima verossimilhança

Dado que os erros do modelo logístico não satisfazem pressupostos da regressão linear, o método de mínimos quadrados não fornece estimativas ótimas. Por conseguinte, utiliza-se o método de máxima verossimilhança para obter tais estimativas, que tem por objetivo estimar os coeficientes que maximizem a função de verossimilhança.

Considerando $\mathbf{Y} = (y_1, \ldots, y_n)$ como um vetor de variável resposta, onde cada y_i apresenta valores 1 ou 0 (sucesso e fracasso, respectivamente), a função de verossimilhança é a densidade conjunta dos $y's$, representado por $L(\theta|Y)$. Logo,

$$L(\theta|Y) = p(y_1|\theta) \times \cdots \times p(y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

Como a distribuição assumida pela variavel resposta na regressão logística, é uma distribuição bernoulli, temos que:

$$p(y|\theta) = \theta^y \times (1-\theta)^{1-y}, \qquad \text{y = 0,1}$$

Com isso, a função de verossimilhança é dada por:

$$L(\theta|y_1, ..., y_n) = \prod_{i=1}^{n} \theta^{y_i} \times (1-\theta)^{1-y_i}$$

Considerando que $\theta(x)$ é uma função em relação à matriz de covariáveis observadas, representada pelos vetores coluna $\mathbf{X} = (1, x_1, \ldots, x_p)$, consequentemente, a função de verossimilhança maximizada pelo vetor de parâmetros estimados $\beta = (\beta_0, \ldots, \beta_p)$ para o modelo logístico, é representada por:

$$L(\beta|Y) = \prod_{i=1}^{n} \theta(x_i)^{y_i} \times (1 - \theta(x_i))^{(1-y_i)}$$

$$\text{com } \theta(x_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Por questões de conveniência analítica utiliza-se a função de log-verossimilhança, denotada por $l(\beta; y)$, para obter o vetor de parâmetros estimáveis. Tal função é obtida ao aplicar o logaritmo neperiano na função de verossimilhança, de modo que:

$$l(\beta|Y) = ln[L(\beta|Y)]$$

Logo, no caso da regressão logística obtem-se,

$$l(\beta|Y) = \sum_{i=1}^{n} [y_i ln[\theta(x_i)] + (1 - y_i) ln[1 - \theta(x_i)]]$$

O vetor de parâmetros que maximiza $l(\beta|Y)$ e consequentemente $L(\beta|Y)$, é obtido igualando a zero as derivadas parciais (em relação a cada parâmetro) de $l(\beta|Y)$, resultando em um sistema de equações não-lineares. Para resolver esse sistema, recorre-se ao uso de métodos iterativos.

2.5 Deviance

O teste de qualidade de ajuste deviance (*deviance goodness of fit test*) mede a qualidade de ajuste entre dois modelo, considerando um modelo completo, com mais parâmetros, e um modelo reduzido, com menos parâmetros, de modo a identificar o modelo mais indicado.

Segundo Souza (2006), o processo de ajuste de um modelo consiste em propor um pequeno número de parâmetros, de maneira que resuma toda a informação da amostra. Dentre as várias estatísticas existentes que propõem-se a medir a discrepância entre os dois

modelos (completo e reduzido), destaca-se a estatística *deviance*, baseada na função de verossimilhança, proposta por Nelder e Wedderburn (1972). A proposta dos autores é baseada na comparação do valor da verossimilhança de um certo modelo proposto, com P parâmetros, com um modelo saturado. O modelo saturado é ajustado por um número de parâmetros até o equivalente ao tamanho da amostra, modelo este, que impõe toda a variação ao componente sistemático, reproduzindo exatamente os dados (Souza, 2006). Logo, a abordagem dos autores consiste em tomar o negativo de duas vezes o logaritmo natural da verossimilhança. Sendo sempre positiva, um melhor ajuste é representado por um menor valor. Logo, a estatística *deviance*, é definida como:

$$D = -2\ln[L(\widehat{\beta_0}, \ldots, \widehat{\beta_p})]$$

Caracterizando então, um teste de razão de verossimilhança generalizado. Ao considerar as adaptações para modelos logísticos, a estatística *deviance*, segue da seguinte maneira:

$$D = -2\sum_{i=1}^{n}[y_i ln(\theta) + (1 - y_i) ln(1 - \theta)]$$

A estatística *deviance* possui distribuição assintótica dada por (Collet, 1991):

$$D \sim \chi^2_{n-p}$$

onde p é o numero de parâmetros no modelo.

