2

Mineração de dados para identificar atributos genéticos associados à características de interesse econômico à pecuária

André Gonzaga
Maurício de A. Mudadu
Roberto H. Higa
Eduardo Hruschka

PESQUISADORES DA área de melhoramento genético possuem cada vez mais acesso a dados genéticos e genômicos e demandam por um método ou ferramenta robusta que atendam às suas necessidades na descoberta de conhecimento. Esse trabalho investiga algoritmos e métodos de mineração de dados aplicados ao contexto de melhoramento genético bovino, concentrando-se fundamentalmente na aplicação e no aperfeiçoamento de algoritmos para seleção de atributos, buscando identificar os atributos genéticos associados às características fenotípicas de um indivíduo em bases de dados simuladas. Para tanto, foram utilizados algoritmos existentes, baseados em agrupamento de dados e em teoria de informação, bem como foi desenvolvido um novo método para seleção de atributos e que é baseado no conceito de janela

deslizante. Os resultados gerados são compatíveis com o domínio do problema e equivalentes aos obtidos por algoritmos tradicionais da área de bioinformática, os quais são usualmente mais caros computacionalmente.

2.1 Introdução

O Brasil possui o maior rebanho bovino comercial do mundo, com mais de 200 milhões de cabeças. No ano de 2013 abateu por volta de 43 milhões de cabeças e exportou 2 milhões de toneladas de equivalente carcaça (ABIEC, 2014). A identificação dos atributos genéticos que determinam características como maciez da carne e porcentagem de gordura no leite podem ter valor econômico direto nesse mercado.

Uma das maneiras tradicionais de se realizar melhoramento de bovinos é por seleção de animais, direcionando os cruzamentos com o uso de informações de desempenho, *pedigree* e cálculo do valor ou mérito genético de um animal (FALCONER; MACKAY, 1996). Alternativamente é possível o uso de atributos genômicos (marcadores SNP, do inglês *single nucleotide polymorphism*, ou polimorfismo de um único nucleotídeo) para assistir na seleção de animais e também para realizar seleção e predição genômica (MEUWISSEN; HAYES; GODDARD, 2001).

O genoma bovino é composto por 31 cromossomos com milhares de genes, que são responsáveis pela expressão de características individuais (LIU et al., 2009), as quais podem variar desde a tendência em desenvolver uma determinada doença, até a capacidade de crescimento do animal. Em termos práticos, torna-se interessante identificar quais regiões do genoma bovino estão relacionadas às características de interesse econômico (essas regiões também são denominadas QTL – Quantitative Trait Loci) para a produção de bovinos de corte e de leite, a fim de tornar possível estimar

o mérito genético-genômico de um determinado indivíduo (HAYES et al., 2009).

Avanços na tecnologia de sequenciamento de DNA e genotipagem levaram ao desenvolvimento de *chip* de DNA de alta densidade que podem ser usados para caracterizar centenas de milhares de marcadores SNP em um único ensaio. A busca e identificação da importância de um dado SNP para uma dada característica de desempenho de um animal (ou os fenótipos, usando terminologia biológica) é feita por meio de testes de associação. Essa busca por marcadores relevantes pode ser ser realizada em todo o genoma via estudos amplos de associação (GWAS – Genome-wide Association Studies) (BALDING, 2006).

Existem diversas metodologias para realizar GWAS. As mais comuns são métodos paramétricos de regressão linear e logística, nos quais são usadas funções para regredir os fenótipos (características de desempenho dos animais) aos atributos genéticos (genótipos ou conjunto de marcadores SNP, usados como covariáveis). Essas funções podem ser interpretadas como uma aproximação para os valores genéticos verdadeiros e desconhecidos e são geralmente modelos matemáticos que envolvem os genótipos, as interações entre os genótipos e condições ambientais (CAMPOS et al., 2013). É importante lembrar que métodos paramétricos usualmente assumem normalidade, linearidade e ausência de colinearidade, o que nem sempre é a regra nesse contexto.

Metodologias não paramétricas são uma alternativa aos métodos tradicionais, principalmente na tentativa de se modelar interações não-lineares que não são capturadas pelos métodos paramétricos. Interações não-lineares são consequência de uma relação mais complexa entre o fenótipo e os genótipos, como, por exemplo, heterogeneidade de *locus* (diferentes trechos do genoma levando ao mesmo fenótipo), fenocópia (fenótipos determinados

exclusivamente pelo ambiente e sem base genética) e epistasia (interações entre genes) (MOORE; ASSELBERGS; WILLIAMS, 2010). Entre as metodologias não-paramétricas utilizadas nesse contexto estão aquelas baseadas em técnicas de mineração de dados e o de aprendizado de máquina, que são métodos computacionais, os quais são essencialmente métodos de modelagem computacional – por exemplo, árvores de classificação e decisão, redes neurais, *support vector machines* (HOWARD; CARRIQUIRY; BEAVIS, 2014).

O presente estudo reporta o uso de GWAS em bases de dados simuladas de fenótipos e genótipos voltadas para o contexto da pecuária, de forma a testar e comparar algumas metodologias não-paramétricas com metodologias paramétricas tradicionais.

2.1.1 Seleção genômica e GWAS

O uso da genômica no melhoramento animal vem sendo adotado por criadores e produtores nos últimos anos (ROLF et al., 2014). A maioria das características de importância econômica na pecuária são quantitativas ou poligênicas, ou seja, consequência da ação de muitos genes que contribuem cada um com uma pequena parcela para a variância fenotípica. Em 2001, Meuwissen, Hayes e Goddard (2001) propôs o uso de marcadores SNP no cálculo mais acurado do mérito genético de animais, em um método denominado seleção genômica. Essa metodologia é indicada principalmente para características em que a seleção genética tradicional é pouco eficiente, como no caso de características complexas e de baixa herdabilidade, assim como características obtidas tardiamente como dados de carcaça e eficiência alimentar.

O termo GWAS foi cunhado devido à distribuição quase homogênea desses marcadores pelo genoma de forma a cobrir grande parte de sua extensão. Dessa forma esses marcadores podem servir como ferramenta para capturar de forma eficiente

o efeito poligênico de características quantitativas, em testes de associação. GWAS requer três elementos: (i) muitas amostras, ou animais, se possível de diversas populações; (ii) marcadores genéticos que cubram grande parte do genoma e (iii) métodos analíticos poderosos o suficiente para identificar sem viés a associação entre marcadores e os fenótipos (CANTOR; LANGE; SINSHEIMER, 2010).

