Fabrízzio Condé de Oliveira

Um método para seleção de atributos em dados genômicos

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Orientador: Prof. D.Sc. Carlos Cristiano Hasenclever Borges

Coorientador: Prof. D.Sc. Wagner Antonio Arbex

Juiz de Fora

2015

Oliveira, Fabrízzio Condé de

Um método para seleção de atributos em dados genômicos/Fabrízzio Condé de Oliveira. – Juiz de Fora: ICE/Engenharia, 2015.

XXIII, 273 p.: il.; 29, 7cm.

Orientador: Carlos Cristiano Hasenclever Borges

Coorientador: Wagner Antonio Arbex

Tese (doutorado) – ICE/Engenharia/Programa de Modelagem Computacional, 2015.

Referências Bibliográficas: p. 247 – 267.

1. Estudos de Associação em Escala Genômica. 2. Polimorfismos de Base Única. 3. Máquina de Vetores Suporte. 4. Florestas Aleatórias. 5. Algoritmos Genéticos. I. Borges, Carlos Cristiano Hasenclever *et al.*. II. Universidade Federal de Juiz de Fora, MMC, Programa de Modelagem Computacional.

Fabrízzio Condé de Oliveira

Um método para seleção de atributos em dados genômicos

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Aprovada em 26 de Novembro de 2015.

BANCA EXAMINADORA

Prof. D.Sc. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Wagner Antonio Arbex - Coorientador
Universidade Federal de Juiz de Fora

Profa. D.Sc Priscila Vanessa Zabala Capriles Goliatt
Universidade Federal de Juiz de Fora

Prof. D.Sc Raul Fonseca Neto
Universidade Federal de Juiz de Fora

Prof. D.Sc Fabyano Fonseca e Silva
Universidade Federal de Viçosa

Prof. D.Sc Moysés Nascimento
Universidade Federal de Viçosa

À minha filha Alícia, à minha esposa Flávia, à minha mãe Maria do Carmo, ao meu padrasto Francisco, à minha sogra Maria Arlete, ao meu avô Luiz Gonzaga e à minha avó Esther in memoriam.

AGRADECIMENTOS

À Deus pelo dom da vida.

À Universidade Federal de Juiz de Fora, pela oportunidade de realização do doutorado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pela concessão da bolsa de estudos.

À Embrapa Gado de Leite pelo apoio na realização deste estudo.

Ao meu orientador professor Carlos Cristiano Hasenclever Borges pela confiança, pelo apoio nos momentos mais difíceis, pelos ensinamentos profissionais e de vida, pela amizade e pela competência com que me orientou.

Ao meu coorientador professor Wagner Arbex pela confiança, por todas as oportunidades dadas ao longo de nossa amizade, pelo apoio nos momentos mais difíceis e pela competência com que me orientou.

Ao pesquisador Marcos Vinicius Gualberto Barbosa da Silva pelos ensinamentos de Genética, pelas críticas construtivas e pelo apoio constante para o desenvolvimento desse trabalho.

À pesquisadora Fernanda Nascimento Almeida pelos conselhos, pelos ensinamentos e pela amizade.

Aos professores Henrique Hippert e Saul de Castro Leite pelos ensinamentos e pela constante ajuda ao longo desse trabalho.

Aos amigos Aldemon Lage Bonifácio e Bruno Zonovelli da Silva pela amizade, por toda ajuda, pelo companherismo, pelo agradável convívio e pelas discussões enriquecedoras e acaloradas, as quais tornaram essa pesquisa extremamente empolgante.

A todos os alunos do mestrado e do doutorado em Modelagem Computacional por tornar o ambiente mais agradável.

Aos professores do Programa de Pós-graduação em Modelagem Computacional pelos ensinamentos, pelo apoio e pelo convívio.

A todos que contribuíram direta ou indiretamente para a conclusão desse trabalho.

'Se um homem tem um talento e não tem capacidade de usá-lo, ele fracassou. Se ele tem um talento e usa somente a metade deste, ele fracassou parcialmente. Se ele tem um talento e de certa forma aprende a usá-lo em sua totalidade, ele triunfou gloriosamente e obteve uma satisfação e um triunfo que poucos homens conhecerão.