2.6 Deviance parcial

Para o cálculo dos efeitos marginais para uma dada covariável, utilizados na seleção por torneios neste estudo, calcula-se então as estimativas de máxima verossimilhança e a deviance dos modelos sem a covariável e com a covariável. A diferença entre a deviance dos modelos, denominada *deviance* parcial, é dada por:

$$DEV(X_1, X_2, \ldots, X_p | X_1, X_2, \ldots, X_{p-1}) = DEV(X_1, X_2, \ldots, X_{p-1}) - DEV(X_1, X_2, \ldots, X_p)$$

A *deviance* parcial, segue assintoticamente distribuição qui quadrado, com graus de liberdade referente a diferença do número de parâmetros nos modelos, possibilitando testar

a significância dos mesmos. Caso a deviance dos modelos sejam próximas, conclui-se que os parâmetros a mais não contribuem no modelo. Caso contrário, indica que os parâmetros são significativos, sugerindo a manutenção dos mesmos. Utilizando a aproximação para a distribuição qui-quadrado, uma covariável é considerada significativa, caso:

$$DEV(X_1, X_2, \ldots, X_p | X_1, X_2, \ldots, X_{p-1}) > \chi^2_{1-\alpha, n-p}$$

Assim, é possível estimar a contribuição dos parâmetros no modelo e concluir a respeito da manutenção de uma certa covariável.

2.7 Critério de informação akaike (AIC)

O critério de informação Akaike (Akaike, 1973) é utilizado para a caracterização de modelos apropriados, baseado na adequação aos dados e na ordem do modelo. O critério leva em consideração o valor maximizado da função logaritmo de verossimilhança e uma penalização referente ao número de parâmetros no modelo. Logo, o AIC é dado por:

$$AIC = -2lnL(\widehat{\theta}) + 2(p),$$

$L(\widehat{\theta})$ = função de máxima verossimilhança,
p = número de parâmetros no modelo.

Apesar do critério não possuir um valor máximo ou mínimo, sua interpretação é relativamente fácil, onde um menor valor do AIC calculado consiste em um melhor ajuste para o modelo.

3 METODOLOGIA

3.1 Material

A base de dados utilizada neste estudo foi cedida pela EMBRAPA (Cenargen) e coletada em março de 2014, através de marcadores moleculares do tipo SNP, utilizando-se painel de alta densidade Illumina Ovine HD assay realizados pela empresa Illumina (San Diego, CA, USA) especializada em genotipagem em alta tecnologia, resultando em 54.412 SNPs genotipados de 175 ovinos.

Na base original, cada linha continha a informação alélica de cada SNP para cada observação (ovelha). Sendo assim, cada observação era repetida 54.412 vezes, resultando em aproximadamente 9,5 milhões de linhas. A base foi transformada para um modelo mais conveniente, sendo as observações representadas por linhas (apenas uma linha para cada observção) e os SNPs representados por colunas (apenas uma coluna para cada SNP), formando uma tabela de 175 linhas e 54.412 colunas.

A identificação dos níveis alélicos em cada SNP é realizada por marcadores moleculares, através de um sinal emitido após o processamento em laboratório de fragmentos do DNA. Com isso, a qualidade da amostra depende deste sinal, de modo que um sinal de baixa intensidade compromete a identificação da amostra, causando a existência de valores faltantes na base de dados.

Inicialmente, os dados foram submetidos a um controle de qualidade. Neste processo, foram retiradas observações (ovelhas) que não possuem informação referente à cor da pelagem (variável resposta) e SNPs (covariáveis) em que não foi possível identificar pelo menos 90% dos alelos, ou seja, covariáveis que apresentaram mais de 10% de valores perdidos, que apresentaram apenas um tipo de alelo, ocasionando falta de variabilidade e SNPs totalmente idênticos. Este primeiro processo resultou em uma base de 61 observações e 52.622 SNPs.

3.2 Métodos

O presente estudo abordou o problema proposto por três maneiras; (1) modelos logísticos simples considerando 1 SNP por vez, denominados modelos univariados, (2) via seleção por torneios e (3) através da escolha de modelos multivariados (que consideram vários SNPs simultaneamente) por indicativas das análises univariadas.

Os modelos logísticos foram montados com a variável resposta sendo a cor da pelagem (1 para vermelho e 0 para não vermelho) e as covariáveis, ou variáveis preditoras, os SNPs. Como os SNPs possuem até três níveis alélicos (AA, AB ou BB), fez-se necessário o uso de variáveis que indicassem a presença ou ausência dos respectivos níveis, conhecidas como "variáveis dummy". Para um SNP com os três níveis possíveis, por exemplo, uma categoria (AA) é incorporada ao intercepto (não havendo um parâmetro isolado para seu efeito) e cria-se dois parâmetros para as demais (AB e BB). Logo, a representação dicotômica dos parâmetros se daria de forma:

Tabela 1 – Representação dicotômica dos parâmetros do modelo.

	x_1	x_2
AA	0	0
AB	1	0
BB	0	1

.