2.1.2 Atributos genéticos: marcadores SNP

Os atributos genéticos submetidos à mineração de dados nesse projeto são os marcadores genéticos do tipo SNP. Esses marcadores são um tipo de polimorfismo em que há alteração em um único nucleotídeo do DNA: A, T, C ou G, cujas letras são referentes às bases nitrogenadas adenina, guanina, citosina ou timina.

Em outras palavras, SNP é uma variação pontual na sequência de DNA e que mantém frequência mensurável em uma população. Os SNPs são os mais abundantes de todos os marcadores genéticos existentes, pois são distribuídos uniformemente por todo o genoma do indivíduo. Estima-se que existam mais de 10 milhões de marcadores SNP no genoma humano e que pelo menos 300 mil deles implicam variações genéticas significantes (SHAH; KUSIAK, 2004).

Graças à interação gênica, SNPs podem servir de marcadores para determinar regiões do genoma que estejam associadas à características fenotípicas de interesse econômico como: maciez da carne, tamanho da área de olho de lombo, produção de leite etc.. Essa interação gênica, também denominada desequilíbrio de ligação (DL), indica que marcadores SNP podem não ter efeito direto em um fenótipo, mas podem estar associados a um trecho do

genoma responsável por uma característica, servindo de “marca” para essa região.

Outra conclusão que pode ser tirada do DL é que o conjunto de marcadores SNP é redundante, ou seja, marcadores SNP podem estar associados entre si, de forma que também poderão estar associados simultaneamente a um dado fenótipo. Uma metodologia para se eliminar essa redundância é a construção de blocos de SNP que estão em DL. Tais blocos podem então ser usados nos testes de associação, ou então para selecionar marcadores SNP referência dentre vários de uma região, denominados *tag* SNP (BALDING, 2006).

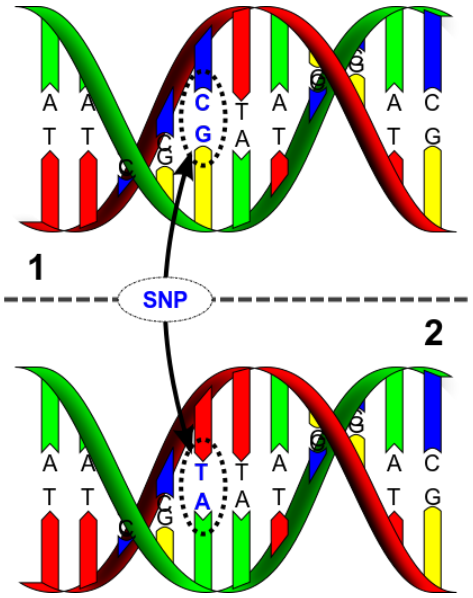


Figura 2.1 – Representação de um polimorfismo de nucleotídeo simples. Fonte: Wikipédia Commons (HALL, 2007).

2.1.3 QTL-MAS: simulação de dados genéticos

A acurácia e poder de métodos de seleção genômica e GWAS podem ser avaliados utilizando dados genômicos simulados (DAETWYLER et al., 2013). As bases de dados utilizadas nesse trabalho são oriundas de dados simulados, disponibilizados

no *workshop* QTL-MAS dos anos 2012 (PANICO et al., 2012) e 2011 (ROY et al., 2011), o qual reúne diversos especialistas em seleção genética de plantas e animais e também fornece dados biológicos artificiais com o conjunto dos parâmetros genéticos utilizados na simulação. Essa base é composta por um conjunto de vários animais (instâncias) com milhares de SNPs associados à uma característica quantitativa ou qualitativa (classe).

A Tabela 2.1 exemplifica as bases de dados utilizadas nesse trabalho. Note que os atributos genéticos são categóricos e podem assumir os valores 11, 12, 21 ou 22, representando as possibilidades de polimorfismo no alelo em questão, considerando que os marcadores SNP são bialélicos. Já o fenótipo é um atributo quantitativo, podendo assumir qualquer valor real ($\mathbb{R}$), representando a quantificação de alguma característica de interesse no indivíduo.

A ordem dos marcadores SNP é de suma importância, pois eles são distribuídos uniformemente pelo genoma simulado, o que é uma aproximação da realidade. Dessa forma, nos dados simulados, a distância entre dois SNP consecutivos permanece constante em todo o cromossomo, ou seja, o SNP 1 é seguido pelo SNP 2 que por sua vez é seguido pelo SNP 3, e assim por diante.

Tabela 2.1 – Representação genérica das bases de dados com N SNPs e M indivíduos.

Indiv.	SNP 1	SNP 2	...	SNP N	Fenótipo
1	{11,12,21,22}	{11,12,21,22}		{11,12,21,22}	$\mathbb{R}$
2	{11,12,21,22}	{11,12,21,22}		{11,12,21,22}	$\mathbb{R}$
...	{11,12,21,22}	{11,12,21,22}		{11,12,21,22}	$\mathbb{R}$
M	{11,12,21,22}	{11,12,21,22}		{11,12,21,22}	$\mathbb{R}$

2.2 Metodologia

Esta seção apresenta, resumidamente, uma visão geral da teoria sobre os métodos tradicionais (paramétricos) e os méto-

dos computacionais (não-paramétricos) que foram utilizados para realizar GWAS com os dados genômicos e fenotípicos simulados do QTL-MAS 2011 e 2012.

2.2.1 Métodos paramétricos

Para se calcular a associação de marcadores genéticos a uma dada característica em um GWAS, pode-se usar um modelo linear paramétrico tradicional em uma regressão:

$$y = \mathbf{X}b + \sum_{i=1}^{m} z_i g_i + e,$$

onde y é um vetor com valores fenotípicos dos animais, b é um vetor de efeitos fixos e X a matriz de incidência relacionando o vetor de efeitos b em y, g_i é o efeito de cada marcador variando de i a m, z_i é o vetor de genótipos de cada indivíduo por marcador i, e o último termo, e, é um vetor de erros residuais aleatórios.

Baseado nesse modelo, diversas outras metodologias foram derivadas como as metodologias de regressão *ridge* BLUP (*best linear unbiased prediction*), rrBLUP (*ridge regression* BLUP) e gBLUP (*genomic* BLUP) (ZHANG; ZHANG; DING, 2011) assim como metodologias bayesianas como o BAYES LASSO, Bayes A, Bayes B, Bayes C e Bayes Cπ (HOWARD; CARRIQUIRY; BEAVIS, 2014).