Thomas Wolfe '

RESUMO

Estudos de associação em escala genômica buscam encontrar marcadores moleculares do tipo SNP que estão associados direta ou indiretamente a um fenótipo em questão tais como, uma ou mais características do indivíduo ou, até mesmo, uma doença. O SNP pode ser a própria mutação causal ou pode estar correlacionado com a mesma por serem herdados juntos. Para identificar a região causadora ou promotora do fenótipo, a qual não é conhecida *a priori*, milhares ou milhões de SNPs são genotipados em amostras compostas de centenas ou milhares de indivíduos. Com isso, surge o desafio de selecionar os SNPs mais informativos no conjunto de dados genotípico, onde o número de atributos é, geralmente, muito superior ao número de indivíduos, com a possibilidade de que existam atributos altamente correlacionados e, ainda, podendo haver interações entre pares, trios ou combinações de SNPs de quaisquer ordens. Os métodos mais usados em estudos de associação em escala genômica utilizam o valor-p de cada SNP em testes estatísticos de hipóteses, baseados em regressão para fenótipos contínuos e baseados nos testes qui-quadrado ou similares em classificação para fenótipos discretos, como filtro para selecionar os SNPs mais significativos. Entretanto, essa classe de métodos captura somente SNPs com efeitos aditivos, pois a relação adotada é linear. Na tentativa de superar as limitações de procedimentos já estabelecidos, este trabalho propõe um novo método de seleção de SNPs baseado em técnicas de Aprendizado de Máquina e Inteligência Computacional denominado SNP *Markers Selector* (SMS). O modelo é construído a partir de uma abordagem que divide o problema de seleção de SNPs em três fases distintas: a primeira relacionada à análise de relevância dos marcadores, a segunda responsável pela definição do conjunto de marcadores relevantes que serão considerados por meio de uma estratégia de corte com base em um limite de relevância dos marcadores e, finalmente, uma fase para o refinamento do processo de corte, geralmente para diminuir marcadores falsos-positivos. No SMS, essas três etapas, foram implementadas utilizando-se Florestas Aleatórias, Máquina de Vetores Suporte e Algoritmos Genéticos respectivamente. O SMS objetiva a criação de um fluxo de trabalho que maximize o potencial de seleção do modelo através de etapas complementares. Assim, espera-se aumentar o potencial do SMS capturar efeitos aditivos e/ou não-aditivos com interação moderada entre pares e trios de SNPs, ou até mesmo, interações de ordens superiores com efeitos que sejam

minimamente detectáveis. O SMS pode ser aplicado tanto em problemas de regressão (fenótipo contínuo) quanto de classificação (fenótipo discreto). Experimentos numéricos foram realizados para avaliação do potencial da estratégia apresentada, com o método sendo aplicado em sete conjuntos de dados simulados e em uma base de dados real, onde a capacidade de produção de leite predita de vacas leiteiras foi medida como fenótipo contínuo. Além disso, o método proposto foi comparado com os métodos baseados no valor-p e com o Lasso Bayesiano apresentando, de forma geral, melhores resultados do ponto de vista de SNPs verdadeiros-positivos nos dados simulados com efeitos aditivos juntamente com interações entre pares e trios de SNPs. No conjunto de dados reais, baseado em 56.947 SNPs e um único fenótipo relativo à produção de leite, o método identificou 245 QTLs associados à produção e à composição do leite e 90 genes candidatos associados à mastite, à produção e à composição do leite, sendo esses QTLs e genes identificados por estudos anteriores utilizando outros métodos de seleção. Assim, o método demonstrou ser competitivo frente aos métodos utilizados para comparação em cenários complexos, com dados simulados ou reais, o que indica seu potencial para estudos de associação em escala genômica em humanos, animais e vegetais.

Palavras-chave: Estudos de Associação em Escala Genômica. Polimorfismos de Base Única. Máquina de Vetores Suporte. Florestas Aleatórias. Algoritmos Genéticos.