3.2.1 Modelos univariados

Modelos univariados foram montados, buscando a relação entre cada SNP e o fenótipo em questão, considerando seus níveis alélicos. A esperança de Y_i será dada por:

$$E[Y_i] = \theta_i = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

$Y_i = 1$ para pelagem vermelha e 0 para pelagem não vermelha,

$\beta_0 =$ Intercepto,

$\sum_{j=1}^{2} \beta_j x_j =$ Efeito do j-ésimo nível alélico.

3.2.2 Seleção por torneios

Para a abordagem via seleção por torneios (como descrito na seção 2.2.1), foram montados modelos multivariados. Em um processo de modelagem, quando observações apresentam valor faltante, elas são excluídas da matriz de delineamento. Como a seleção por torneios é um processo automatizado e formado por grupos aleatórios, eventuais combinações de grupos com presença de valores faltantes acarretariam no cancelamento de linhas da matriz de delineamento, de maneira que algumas covariáveis ficariam com apenas um nível, sem variabilidade, impossibilitando a realização de contrastes. Para tanto, foram excluídas todos as covariáveis que apresentaram algum valor faltante, restando 1.522 SNPs.

O modelo multivariado segue de forma similar ao univariado, porém, representado em notações matriciais. A esperança de Y será dada por:

$$E[Y] = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

$Y_{61\times1}$ = Vetor correspondente à cor da pelagem,

$X_{61\times k}$ = Matriz de delineamento (primeira coluna preenchida por 1's, referente ao intercepto, e as demais preenchidas por variáveis dummy, referentes aos $k-1$ parâmetros estimados para os demais níveis),

$\beta_{k\times1}$ = vetor de parâmetros estimados.

3.2.3 Modelos multivariados montados a partir de indicativas de análises univariadas

Para esta abordagem, os p-valores obtidos através dos modelos univariados foram analisados graficamente, com o intuito de identificar regiões com vários SNPs promissores. Estas regiões, caracterizam-se quando vários SNPs significativos (no modelos univariados) encontram-se próximos no decorrer do genoma, causando uma concentração de SNPs significativos em uma certa área genômica. Com isso, para os cromossomos que obtiveram indício de tais regiões, foram montados modelos multivariados (como descrito anteriormente) considerando os SNPs mais significativos. Devido à possível correlação entre SNPs próximos, os modelos passaram pelo mesmo processo de seleção de variáveis utilizada na seleção por

torneios, de maneira a restar apenas SNPs com contribuições marginais significativas.

3.3 Método de estimação

Os parâmetros estimados pelos métodos descritos foram obtidos pelo critério de máxima verossimilhança (vide seção 2.4). O trabalho foi realizado no ambiente R de computação estatística (R Core Team, 2014), de modo que os modelos foram construídos utilizando a função *glm* e os p-valores calculados através do uso da função *Anova*, presente no pacote *car*.

4 RESULTADOS E DISCUSSÃO

4.1 Análise univariada

Os resultados de modelos logísticos univariados possibilitam uma análise visual, ao relacionar o p-valor obtido para cada SNP com sua região cromossômica. No gráfico de *Manhattan*, o eixo horizontal consiste na região cromossômica, de modo que a alternância das cores indicam diferentes cromossomos e pontos mais próximos estão fisicamente mais próximos no genoma. O eixo vertical corresponde ao negativo do $\log_{10}$ do p-valor obtido pela análise univariada ($-\log_{10}$(p-valor)), onde menores p-valores correspondem a maiores valores na escala referida. Também constam no gráfico, linhas sugestivas de significância em vermelho (p-valor = 1e-07) e azul (p-valor = 1e-04). Abaixo, segue o gráfico de Manhattan para os 52.622 modelos univariados.