Tais modelos de regressão testam um SNP por vez, podendo gerar falsos positivos devido ao viés de testes múltiplos, porque ignoram a interação entre os marcadores SNP (MOORE; ASSELBERGS; WILLIAMS, 2010). O uso de blocos de SNP ou de *tag* SNP, ao invés de SNP únicos nos testes de associação pode ser uma alternativa para reduzir esse problema. Para montar os blocos de SNP e calcular os *tag* SNP é necessário calcular o nível de DL entre todos os SNP, par a par. Um dos softwares mais usados para realizar essa tarefa é o Haploview (BARRETT et al., 2005). Antes

de se montar os blocos é necessário reconstruir a “fase de ligação”, que nada mais é do que estimar de qual cromossomo, materno ou paterno, um dado SNP se originou. O Beagle é um software muito utilizado nesse intuito (BROWNING; BROWNING, 2007).

Algumas metodologias paramétricas estimam o efeito de cada SNP no fenótipo (% da variância fenotípica explicada) assim como há metodologias que geram *p-valor*, referentes ao teste de uma hipótese nula de não-associação. Métodos não-paramétricos não fazem uso de tais estatísticas e muitas vezes o parâmetro de medida de associação de um SNP a um fenótipo são scores que medem a importância de um dado SNP para explicar o(s) modelo(s) de classificação. Essa “importância” é geralmente medida como a frequência com que um dado SNP aparece nas repostas dos modelos computacionais usados na classificação.

2.2.1.1 PLINK

O PLINK é uma das ferramentas mais conhecidas e eficientes para se manipular dados de SNP (PURCELL et al., 2007; CLARKE et al., 2011) além de também realizar GWAS. Dentre as várias metodologias disponíveis para realizar estudos de associação com características quantitativas nesse software, encontram-se as regressões lineares. Nesse caso PLINK retorna valores dos coeficientes de regressão (β) e *p-valor* para estatísticas T individuais para cada SNP. Também é possível realizar correções para viés de múltiplos testes, como ajustes de Bonferroni, Sidak, False Discovery Rate (FDR), etc., além de procedimentos de permutação para gerar níveis de significância empiricamente. O PLINK também permite utilizar blocos de SNP para realizar testes de associação.

2.2.1.2 Gensel

O Gensel é um software usado para seleção genômica no qual foram implementados diversos métodos bayesianos, como Bayes A, B, C e Cπ. O software não é aberto ao público mas pode ser acessado por colaboradores da Iowa State University (FERNANDO; GARRICK, 2009). Ao contrário do PLINK que gera *p-valor* para cada SNP, indicando se há associação com o fenótipo, um dos arquivos resposta gerados pelo Gensel indica a porcentagem de variância fenotípica explicada por blocos de SNP, no caso indicando regiões que são possíveis QTL.

2.2.2 Mineração de dados e aprendizado de máquina

A mineração de dados envolve basicamente três etapas: preparação de dados, utilização de um algoritmo para reconhecimento de padrões e análise das informações obtidas (BIGUS, 1996). Estas três etapas são correlacionadas e interdependentes, de tal forma que a abordagem ideal para extrair informações relevantes em bancos de dados consiste em considerar as inter-relações entre cada uma das etapas e sua influência no resultado final. Em linhas gerais, a preparação de dados envolve a seleção de atributos e de objetos (registros) adequados para a mineração, a representação e o pré-processamento dos dados para as ferramentas de mineração, e a purificação dos dados (eliminação de ruído).

Usualmente, o principal objetivo no processo de seleção de atributos é melhorar o desempenho dos algoritmos de modelagem (WITTEN; FRANK, 2005), reduzindo o tamanho da base de dados por meio da remoção de atributos irrelevantes e redundantes. Resumidamente, existem três abordagens fundamentais para selecionar atributos: *filter*, *wrapper* e *embedded* (GUYON; ELISSEEFF, 2003). Filtros baseiam-se apenas nas propriedades intrínsecas dos dados e são computacionalmente mais baratos. Métodos *wrapper*

dependem de um algoritmo de aprendizado para buscar subconjuntos de atributos que sejam representativos da base de dados como um todo, sendo computacionalmente custosos. Já nos métodos do tipo *embedded* a seleção de atributos ocorre como parte integrante do algoritmo indutor do classificador. Ademais, existem também métodos híbridos, que combinam características dos três métodos mencionados.

2.2.2.1 Seleção de atributos via agrupamento de dados

O processo de agrupamento de dados envolve a identificação de um conjunto de categorias – também chamadas de grupos ou de clusters – que descrevam um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), objetivando-se maximizar a homogeneidade entre os objetos de um mesmo grupo e, concomitantemente, maximizar a heterogeneidade entre objetos de grupos distintos.

Existem várias técnicas utilizadas para agrupamento de dados (*clustering*). No contexto desse projeto um dos algoritmos investigados é o filtro baseado em silhueta (SSF) (COVOES; HRUSCHKA, 2011), que se baseia na identificação de grupos de atributos correlacionados. Este algoritmo é bem fundamentado teoricamente, e pode ser visto como um aperfeiçoamento de dois algoritmos bem conhecidos – ACA (AU et al., 2005) e MMP (MITRA; MURTHY; PAL, 2002). Basicamente, o algoritmo agrupa atributos que sejam correlacionados entre si e depois seleciona apenas o atributo que seja o mais representativo de cada grupo.

Mais especificamente, o SSF supre algumas limitações encontradas no ACA e no MMP, bem como permite incorporar a informação da classe ao processo de agrupamento de atributos. Esta característica é de fundamental importância nessa pesquisa, pois

permite identificar grupos de atributos que sejam simultaneamente correlacionados entre si e com o atributo meta (classe).

2.2.2.2 Algoritmo baseado em ganho de informação: InfoGain

A partir da teoria da informação (COVER; THOMAS, 2006), o algoritmo denominado de InfoGain (WITTEN; FRANK, 2005) utiliza as medidas de entropia e informação mútua para selecionar os atributos mais relevantes em relação à classe principal de uma base de dados.