ABSTRACT

Genome-wide association studies have as main objective to discovery SNP type molecular markers associated directly or indirectly to a specific phenotype related to one or more characteristics of an individual or even a disease. The SNP could be the causative mutation itself or correlated with the causative mutation due to common inheritance. Aiming to identify the causal or promoter region of the phenotype, which is unknown *a priori*, thousands or millions of SNPs are genotyped in samples composed of hundreds or thousands of individuals. Therefore, emerges the necessity to confront a challenge of selecting the most informative SNPs in genotype data set where the number of attributes are, usually, much higher than the number of individuals. Besides, the possibility of highly correlated attributes should be considered, as well as interactions between pairs, trios or combinations of high order SNPs. The most usual methods applied on genome-wide association studies adopt the p-value of each SNP as a filter to select the SNPs most significant. For continuous phenotypes the statistical regression-based hypothesis test is used and the Chi-Square test or similar for classification of discrete phenotypes. However, this class of methods capture only SNPs with additive effects, due to the linear relationship considered. In an attempt to overcome the limitations of established procedures, this work proposes a new SNPs selection method, named SNP Markers Selector (SMS), based on Machine Learning and Computational Intelligence strategies. The model is built considering an approach which divides the SNPs selection problem in three distinct phases: the first related to the evaluation of the markers relevance, a second responsible for the definition of the set of the relevant markers that will be considered by means of a cut strategy based on a threshold of markers relevance and, finally, a phase for the refinement of the cut process, usually to diminish false-positive markers. In the SMS, these three steps were implemented using Random Forests, Support Vector Machine and Genetic Algorithms, respectively. The SMS intends to create a workflow that maximizes the SNPs selection potential of the model due to the adoption of steps considered complementary. In this way, there is an increasing expectation on the performance of the SMS to capture additive effects, moderate non-additive interaction between pairs and trios of SNPs, or even, higher order interactions with minimally detectable effects. The SMS can be applied both in regression problems (continuous phenotype) as in classification problems

(discrete phenotype). Numerical experiments were performed to evaluate the potential of the strategy, with the method being applied in seven sets of simulated data and in a real data set, where milk production capacity predicated of dairy cows was measured as continuous phenotype. Besides, the comparison of the proposed method with methods based on p-value and Lasso Bayesian technique indicate, in general, competitive results from the point of view of true-positive SNPs using simulated data set with additive effects in conjunction with interactions of pairs and trios of SNPs. In the real data, based on 56,947 SNPs and a single phenotype of milk production, the method identified 245 QTLs associated with milk production and composition and 90 candidate genes associated with mastitis, milk production and composition, standing out that these QTLs and genes were identified by previous studies using other selection methods. Thus, the experiments showed the potential of the method in relation to other strategies when complex scenarios with simulated or real data are adopted, indicating that the workflow developed to guide the construction of the method should be considered for genome-wide asociation studies in humans, animals and plants.

Keywords: Genome-wide association studies. Single Nucleotide Polymorphisms. Support Vector Machine. Random Forests. Genetic Algorithms.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

1 Introdução

1.1 Caracterização do Problema

O genoma humano contém aproximadamente três bilhões de pares de base, sendo a maior parte composta por regiões não-codificantes ou íntrons[1] (LAIRD; LANGE, 2011). Com o sequenciamento completo do genoma humano, imaginou-se que o entendimento da relação genótipo-fenótipo seria relativamente rápido, porém, isso ainda não ocorreu devido à complexidade da interação genótipo-ambiente-fenótipo. Os genes[2] são as unidades fundamentais de hereditariedade, logo, o mapeamento deles é imprescindível para elucidar a determinação de diversos fenótipos tanto relacionados às doenças quanto aos traços benéficos. Estimativas atuais sugerem que existem entre 20.000 a 30.000 genes distribuídos no genoma (LAIRD; LANGE, 2011).

Como a quantidade de fenótipos nos seres vivos é elevada, então existem fenótipos distintos que são influenciados por vários genes em comum, apesar de alguns fenótipos serem determinados por um ou poucos genes. Assim, buscar genes responsáveis por determinados fenótipos, ou, até mesmo, variações nos genes que produzam variações em determinada característica é uma tarefa de grande relevância para Ciências Biológicas. Segundo Silva (2002), a identificação de genes pode levar a aplicações úteis em várias áreas, tais como a identificação de genes relacionados às doenças cardíacas ou a diabetes, e no aumento da eficiência de seleção no melhoramento animal, especialmente, em características de baixa herdabilidade[3] ou naquelas que somente podem ser medidas após o abate do animal ou em apenas um sexo. Em conjunto aos genes, múltiplos fatores ambientais também interferem no processo biológico para a determinação do fenótipo.