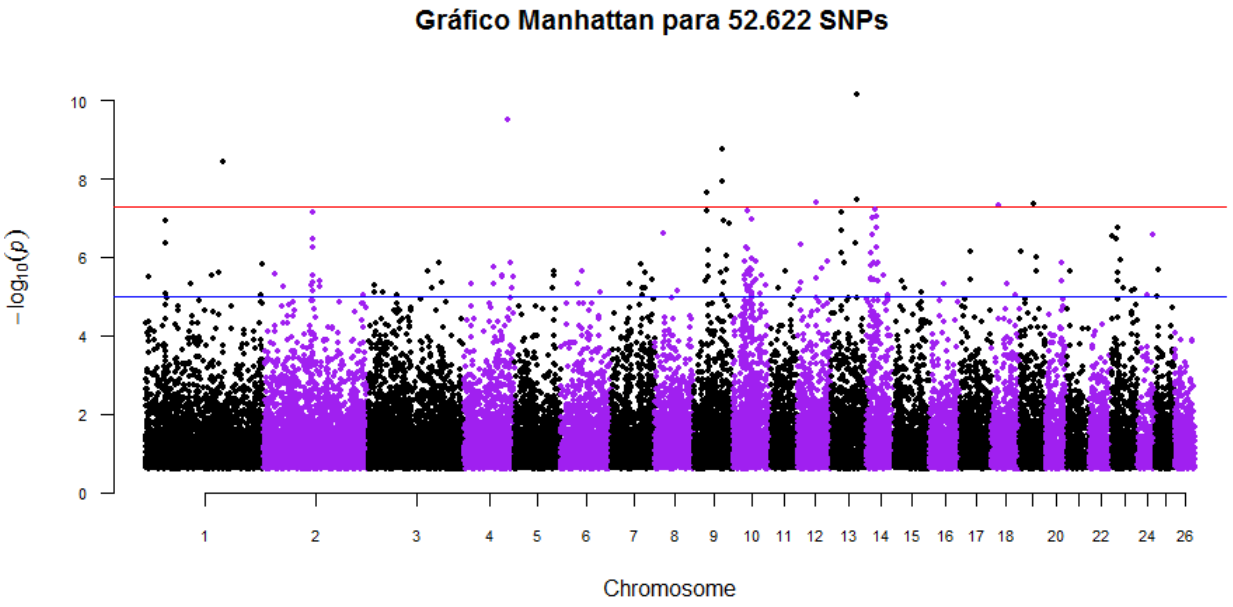

Figura 6 – Resultado de modelos logísticos univariados

Alguns estudos da área, selecionam como candidatos cerca de 20 a 30 SNPs. Utilizando este critério, selecionando os 30 SNPs com menor p-valor, obtém-se candidatos nos cromossomos 1, 2, 4, 9, 10, 12, 13, 14, 18 e 19.

4.2 Seleção por torneios

Ao retirar SNPs com presença de valores faltantes, sobraram 1.522 covariáveis. Abaixo, segue o gráfico de Manhattan para os respectivos 1.522 SNPs, a título de visualização.

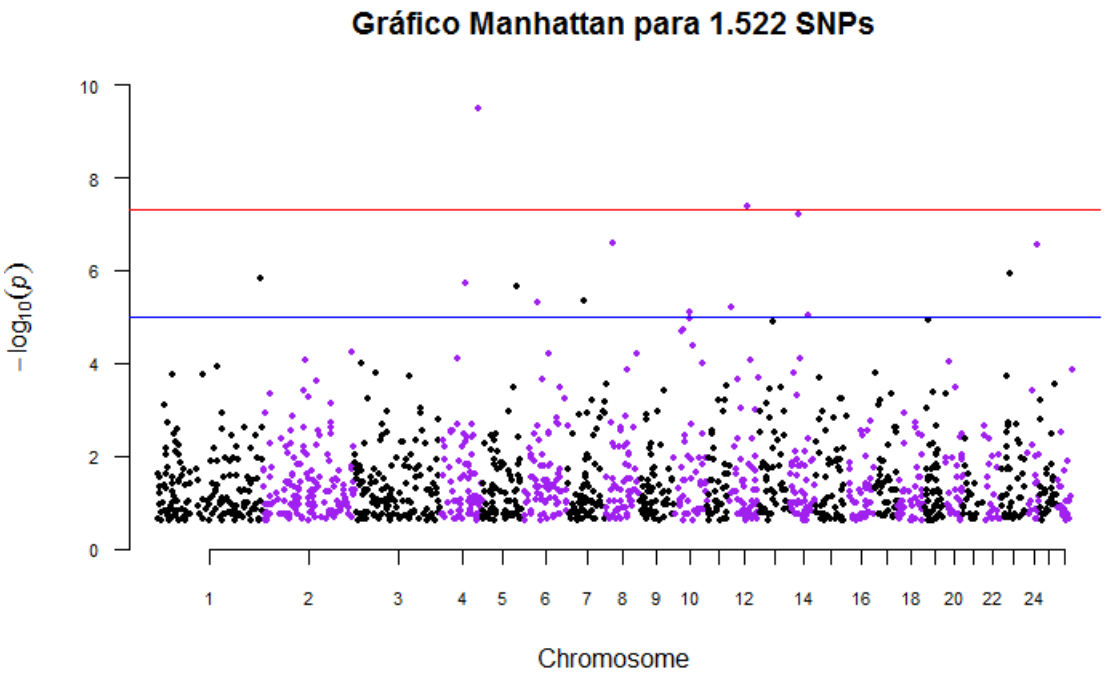

Figura 7 – Resultado de modelos logísticos univariados para SNPs que participaram da seleção por torneios

Nota-se portanto, a diferença da densidade de pontos em relação à Figura 6, levantando a questão da possível exclusão de SNPs influentes.