A entropia mede o grau de incerteza em relação aos valores que uma variável aleatória discreta pode assumir. A Expressão 2.1 representa o cálculo da entropia H para uma variável X, sendo x os possíveis valores da variável e p_x a probabilidade da variável X assumir um valor x na base de dados analisada. A entropia pertence ao intervalo $[0, 1]$ no qual 0 (zero) significa ter todas as instâncias com o mesmo valor para essa variável, ou seja, não há incerteza em relação ao seu valor, e 1 significa ter esses valores distribuídos uniformemente entre as instâncias da base de dados, consistindo o grau máximo de incerteza.

$$H(X) = -\sum_{x \epsilon X} p_x \, log_2 \, p_x. \tag{2.1}$$

A informação mútua representada pela Expressão 2.2 utiliza a medida de entropia condicional para avaliar o quanto uma variável contribui para determinar o valor de uma outra variável.

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \, log \frac{p(x,y)}{p(x) \, p(y)} \tag{2.2}$$

O algoritmo InfoGain utiliza diretamente a medida de informação mútua para selecionar os atributos que estejam correlacionados com a informação da classe. Dessa forma, a medida de informação mútua é calculada pela Expressão 2.3 e a seleção dos

atributos é realizada por meio do ranqueamento dessa medida para cada um dos atributos avaliados.

$$\mathsf{InfoGain}(\mathsf{Classe}, \mathsf{Atributo}) = \mathsf{H}(\mathsf{Classe}) - \mathsf{H}(\mathsf{Classe}|\mathsf{Atributo}) \quad (2.3)$$

2.2.2.3 Algoritmo baseado em janela deslizante

O algoritmo baseado no conceito de janela deslizante foi desenvolvido particularmente para o contexto dessa pesquisa com o objetivo de avaliar o mérito de atributos sequenciais, como é o caso dos SNP.

O método consiste basicamente em criar um subconjunto de atributos consecutivos, representando uma região de interesse no cromossomo. Esse subconjunto funciona como uma janela deslizante. A cada iteração do algoritmo o subconjunto é utilizado para classificar os dados e, em seguida, o algoritmo percorre o vetor de atributos adicionando, e removendo atributos sequencialmente, como mostra a Figura 2.2. Seleciona-se, então, o atributo central dos subconjuntos que apresentaram a melhor acurácia de classificação.

$$\underbrace{snp1 \;\; snp2 \;\; snp3 \;\; snp4 \;\; snp5}\;\; \text{snp6} \;\; \text{snp7} \;\; ...$$

$$\text{snp1} \;\; \underbrace{snp2 \;\; snp3 \;\; snp4 \;\; snp5 \;\; snp6}\;\; \text{snp7} \;\; ...$$

$$\text{snp1} \;\; \text{snp2} \;\; \underbrace{snp3 \;\; snp4 \;\; snp5 \;\; snp6 \;\; snp7}\;\; ...$$

Figura 2.2 – Funcionamento do algoritmo de janela deslizante.

2.3 Descrição do conjunto de dados

2.3.1 Análise da base de dados

Após a aquisição das bases de dados, realizou-se uma análise estatística da distribuição dos registros coletados. Na Tabela 2.2 observa-se que a quantidade de padrões genéticos[1] em relação ao número total de atributos é desproporcional, visto que dos 10 mil atributos totais menos de 1% são considerados padrões genéticos que influenciam o fenótipo.

Tabela 2.2 – Características das bases de dados utilizadas.

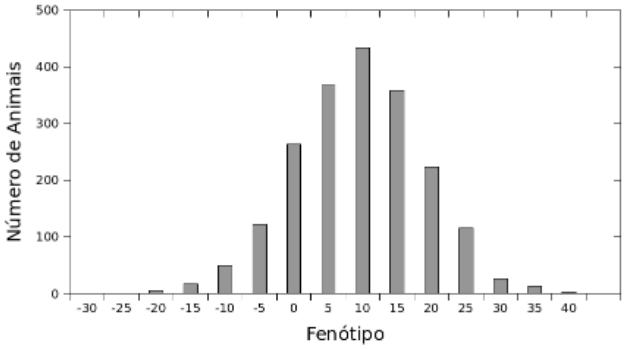
QTL-MAS	Fenótipo	Indivíduos	SNP	Padrões Genéticos
2011	[-30;40]	3 100	9 990	8
2012	[-600;600]	4 100	10 000	50

A principal diferença entre as duas bases de dados é quantidade de padrões genéticos simulados. No conjunto de dados de 2011, apenas oito dos dez mil marcadores genéticos são responsáveis pelo fenótipo dos animais, enquanto na simulação de 2012, cinquenta atributos estão relacionados à manifestação das características. Vale ressaltar que o fenótipo observado nos animais em cada uma das bases de dados é um atributo numérico que compreende diferentes intervalos, como pode ser visto na Tabela 2.2.

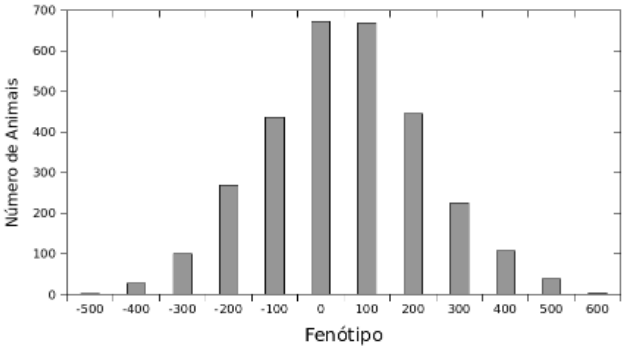
Sabe-se previamente que existem atributos genéticos que contribuem positivamente para uma característica e outros que contribuem negativamente, ou seja, dentre os padrões genéticos gerados na simulação, alguns marcadores são responsáveis por aumentar o valor fenotípico e outros por diminuir. Dessa forma, podem existir animais que possuem apenas os componentes que contribuem positivamente para uma característica, animais que possuem apenas

[1] Padrões genéticos são os atributos que estão correlacionados ao fenótipo (classe).

os componentes negativos, ou mais comumente, os animais que possuem ambos atributos genéticos e acabam por desenvolver um fenótipo médio, como pode ser visto nos histogramas da Figura 2.3.



(a) QTL-MAS 2011



(b) QTL-MAS 2012

Figura 2.3 – Histograma das bases de dados.