[1] Regiões não-codificantes ou íntrons são regiões que não codificam proteínas, mas podem estar envolvidos na regulação da produção de proteínas (LAIRD; LANGE, 2011). Os íntrons são regiões não-codificantes do DNA que são removidos do RNA mensageiro para a produção do RNA maduro (PIERCE, 2013).

[2] Gene é normalmente definido como um segmento de DNA que contém as instruções para produzir uma determinada proteína, embora essa definição sirva para a maioria dos genes, vários deles produzem moléculas de RNA ao invés de proteínas como produto final (ALBERTS et al., 2010). A grosso modo, os genes são as unidades funcionais para a hereditariedade.

[3] A **herdabilidade em sentido amplo** (H^2) é a variância fenotípica devida à variância genética e é calculada dividindo a variância genética pela variância fenotípica (PIERCE, 2013). A **herdabilidade em sentido restrito** (h^2) é igual à variância genética aditiva dividida pela variância fenotípica (PIERCE, 2013).

O desenvolvimento de doenças comuns resultam de complexas interações entre numerosos fatores ambientais e alelos[4] de muitos genes (WANG et al., 2005). Identificar os alelos de genes que afetam o risco de desenvolvimento da doença ajudará a entender a etiologia da doença (WANG et al., 2005).

Polimorfismos de base única (do inglês, *Single Nucleotide Polymorphisms* - SNPs) são uma forma abundante de variação genômica, que se diferem das variações raras denominadas mutações, ou seja, são mutações não-deletérias que se estabeleceram na população com uma frequência mínima de 1% (BROOKES, 1999). Por meio do uso de SNPs como marcadores de regiões no genoma, estudos de associação em escala genômica (do inglês, *Genome-Wide Association Studies* - GWAS) buscam identificar *loci* associados com a característica de interesse (GONDRO; WERF; HAYES, 2013). O pressuposto básico para os estudos de associação em escala genômica é que a característica avaliada pode ser explicada a partir desse tipo de marcador genético. Assim, considera-se que existirão SNPs com alto desequilíbrio de ligação[5] (atributos altamente correlacionados) com o verdadeiro *locus* da mutação causal (*Quantitative Trait Loci* - QTL). Portanto, não é necessário encontrar o verdadeiro *locus* da variante causal, pois algum SNP avaliado terá grande parte da informação da mesma, podendo o SNP ser a própria mutação causadora do fenótipo.

A grande contribuição do GWAS para humanos é a descoberta de quais genes podem aumentar ou diminuir o risco de desenvolvimento de determinada doença. Em conjunto, pode-se agregar covariáveis ambientais às variáveis genômicas (marcadores SNP) para mapear a interação gene-ambiente sobre os atributos fenotópicos mensuráveis. Nota-se pela Figura 1.1, que esse ramo da genômica tem despertado grande interesse pela comunidade científica.

Para o processo de previsão do valor genético genômico usado em Seleção Genômica (do inglês, *Genome-Wide Selection* - GWS) para selecionar animais superiores em relação à alguma característica em programas de melhoramento genético, a seleção de marcadores SNP informativos para produção de *chips*[6] de genotipagem com baixa densidade podem ser viáveis para determinados tipos de arquiteturas referente ao traço genético avaliado, reduzindo drasticamente o custo de genotipagem (MORSER; HAYES; RAADSMA, 2010).

[4] Alelos ou genes alelos são as várias formas de um gene (PIERCE, 2010)

[5] Associação não-aleatória entre SNPs (PIERCE, 2013).

[6] *Chips* de DNA de alta densidade foram criados para genotipar de dezenas de milhares até centenas de milhares de marcadores SNP em um único ensaio (CAETANO, 2009).