Esses 1.522 SNPs foram divididos aleatoriamente em 76 grupos, sendo 2 grupos com 21 elementos e 74 grupos com 20 elementos, totalizando os 1.522 SNPs. O algoritmo de torneio foi aplicado, de modo a restar apenas um elemento em cada grupo, ou seja, 76 elementos. Estes 76 foram novamente sorteados para uma segunda rodada, porém em 10 grupos; 2 com 6 elementos e 74 com 8 elementos. Novamente, torneios foram realizados restando apenas 1 por grupo, totalizando então, 10 SNPs "campeões". O procedimento foi repetido 10 vezes mudando a semente de aleatorização, gerando uma lista final com 100 elementos. Ao realizar uma análise de frequência nesta lista, foram selecionados elementos

com frequência maior ou igual a 4, ou seja, selecionados em ao menos 4 dos 10 torneios realizados, o que resultou em 8 SNPs.

Ao construir um modelo com os 8 SNPs "campeões", alguns pareciam não contribuir marginalmente no grupo devido a possíveis correlações. O critério de efeito marginal foi utilizado, assim como na seleção por torneios, de modo a manter apenas SNPs com contribuições significativas, restando então apenas três, formando um modelo final.

Observa-se por meio de tabelas de contingência, que estes três possuem o caso de separação perfeita em algumas categorias, indicando um bom poder de identificação do método, dentre os 1.522 possíveis. Abaixo, segue a tabela de contingência dos três SNPs finais relacionando nível alélico com cor da pelagem, onde 1 representa pelagem vermelha e 0 representa pelagem não vermelha.

Tabela 2 – Frequência Alélica dos SNPs: (a) OAR7_45426116.1, (b) s24425.1 e (c) s50152.1

(a)

Alelo	0	1
AA	9	18
AB	18	4
BB	12	0

(b)

Alelo	0	1
AA	26	0
AB	13	13
BB	0	9

(c)

Alelo	0	1
AA	0	13
AB	16	5
BB	23	4

4.3 Modelos multivariados montados a partir de indicativas de análises univariadas

A análise visual proporcionada pelo gráfico de *Manhattan* permite a identificação de possíveis pontos influentes que estejam em regiões próximas no decorrer do genoma. Por razões genéticas, é coerente esperar que SNPs próximos atuem em conjunto no controle de uma certa característica fenotípica (também pode ocorrer o mesmo para SNPs distantes, apesar de mais raro).

Buscou-se então a identificação de regiões com vários pontos promissores, formando grupos de covariáveis para modelos multivariados. Os cromossomos 2, 9, 10, 13, 14 e 23 obtiveram indícios de serem regiões promissoras, conforme Figura 8. De tal maneira,

os SNPs mais promissores destes cromossomos foram identificados, formando um modelo logístico multivariado para cada cromossomo. SNPs muito próximos podem apresentar alta correlação. Em tal cenário, foi utilizado o que apresentasse menor p-valor univariado (com o maior pico no gráfico) para o modelo múltiplo. Os modelos passaram pelo mesmo processo descrito anteriormente, reduzindo o número de covariáveis. Cada cromossomo, ficou então com a seguinte quantidade de SNPs em seu modelo final:

Tabela 3 – Número de covariáveis no modelo final, por cromossomo.

Cromossomo	Número de SNPs
2	4
9	4
10	5
13	3
14	6
23	4

Na Figura 8, destacam-se em verde e vermelho, nos cromossomos 2, 9, 10, 13, 14 e 23, SNPs próximos e significativos, indicando regiões promissoras.