Dentre os parâmetros de simulação utilizados na criação dos padrões genéticos, existem alguns atributos que se destacam no quesito “capacidade de influência no fenótipo” e também na frequência de ocorrência na população (Tabela 2.3a). Um atributo genético que esteja presente em 100% da população torna-se impossível de ser identificado, mesmo que tenha um alto poder de influência no fenótipo, pois todos os animais teriam essa variante, tanto os indivíduos geneticamente favorecidos, quanto os desfavorecidos, tornando impossível classificá-los. O mesmo vale para

atributos que possuem uma influência mínima na característica (Tabela 2.3b).

Tabela 2.3 – Exemplos de SNP presentes nas simulações.

(a) Exemplo de SNP importantes

SNP	Efeito	Frequência
6499-6500	+74	47%
2673-2674	-42	65%

(b) Exemplo de SNP irrelevantes

SNP	Efeito	Frequência
3499-3500	+5	100%
2673-2674	-0.73	66%

2.3.2 Descrição do problema

O problema tratado consiste basicamente na seleção de atributos que sejam correlacionados a uma classe principal. Os atributos são componentes genéticos representados pelos SNP e a classe principal é a quantificação de um fenótipo observado no indivíduo, representado por um atributo numérico, como pode ser observado na Tabela 2.1.

A principal complexidade do problema é dada pelo alto número de atributos e o pequeno número de instâncias, caracterizando um caso da denominada “maldição da dimensionalidade” (BELLMAN, 1961), problema no qual os dados se tornam esparsos e pouco representativos. Outra particularidade da base de dados que também contribui para complexidade do problema se refere à grande quantidade de atributos irrelevantes visto que, no contexto biológico, dentre todos os componentes genéticos presentes no genoma de um indivíduo apenas alguns deles estão efetivamente associados ao seu fenótipo.

Devido à natureza do problema existe também a questão de inconsistência dos dados, pois o desenvolvimento de uma característica não é restrita apenas ao genoma do indivíduo, mas também às condições ambientais (GRIFFITHS et al., 2007). Assim sendo, não é raro obter indivíduos que possuem características contraditórias ao seu genótipo, tornando a base de dados suscetível

a ruídos causados por esses casos, como pode ser observado na Figura 2.4, na qual as barras azuis representam o fenótipo apresentado pelo indivíduo e as barras vermelhas representam o fenótipo esperado de acordo com o seu genoma. Pode-se inferir essa informação através dos parâmetros de simulação utilizados no conjunto dos dados do QTL-MAS 2012.

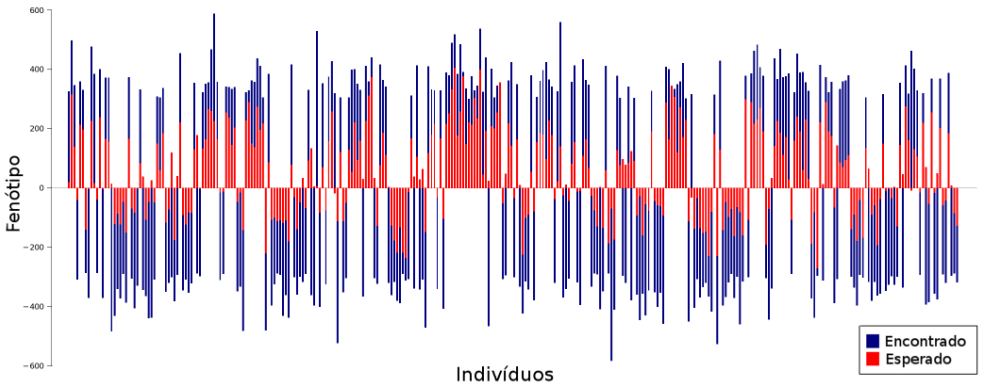


Figura 2.4 – Exemplo de ruídos na base de dados QTL-MAS 2012.

2.3.3 Metodologia dos métodos paramétricos

2.3.3.1 PLINK

PLINK foi utilizado inicialmente para realizar filtros de controle de qualidade nos dados: MAF (frequência do alelo menos frequente, ou do inglês *minimum allele frequency*) 1%, *call rate* para SNP 5% e amostras 10% . Para realizar GWAS em ambas as bases de dados (2011 e 2012), utilizou-se PLINK usando regressão linear. Foram adotadas duas estratégias: (i) testando um SNP de cada vez (*single SNP*) e (ii) testando blocos de SNPs. Em (i) foi usado o comando "*–linear*" e foi realizado um procedimento de permutação adaptativa para se obter *p-valor* com significância empírica, com o comando "*–perm*". Em (ii) foi usado o comando "*–hap-linear*" e 1.000 permutações com o comando "*–mperm 1000*".

Em ambas as estratégias, os resultados foram ordenados pelos *p-valor* obtidos com permutação e os 50 SNP e blocos de SNP com menores *p-valor* foram usados nas análises. Na estratégia (ii) o SNP do centro do bloco foi utilizado nos cálculos de Precision e Recall.

Para se obter os blocos de SNP foi usado o software Beagle para reconstruir a fase de ligação e o software Haploview para construir os blocos com o comando *-blockoutput GAB*.

2.3.3.2 GenSel

O GenSel foi utilizado com o método Bayes A e os parâmetros: chainLength = 41.000; burnin = 1.000; probFixed = 0,95; varGenotypic = 1; varResidual = 1; nuRes = 10; degreesFreedomEffectVar = 4; outputFreq = 100 e windowBV = *yes*, sendo que o último comando habilita a busca por blocos de SNP associados (QTL). Apenas os resultados por blocos foram utilizados nas análises, após ordená-los pela % de variância do fenótipo explicada por cada um, e os 50 blocos mais influentes foram utilizados. A mesma estratégia utilizada com o PLINK foi adotada para cálculo de Precision e Recall.

2.3.4 Metodologia dos algoritmos computacionais

A preparação dos dados limitou-se a obter rótulos de classe, a partir de uma variável associada ao fenótipo e que originalmente é representada por valores contínuos, em um atributo binário. Essa transformação aconteceu por meio de uma seleção dos animais (instâncias) que possuíam valores extremos em relação ao fenótipo em questão, ou seja, separamos os animais em dois grupos. Um grupo é formado por todos os animais cujo fenótipo se destaca pelo lado positivo e o outro grupo é formado por todos os animais que se destacam pelo lado negativo. Desta forma, foram

selecionados apenas os animais cujo mérito genético é muito alto ou muito baixo, eliminando o risco de se ter animais que possuem tanto os atributos que contribuem positivamente quantos os que contribuem negativamente.