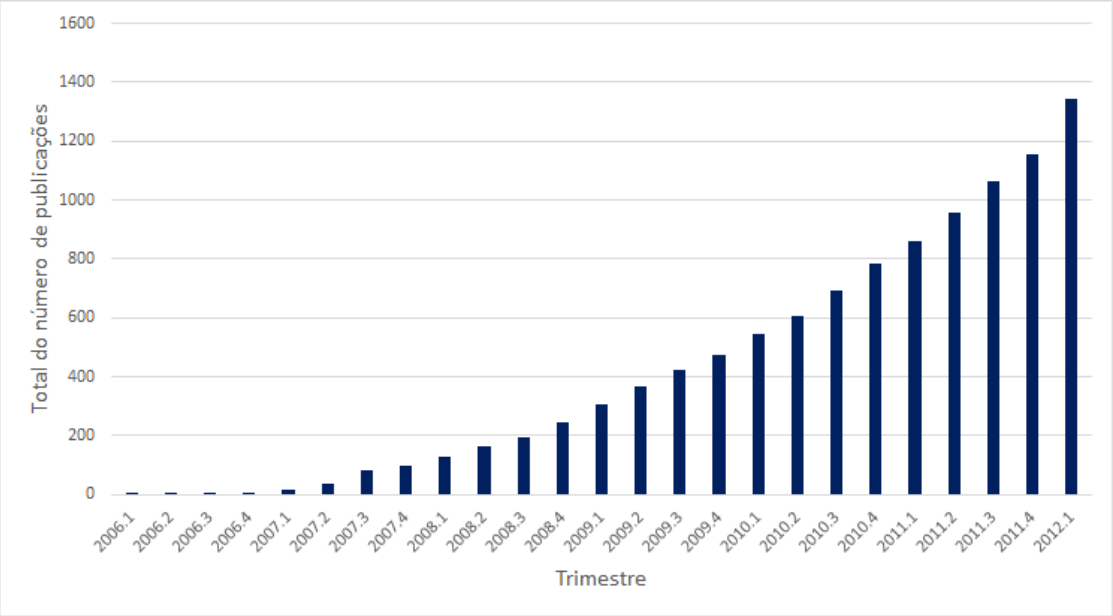


Figura 1.1 Número de publicações sobre GWAS entre 2005 e 2012.

National Human Genome Research Institute-NHGRI. Adaptado de Welter et al. (2014).

Assim, um ensaio de baixa densidade de SNPs espaçados pode fornecer uma precisão suficiente de previsão para avaliação de animais, desde que o conteúdo da informação do subconjunto de SNPs seja suficiente para estimar os efeitos de haplótipos distintos (MORSER; HAYES; RAADSMA, 2010).

Outra aplicação é na rastreabilidade de carne bovina comercializada. A partir da genotipagem de uma amostra de carne de um animal com *chips* economicamente acessíveis compostos de poucas dezenas de marcadores genéticos, pode-se descobrir rapidamente a origem desse indivíduo, a partir de um banco de dados adequado, diminuindo a possibilidade de fraudes nas informações de cada animal. Com o objetivo de fornecer um conjunto de SNPs úteis para a rastreabilidade de carne de bovino, Orru et al. (2009) testou 63 SNPs para a capacidade de identificar indivíduos isolados em seis raças européias de gado. Dezoito SNPs altamente informativos localizados em diferentes genes, foram selecionados. Ao utilizar este painel de SNPs a probabilidade de que um indivíduo esteja incorretamente classificado varia de 1,39 para 0,07 de 1 milhão, dependendo da raça.

GWAS exploram o desequilíbrio de ligação, que são as associações em nível de população entre os marcadores e mutações causais, também chamado de *locus* de características quantitativas (GONDRO; WERF; HAYES, 2013). Estas associações surgem porque há pequenos segmentos do cromossomo na população atual a partir da

qual existe um ancestral comum. Estes segmentos cromossômicos, que remontam a um mesmo ancestral comum sem intervenção de recombinação, criam alelos de marcadores idênticos (GONDRO; WERF; HAYES, 2013). Se houver algum QTL em um segmento do cromossomo, os marcadores também carregam os alelos idênticos ao do QTL. Há uma série de metodologias estatísticas para explorar essas associações.