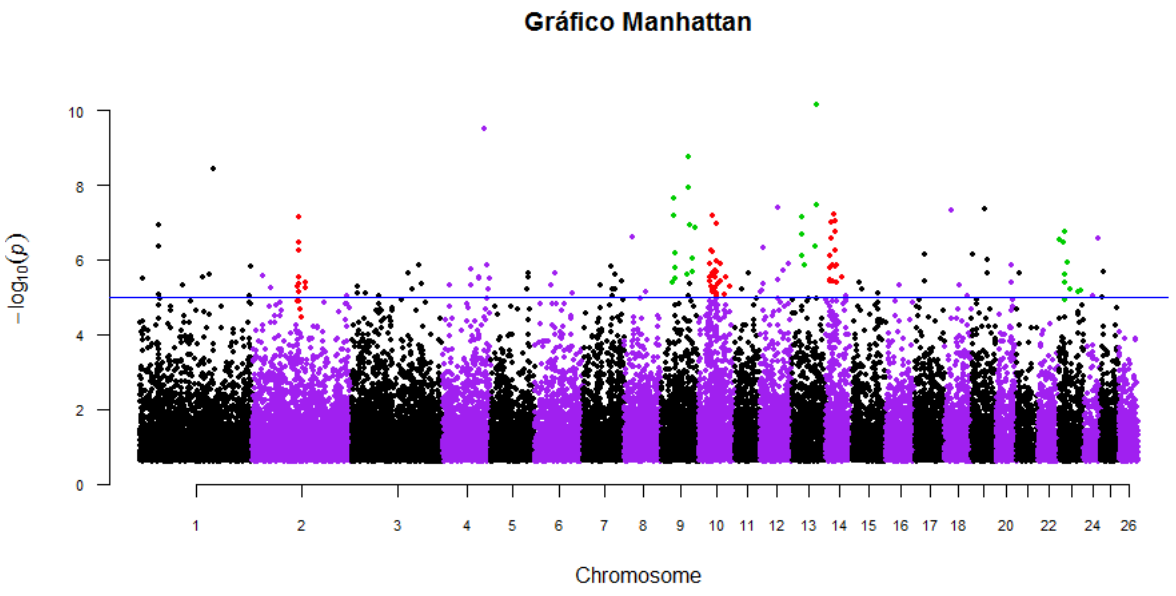

Figura 8 – Em verde, SNPs selecionados para os modelos finais.

4.4 Comparação de resultados

Após estabelecidos os modelos por cromossomos e pela seleção por torneios, os mesmos foram comparados através do *AIC* e do poder de predição, conforme Tabela 4. Para o cálculo do poder de predição de cada modelo, a base de dados foi divida em dois; uma base treino com 45 observações e uma base teste com 16 observações. Os parâmetros foram estimados a partir da base treino, para então, serem utilizados na base teste, calculando o valor predito. Após os valores preditos serem calculados, os mesmos foram comparados com os valores reais, obtendo-se uma taxa de acerto de predição.

Tabela 4 – Comparação dos modelos pelo AIC e poder de predição.

Modelo	AIC	Predição correta
Crm. 2	44, 9	73%
Crm. 9	18, 0	100%
Crm. 10	28, 9	90%
Crm. 13	20, 28	100%
Crm. 14	26, 0	43%
Crm. 23	32, 4	92%
Torneios	25, 6	100%

Ao analisar os modelos entre si, destacam-se os que são formados por SNPs pertencentes aos cromossomos 9 e 13, pelo baixo AIC e poder de predição de 100%. Todavia, o modelo construído através da seleção por torneios também apresenta uma taxa de acerto de predição de 100%, demonstrando a capacidade de construção de um bom modelo de predição, dentre milhares covariáveis iniciais.

Ao observar a frequência dos SNPs selecionados pelos cromossomos 9 e 13, constam-se casos de separação perfeita, fenômeno coerente ao se tratar de assuntos genéticos.

Tabela 5 – SNPs do cromossomo 9; (a) OAR9_7361986.1, (b) s15620.1, (c) OAR9_34630051.1 e (d) OAR9_29671303.1

(a)

Alelo	0	1
AA	6	21
AB	21	1
BB	9	0

(b)

Alelo	0	1
AA	17	2
AB	17	2
BB	3	18

(c)

Alelo	0	1
AA	24	3
AB	11	7
BB	0	12

(d)

Alelo	0	1
AA	1	0
AB	0	10
BB	36	11

Tabela 6 – SNPs do cromossomo 13; (a) OAR13_64325996.1, (b) OAR13_24096002.1, (c) s15466.1

(a)

Alelo	0	1
AA	36	5
AB	2	15
BB	1	2

(b)

Alelo	02	1
AA	18	2
AB	20	8
BB	0	12

(c)

Alelo	0	1
AA	25	3
AB	10	8
BB	0	10

As indicativas encontradas nos cromossomos 9 e 13 podem ser levadas adiantes paraa estudos mais aprofundados e discutidas com geneticistas da área. Caso a literatura genômica referente a ovinos chegue a conclusões ou indícios de que os SNPs responsáveis pela pelagem encontram-se em diferentes cromossomos, a seleção por torneios aplicada em uma base mais consistente, pode ser uma boa abordagem.

5 CONSIDERAÇÕES FINAIS

Este trabalho buscou associações entre genótipo e fenótipo através de ferramentas estatísticas e computacionais viáveis e bem difundidas, a fim de contornar obstáculos encontrados. Foram utilizadas três abordagens, resultando em diferentes números de SNPs, conforme Tabela 7.