A quantidade de animais que constituem os grupos é relativa pois seria necessário estabelecer um limiar para o qual o valor do fenótipo seria considerado extremo positivo ou extremo negativo. Uma alternativa à da escolha de um limiar seria dividir em grupos de mesmo tamanho. Por exemplo, em um caso de teste de 50 animais, metade seriam os indivíduos que desenvolveram os maiores valores para o fenótipo e a outra metade seriam os animais que possuem o fenótipo com os menores valores. No entanto o tamanho do grupo ainda é um valor arbitrário, assim sendo optou-se por criar seis diferentes partições com tamanhos distintos, sendo elas: 10, 20, 40, 80, 160 e 320.

A Figura 2.5 exemplifica o método utilizado para a seleção dos indivíduos candidatos à mineração. A criação de um grupo com 10 animais seria equivalente à selecionar apenas os indivíduos localizados nos extremos do histograma. Conforme o tamanho do grupo aumenta, os animais pertencentes ao grupo tendem a estar localizados mais próximos do centro.

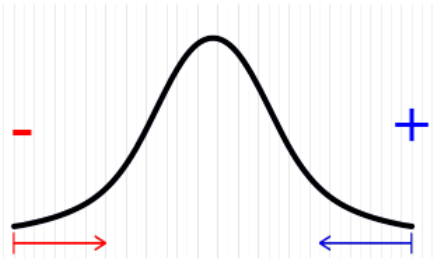


Figura 2.5 – Método para seleção dos indivíduos candidatos à mineração referenciando o histograma de fenótipos da Figura 2.3.

Seleção de atributos

Os experimentos com seleção de atributos foram realizados utilizando três algoritmos distintos que se encaixam no contexto do problema e são computacionalmente eficientes:

1. **Information Gain** (COVER; THOMAS, 2006): Calcula o mérito de um atributo em relação à uma classe principal utilizando a entropia condicional. Os atributos selecionados foram os 50 com maiores méritos.

2. **SSF** (COVOES; HRUSCHKA, 2011): Realiza a redução de dimensionalidade da base de dados agrupando atributos que são semelhantes. Os atributos selecionados são aqueles que melhor representam o grupo ao qual pertencem. A semelhança entre os atributos é determinada por uma medida de distância que pode ou não considerar a classe principal. O número de grupos pode ser fixado a *priori* ou estimado automaticamente a partir dos dados. Neste trabalho, optou-se por fixar o número de grupos em 50 e utilizar uma medida de distância que considera a classe principal.

3. **Janela deslizante:** Para executar os experimentos de seleção de atributos utilizamos uma janela de tamanho igual a 5, que equivale a avaliar o mérito de uma região no cromossomo com cinco SNPs. O classificador utilizado foi o KNN (K-Nearest Neighbour) (WANG; ZUCKER; JEAN-DANIEL, 2000), que consiste em classificar uma instância de acordo com os seus K vizinhos mais próximos, ou seja, aqueles indivíduos que mais se parecem com o objeto em questão, de acordo com seus atributos.

2.3.5 Medidas de qualidade: Precision e Recall

Para avaliar a qualidade dos resultados foram utilizadas as medidas Precision e Recall definidas por (POWERS, 2007), adaptando-as para o contexto já mencionado. Assim sendo, a medida Precision contabilizará como verdadeiro positivo todos os SNP relatados pelo algoritmo e que estejam contidos em uma região de interesse[2]. Já o Recall vai desconsiderar SNP relatados que compreendem uma mesma região de interesse, ou seja, mesmo que o algoritmo retorne quatro SNP corretos em uma mesma região apenas um deles será considerado correto visto que pertecem a uma mesma região.

$$\frac{TP}{TP+FP} \qquad \frac{TP}{TP+FN}$$

(a) Precision (b) Recall

Figura 2.6 – Medidas de qualidade.

2.4 Resultados e discussão

Para avaliar os resultados obtidos nos experimentos primeiramente é preciso entender qual tipo de informação um especialista do domínio está interessado em extrair quando se realiza a mineração de dados genéticos. Conforme já discutido anteriormente, pode-se considerar que que os marcadores SNP estão distribuídos de forma contínua e homogeneamente espaçada por todo o genoma de um indivíduo. Dessa maneira, pode-se considerar que marcadores vizinhos pertencem a uma mesma região no genoma. A informação na qual o especialista do domínio está interessado é a posição no genoma que tem efeito sobre o fenótipo. Essa posição pode ser estimada pela associação do fenótipo a um

[2] Definida por um raio de 5 SNP a partir do SNP causador.

ou mais marcadores SNP. Nesse contexto é válido afirmar que qualquer SNP que esteja em uma região próxima, ou em desequilíbrio de ligação com a região do genoma que tem efeito sobre o fenótipo é tida como resposta válida.

Comparando os resultados dos 5 melhores SNPs (TOP 5 SNP) ranqueados pelos algoritmos não-paramétricos com os resultados dos métodos tradicionais paramétricos (Tabelas 2.4 e 2.5) percebe-se que mesmo sem qualquer adequação para interpretar os dados genéticos e sem qualquer outra informação genética, por exemplo *pedigree*, foram obtidos resultados compatíveis ao contexto e bastante semelhantes aos softwares especializados da área. Essa tendência já foi observada em outros estudos (HOWARD; CARRIQUIRY; BEAVIS, 2014).

Tabela 2.4 – Desempenho dos algoritmos no QTL 2011.

Método	TOP 5 SNP	Precisão	Padrões Genéticos
InfoGain	60, 58, 59, 21, 71	60%	1
SSF	60, 71, 153, 4407, 2126	20%	1
Janela Desl.	53, 52, 56, 57, 51	80%	1
PLINK Single	58, 59, 60, 71, 89	60%	1
PLINK SNP Blocks	60, 58, 59, 21, 71	20%	1
GenSel	50, 30, 4106, 3629, 4088	0%	0

Tabela 2.5 – Desempenho dos algoritmos no QTL 2012.