O teste de associação mais comumente usado com fenótipos contínuos é baseado na construção do modelo de regressão linear simples entre o marcador SNP e o fenótipo em questão. Esses modelos de regressão baseados em um único marcador (monoatributo) são os mais usados em GWAS no sentido de que baseiam-se na análise do valor-p de algum teste de hipóteses para associação entre duas variáveis. Por exemplo, com fenótipo dicotômico (codificados como 0 ou 1), pode-se realizar múltiplos testes qui-quadrado independentes entre cada marcador e o fenótipo e, então, ordenar os valores-p desses marcadores de forma crescente. A seguir, defini-se um ponto de corte, de tal forma que todos os marcadores cujos valores-p são menores que esse limite sejam estudados posteriormente para avaliar a função biológica dos genes indicados por esses marcadores. Ou até mesmo, sugerir que regiões do DNA, ainda não estudadas, possam definir genes hipotéticos associados com a característica fenotípica. As vantagens dessa abordagem monoatributo são baixo tempo de processamento, facilidade de uso e de interpretação dos resultados.

1.2 Motivação

A partir do trabalho de Easton e Eeles (2008), com a abordagem monoatributo de testes estatísticos, novos *loci* foram identificados para câncer de mama, câncer de próstata, câncer colorretal, câncer de pulmão e melanoma. Porém, a principal desvantagem é considerar somente o efeito principal de cada marcador sobre algum fenótipo, desconsiderando possíveis sinergias entre os marcadores, como por exemplo, epistasia [7] e codominância [8].

Até o momento, a previsão do risco de doenças em humanos baseada em SNPs

[7] Epistasia (do Grego *epi*, sobre, *stasis*, parada, inibição) é simplesmente a interação entre genes. A epistasia ocorre quando a ação de um gene é modificada por um ou diversos genes que se associam independentemente (PIERCE, 2013).

[8] Codominância é um tipo de interação entre alelos de um gene funcional onde o indivíduo heterozigoto produz os dois fenótipos referentes aos dois homozigotos (PIERCE, 2013).

validados com base nessa metodologia mostrou pouco poder preditivo (MITTAG et al., 2012), apesar de tais SNPs indicarem alta significância de associação com a expressão fenotípica. De acordo com Moore, Asselbergs e Williams (2010), os *loci* identificados por GWAS tipicamente tem muitos efeitos de pequena magnitude. Isso acontece, por exemplo, para o câncer onde o aumento no risco para a suscetibilidade dos alelos nos *loci* descobertos possui na maior parte *odds ratio* [9] menor ou igual a 1,3 (MOORE; ASSELBERGS; WILLIAMS, 2010). Sabendo que o câncer de mama é uma doença com influência genética elevada e que é bastante razoável considerá-la com etiologia simples, Easton et al. (2007) encontrou cinco marcadores significantes, sendo que quatro variantes descobertas estão em genes conhecidos e o outro em um gene hipotético. Considerando um modelo multiplicativo, os cinco *loci* juntos explicam 3,6% do excesso de risco para o desenvolvimento do risco de câncer de mama. Esse fato pode ser explicado devido à variação dos marcadores mais significativos possuir baixo poder explicativo em relação à variação fenotípica.

Como outro exemplo de falha da metodologia mais usual em GWAS, um estudo feito sobre o câncer de pâncreas em Amundadottir et al. (2009) que analisou mais de 500.000 SNPs em uma amostra de 1.896 pacientes com a doença (casos) e 1.939 sem a doença (controles). Os autores avaliaram também uma amostra de replicação com 2.457 casos e 2.654 controles (AMUNDADOTTIR et al., 2009). Um modelo de regressão logística foi aplicado em ambas amostras e identificou um único SNP com uma *odds ratio* igual a 1,2. Este SNP foi localizado em um íntron do gene do grupo sanguíneo *ABO*. Esse resultado confirmou o que estudos epidemiológicos anteriores indicaram que tal grupo sanguíneo está associado com um menor risco de câncer de pâncreas. Além disso, esse achado já foi realizado a mais de 50 anos atrás, não representando originalidade na descoberta (MOORE; ASSELBERGS; WILLIAMS, 2010). Isso mostra limitações da metodologia mais frequentemente empregada em GWAS para a descoberta de novos *loci*, relacionados à suscetibilidade de algumas doenças.