Tabela 7 – Número de covariáveis e cromossomos pertencentes no modelo final, por abordagem.

	Univariado	Multivariados	Torneios
Cromossomos	1, 2, 4, 9, 10, 12, 13, 14, 18 e 19	2, 9, 10, 13, 14 e 23	1, 4, 5, 7, 8, 12 e 23
Total de SNPs	30	26	8

Pode-se observar que há diferença na quantidade total de SNPs candidatos selecionados por cada uma das três abordagens. Os modelos univariados selecionaram 30 SNPs, ao passo que a seleção por torneios permaneceu apenas com 8 e os modelos multavariados por cromossomo, apresentaram um total de 26 SNPs candidatos.

A seleção por torneios demonstrou ser uma técnica interessante para redução de dimensionalidade, pois conseguiu selecionar boas variáveis dentre as milhares iniciais, montando um modelo com bom poder de predição. Seria interessante então, alguma alternativa em que fosse viável o uso da seleção por torneios sem que ocorra o descarte das variáveis com valor faltante, pois o resultado poderia ser ainda melhor se todas as variáveis fossem consideradas. O uso de outras técnicas, estatísticas e computacionais, também é uma alternativa atrativa, visto que não há uma maneira consolidada para abordar o problema.

Em meio a tantos desafios encontrados, o uso de modelos logísticos formados por indicativas de regiões promissoras pode ser uma boa primeira aproximação. O esforço computacional torna-se mínimo e a interpretação visual de resultados de análises univariadas, permite que esta primeira aproximação resulte em contribuições, ao apontar possíveis caminhos a se seguir.

Estudos do tipo, que possuem uma contribuição imensurável para a evolução, necessitam do apoio de profissionais de áreas quantitativas e computacionais. Propõe-se então, o contínuo estudo dos desafios encontrados neste trabalho, para uma maior consolidação de ferramentas utilizadas e avanço em descobertas da área.

REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: INTERNACIONAL SYMPOSIUM OF INFORMATION THEORY, 2., Budapest, 1973. Budapest: Akademiai Kiàdio, p. 267-281, 1973.

ALVES, R. R. Seleção por torneios nas estimativas de associação entre marcadores SNP's e fenótipos, Tese de Doutorado, LAVRAS, 2014.

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. The Annals of Statistics v. 36, Number 1, 199-227.

BRITO, et al. (2014) A Biotecnologia no Melhoramento Genético Animal, geneticamentemelhorado.blogspot.com.br, Brasil, 2014.

CAI, T. and LV, J. (2007). Discussion: The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics 35, 2365-2369

Centro de Estudos Avançados em Economia Aplicada (CEPEA) - ESALQ/USP. Universidade de São Paulo, Piracicaba, cepea.esalq.usp.br

CHEN, Z.; CHEN, J. Tournament screening cum EBIC for feature selection with high-dimensional feature spaces. Science in China Series A: Mathematics, Beijing, v. 52, n. 6, p. 1327-1341, June 2009.

COLLET, D. Modelling Binary Data, 2nd edition, 1991.

CORNELIS, C.C.; AGRAWAL, A.; COLE, J.W.; et al.; The Gene, Environment Association Studies consortium (GENEVA): maximazing the knowledge obtained from GWAS by collaboration across studies of multiple conditions, v.34, 2010.

FAN, J. and LV, J. (2008). Rejoinder: Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society Series B 70, 905-908.

HEATON, M. P.; HARHAY G. P.; BENETT, G. L.; STONE, R. T.; GROSSE, W. M.; CASAS, E.; KEELE, J. W.; SMITH, T. P. L.; CHITKO-MCKOWN, C. G.; LAEGREID, W. M. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mammalian Genome, v.13, 2002.

LEITE, E. E. O agronegócio das peles caprina e ovina. EMBRAPA, CNPGC - Gado de Corte. Campo Grande, MS, 2001.

LI, M.; LI, C.; GUAN, W.; Evaluation of coverage variation of SNP chips for genome-wide association studies, European Journal of Human Genetics, v. 16, 2008.

LÔBO, R.N.B et al. Breeding plan for commercial dairy goat production systems in southern Brazil, 2011.

LUNA, A. La Busca de fatores genéticos associados à resposta ao tratamento do HCV Genótipo tipo 3. Tese para obtenção do título de Doutorado, ICB, Universidade de São

Paulo - USP, São Paulo, SP, 2012.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society Series A (General), Vol. 135, No. 3 (1972), pp. 370-384.

OLIVEIRA, L. S. Seleção de covariáveis para ajuste de regressão logística na análise da abundância de invertebrados edáficos em diferentes agroecossistemas. Dissertação de Mestrado, UFV, MG, Brasil, 2011.