Método	TOP 5 SNP	Precisão	Padrões Genéticos
InfoGain	6499, 1697, 1683, 1682, 528	60%	2
SSF	6499, 1683, 6571, 4674, 9390	40%	2
Janela Desl.	6498, 6497, 6499, 1683, 9592	80%	2
PLINK Single SNP	6499, 6506, 1682, 1683, 6507	60%	3
PLINK SNP Blocks	1683, 1253, 4675, 7283, 1613	20%	1
GenSel	6491, 1691, 1171, 291, 2671	40%	2

Nota-se que a maioria dos algoritmos encontraram os padrões genéticos mais representativos dos dados simulados, por exemplo a região dos SNP 6499-6500 do QTL-MAS 2012 e a região

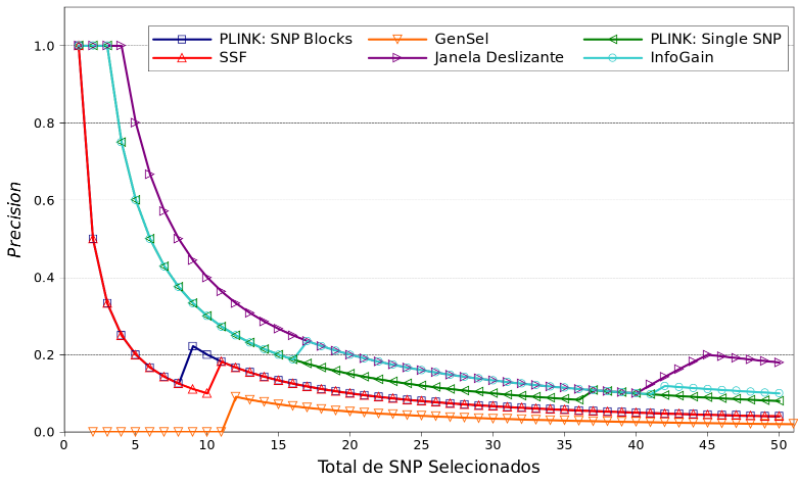
do SNP 57 do QTL-MAS 2011. Algoritmos que utilizam blocos de SNP, no entanto, foram penalizados pela metodologia adotada para avaliar os resultados, como é o caso do GenSel, no qual apesar de ter relatado um bloco de SNP (40 ao 60) que compreendia um padrão genético no QTL-MAS 2011, acabou não sendo contabilizado como acerto, pois foi considerado apenas o SNP médio do bloco, no caso o SNP 50, por exemplo.

Em relação aos gráficos das Figuras 2.7 e 2.8 observa-se que alguns algoritmos apresentam bons resultados em relação ao Precision mas não os mantêm em relação ao Recall, como é o caso da janela deslizante, e vice-versa, como no caso do GenSel. No entanto o algoritmo InfoGain que realiza cálculos mais simples que os demais algoritmos e tem um tempo de execução praticamente instantâneo, apresentou resultados satisfatórios para ambas medidas, mostrando-se uma excelente alternativa para análises cujo tempo de execução é crítico.

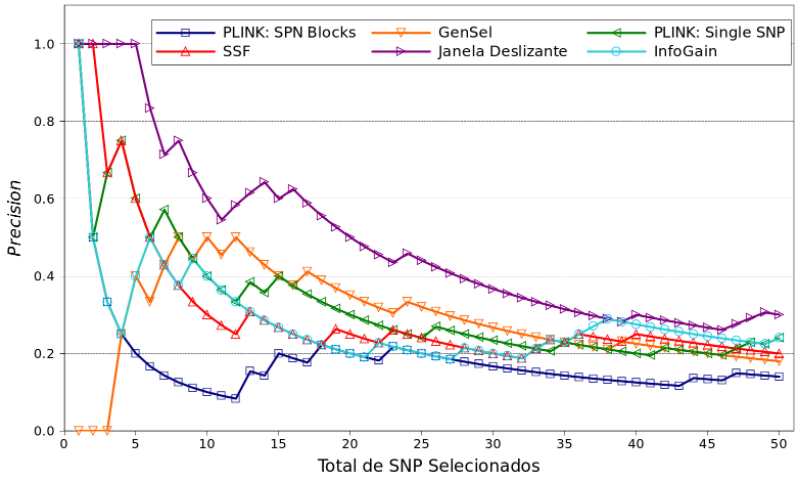
O comportamento dos gráficos de Precision na Figura 2.7 indica que a maioria dos algoritmos tendem a ter uma confiabilidade maior em relação aos atributos mais bem ranqueados durante a seleção, ou seja, quanto mais SNPs selecionados menor é a acurácia deles. Já em relação aos gráficos de Recall o comportamento é o contrário pois conforme o número de SNP selecionados aumentam, maior é a chance de se encontrar um padrão genético.

2.5 Considerações finais

De modo geral todos os algoritmos não-paramétricos apresentaram resultados e tendências semelhantes aos algoritmos paramétricos em relação à seleção dos SNP. Nosso trabalho segue essa tendência e mostra que os métodos não-paramétricos podem ser uma alternativa viável e auxiliar aos métodos tradicionais.