R: A Language and Environment for Statistical Computing. R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2014, http://www.R-project.org

REDE MORADA NOVA, EMBRAPA Caprinos e Ovinos; Sobral/CE, 2015.

RESENDE, et. al.; Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine, v. 190(4); 2012 Apr.

Revista Brasileira de Zootecnia, Viçosa, Brasil, 2009.

SILVA, J. P. da; Uma abordagem Bayesiana para o mapeamento de QTLs utilizando método MCMC com saltos reversíveis. In: Escola Superior de Agricultura Luiz de Queiroz, USP, Piracicaba-SP, 2006. Piraciba: Joseane Padilha da Silva, p. 10-23, 2006.

SILVA, P.H.T. da; FACO, O.; SILVA, K. de M.; LANDIM, A. V. Padrão racial e seu impacto sobre o melhoramento genético da Raça Morada Nova. CNPTIA, EMBRAPA, Campinas, SP, 2013.

SOUZA, E. C. Análise de influência local no modelo de regressão logística. Dissertação de Mestrado, ESALQ, USP, Brasil, 2006.

T, JEHAN.; S, LAKHANPAUL.; Single Nucleotide Polimorphism (SNP) - Methods and applications in plantic genetics: A review, Indian Journal of Biotechnology, 2006.

The Bovine Genome Sequencing and Analysis Consortium Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution

TURNER, S.; ARMSTRONG, L.L.; BRADFORD, Y. et al.Quality control procedures for Genome Wide Association Studies, 2011.

VISSCHER, PM.; HILL, WG.; WRAY, NR.; Heritability in the genomics era–concepts and misconceptions. Nature Review Genetics, 2008

WHITTAKER, JC.; Thompson, R.; Denham, MC.; 2000. Marker assisted selection using ridge regression. Genetics Research 75, 249?252

YEAGER M et. al (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics 39, 645-649.

ANEXOS

Anexo A: Parte da programação utilizada no ambiente R

Programação em R utilizada na seleção por torneios

```
ovelhas <- read.delim("C:/Users/Cayan Portela/Desktop/Monografia/reg.csv",sep=",")

retira_na <- function(x){
  sum(is.na(x))
}

quem_tem_na <- apply(ovelhas,2,retira_na)
dados_limpos <- ovelhas[,-which(quem_tem_na>0)]

grupo <- list()
grupo[[1]] <- sample(2:1567,18)

for (i in 2:87){
  grupo[[i]] <- sample(setdiff(2:1567,unlist(grupo)),18)
  }


candidatos <- grupo[[1]]
modelos <- list()
resultado <- list()

for (i in 1:87){

candidatos <- grupo[[i]]

  while(length(candidatos)>0){
  modelos[[i]] <- glm(Fenotipo2 ~ .,family=binomial,
     data=dados_limpos[,c(candidatos,1568)],maxit=100)
  resultado[[i]] <- Anova(modelos[[i]])
  pior <- which.max(resultado[[i]][,3])
  candidatos <- candidatos[-pior]
  }
}

campeoes <- unlist(lapply(resultado,nomes_colunas))
dados_limpos2 <- dados_limpos[,c("ID",campeoes,"Fenotipo2")]

grupo2 <- list()
grupo2[[1]] <- sample(2:89,8)


for (i in 2:10){
```

```
  grupo2[[i]] <- sample(setdiff(2:89,unlist(grupo2)),8)
}


candidatos2 <- grupo2[[1]]
modelos2 <- list()
resultado2 <- list()

for (i in 1:10){

  candidatos2 <- grupo2[[i]]

  while(length(candidatos2)>0){
    modelos2[[i]] <- glm(Fenotipo2 ~ .,family=binomial,
          data=dados_limpos2[,c(candidatos2,90)],maxit=100)
    resultado2[[i]] <- Anova(modelos2[[i]])
    pior2 <- which.max(resultado2[[i]][,3])
    candidatos2 <- candidatos2[-pior2]
  }
}



campeoes2 <- unlist(lapply(resultado2,nomes_colunas))
dados_limpos3 <- dados_limpos[,c("ID",campeoes_maximo,"Fenotipo2")]

treino <- sample(1:61,45)
teste <- setdiff(1:61,treino)



mod_treino <- glm(Fenotipo2 ~.,family=binomial,
  data=final[treino,c(candidatosmax,"Fenotipo2")],maxit=100)

matriz_teste <- model.matrix(
glm(Fenotipo2~.,family=binomial,
data=final[teste,c(candidatosmax,"Fenotipo2")],maxit=100)
)


ab <- mod_treino$coefficients

betas <- matriz_teste%*%ab

preditos <- exp(betas)/(1+exp(betas))
```