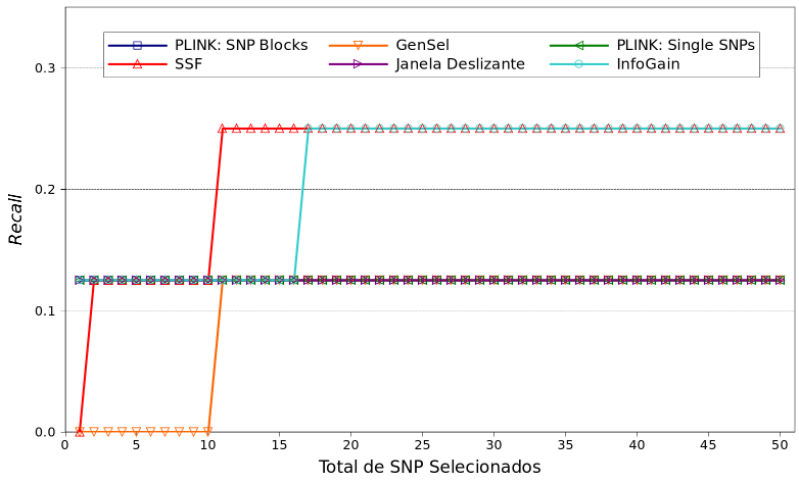


(a) QTL-MAS 2011

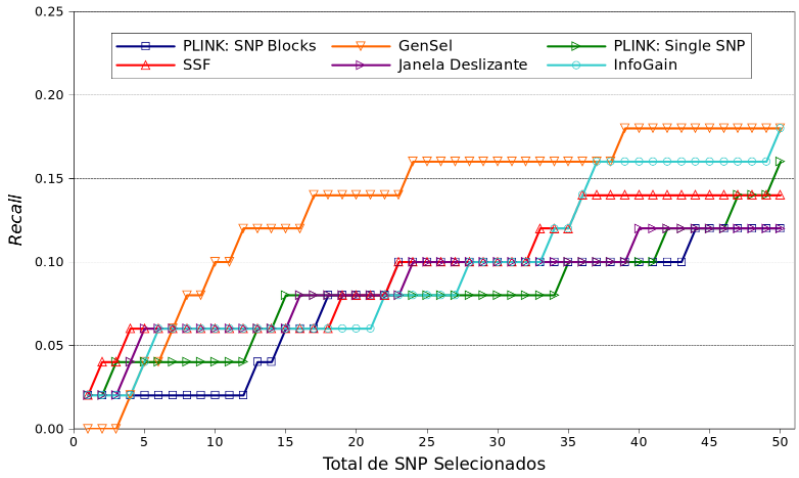


(b) QTL-MAS 2012

Figura 2.7 – Comparação do Precision com diferentes métodos.



(a) QTL-MAS 2011



(b) QTL-MAS 2012

Figura 2.8 – Comparação do Recall com diferentes métodos.

Agradecimentos

Os autores gostariam de agradecer à Patrícia Tholon pela ajuda na parte teórica, à Priscila Neubern pela ajuda com o software GenSel e ao CNPq pelo apoio financeiro.

2.6 Referências

ABIEC. *Estatísticas: balanço da pecuária*. 2014. Disponível em: <http://www.abiec.com.br/>.

AU, W. et al. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE Computer Society, Los Alamitos, CA, USA, v. 2, n. 2, p. 83–101, 2005. ISSN 1545-5963.

BALDING, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, v. 7, n. 10, p. 781–791, Oct 2006.

BARRETT, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, v. 21, n. 2, p. 263–265, Jan 2005.

BELLMAN, R. *Adaptive control processes - A guided tour*. Princeton, New Jersey, U.S.A.: Princeton University Press, 1961. 255 p.

BIGUS, J. P. *Data mining with neural networks: solving business problems from application development to decision support*. Highstown: McGraw-Hill, Inc., 1996.

BROWNING, S. R.; BROWNING, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome

association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, v. 81, n. 5, p. 1084–1097, Nov 2007.

CAMPOS, G. de L. et al. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, v. 193, n. 2, p. 327–345, Feb 2013.

CANTOR, R. M.; LANGE, K.; SINSHEIMER, J. S. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, v. 86, n. 1, p. 6–22, Jan 2010.

CLARKE, G. M. et al. Basic statistical analysis in genetic case-control studies. *Nat Protoc*, v. 6, n. 2, p. 121–133, Feb 2011.

COVER, T. M.; THOMAS, J. A. *Elements of information theory (2. ed.)*. Wiley, 2006. I–XXIII, 1–748 p. ISBN 978-0-471-24195-9. Disponível em: <http://www.elementsofinformationtheory.com/>.

COVOES, T. F.; HRUSCHKA, E. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, v. 181, n. 18, p. 3766 – 3782, 2011. ISSN 0020-0255.

DAETWYLER, H. D. et al. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, v. 193, n. 2, p. 347–365, Feb 2013.

FALCONER, D.; MACKAY, T. *Introduction to Quantitative Genetics*. [S.l.]: Longman, 1996.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. In: *Advances in Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1996. p. 1–34.

FERNANDO, R.; GARRICK, D. *GenSel - User Manual for a portfolia of Genomic Selection related Analyses*. Iowa State

University Animal Breeding & Genetics, 2009. Disponível em: <http://www.biomedcentral.com/content/supplementary/1471-2105-12-186-s1.pdf>.

GRIFFITHS, A. J. F. et al. *Introduction to Genetic Analysis*. 9. ed. [S.l.]: W. H. Freeman, 2007. Hardcover. ISBN 0-7167-6887-9.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 1157–1182, 2003. ISSN 1532-4435.

HALL, D. *Dna-SNP*. Wikipedia Commons, CC BY 2.5, 2007. Acesso em: 11 set. 2014. Disponível em: <http://upload.wikimedia.org/wikipedia/commons/2/2e/Dna-SNP.svg>.

HAYES, B. J. et al. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.*, v. 92, n. 2, p. 433–443, Feb 2009.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)*, v. 4, n. 6, p. 1027–1046, Jun 2014.

LIU, Y. et al. Bos taurus genome assembly. *BMC Genomics*, v. 10, n. 1, p. 180, 2009.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819–1829, Apr 2001.

MITRA, P.; MURTHY, C.; PAL, S. K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 24, n. 3, p. 301–312, 2002. ISSN 0162-8828.

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, v. 26, n. 4, p. 445–455, Feb 2010.

PANICO, B. et al. *QTL-MAS Workshop 2012*. Kassiopea Group srl, 2012. Acesso em: 11 set. 2014. Disponível em: <http://qtl-mas-2012.kassiopeagroup.com/en/index.php>.

POWERS, D. M. W. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Adelaide, Australia, 2007.

PURCELL, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, v. 81, n. 3, p. 559–575, Sep 2007.

ROLF, M. M. et al. Genomics in the united states beef industry. *Livestock Science*, v. 166, n. 0, p. 84 – 93, 2014. Genomics Applied to Livestock Production. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1871141314003138>.

ROY, P. L. et al. *QTL-MAS Workshop 2011*. The Animal Genetics Division at INRA, 2011. Acesso em: 11 set. 2014. Disponível em: <https://colloque4.inra.fr/qtlmas>.

SHAH, S. C.; KUSIAK, A. Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine*, v. 31, n. 3, p. 183–196, 2004. Disponível em: <http://dx.doi.org/10.1016/j.artmed.2004.04.002>.

WANG, J.; ZUCKER; JEAN-DANIEL. Solving multiple-instance problem: A lazy learning approach. In: LANGLEY, P. (Ed.). *17th International Conference on Machine Learning*. [S.l.: s.n.], 2000. p. 1119–1125.

WITTEN, I.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0120884070.

ZHANG, Z.; ZHANG, Q.; DING, X. Advances in genomic selection in domestic animals. *Chinese Science Bulletin*, Chinese Science Bulletin, v. 56, n. 25, p. 2655, 2011. Disponível em: <http://csb.scichina.com:8080/kxtbe/EN/abstract/article_504345.shtml>.