Apesar do GWAS sobre o câncer de mama ter indicado novos *loci*, eles não podem ser usados em testes genéticos para predizer ou prevenir essa doença (MOORE; ASSELBERGS; WILLIAMS, 2010) devido aos classificadores com baixa acurácia

[9] *Odds ratio* (geralmente abreviado "OR") é uma das três principais formas de quantificar quão fortemente a presença ou ausência da propriedade A está associada com a presença ou ausência da propriedade B em uma dada população (CORNFIELD, 1951).

construídos com esses atributos. Desta forma, espera-se que a metodologia mais adotada em GWAS aplicada em doenças comuns, com arquitetura genômica complexa, como câncer de mama "esporádico", isto é, com baixa herdabilidade, e diabetes tipo 2 seja desencorajadora (MOORE; ASSELBERGS; WILLIAMS, 2010). Uma abordagem alternativa é ampliar o número de marcadores, considerando também os que possuem pequenas correlações sobre a característica avaliada. Mas, tal fato cria três problemas, a saber: (i) o número de marcadores pode ser elevado, (ii) com muitos deles sendo correlacionados entre si, gerando correlações espúrias com a característica, e (iii) podem também apresentar associações complexas com o fenótipo indicando interações não-lineares de alta ordem entre os marcadores. Considerando somente os dois primeiros problemas anteriores, de acordo com Gianola, Perez-Enciso e Toro (2003), tal análise demanda a utilização de métodos estatísticos que considerem a seleção de covariáveis (problema de multicolinearidade[10]) e a regularização do processo de estimação (problema de redução de dimensionalidade). Outras técnicas de regressão foram criadas para abordar esses dois problemas como regressão *ridge* e regressão por mínimos quadrados parciais (MORSER; HAYES; RAADSMA, 2010).

No que tange à complexidade na relação genótipo-fenótipo, Moore e Williams (2009) argumentam que a modelagem linear (regressão linear simples) frequentemente usada em problemas de GWAS considera somente um SNP por vez, logo, ignora o contexto genômico (interação gene-gene) e ambiental (interação gene-ambiente) de cada SNP. Para tentar contornar as deficiências anteriores, algoritmos de Aprendizado de Máquina[11] tais como Máquinas de Vetores Suporte (do inglês, *Support Vector Machine* - SVM), considerando múltiplos marcadores em problemas de classificação, vêm demonstrando desempenho satisfatório como em Mittag et al. (2012), Wei et al. (2009) e Ban et al. (2010). Para fenótipos contínuos, Máquinas de Vetores Suporte para Regressão (do inglês, *Support Vector Regression* - SVR) tem demonstrado potencial ao ser aplicado na previsão de doses de varfarina[12] em afro-americanos e apresentaram R^2 de 57,8% entre o valor predito e o observado (COSGUN; DUARTE, 2011).

Caso a estrutura genética subjacente a um fenótipo seja definida por um grande

[10] As variáveis independentes possuem relações lineares exatas ou aproximadamente exatas em problemas de regressão múltipla (GUJARATI, 2006).

[11] O estudo e modelagem computacional dos processos de aprendizagem em suas múltiplas manifestações constitui o problema da Aprendizagem de Máquina (MICHALSKI; CARBONELL; MITCHELL, 2013).

[12] Varfarina é um fármaco do grupo dos anticoagulantes, que é usado na prevenção das tromboses.

número de *Quantitative Trait Loci* (QTL) [13] com pequenos efeitos juntamente com a evolução de novos *chips*, com densidades de 500.000 a 800.000 marcadores, a formação de grandes conjuntos de dados e medidas fenotípicas precisas serão necessárias para realizar a melhoria da acurácia no aumento da densidade de SNPs (HARRIS; JOHNSON, 2010). Consequentemente, é de suma importância que novas metodologias sejam desenvolvidas para tratar adequadamente dados genômicos com alta dimensionalidade sem a eliminação de variáveis relevantes. Portanto, após a identificação do subconjunto de marcadores suficientes e necessários para a explicação do fenótipo, é possível a redução de custos na confecção de *chips* personalizados com menor quantidade de SNPs para a previsão do fenótipo baseados em métodos de Seleção Genômica (GWS).

Moore e Ritchie (2004) têm discutido sobre três grandes desafios a serem superados para obter êxito na identificação de variações genéticas que estão associadas com a saúde ou com a doença em uma abordagem do genoma inteiro. O primeiro desafio é referente aos métodos poderosos de Mineração de Dados e Aprendizado de Máquina que terão de ser desenvolvidos para modelar computacionalmente a relação entre combinações de SNPs, outras variações genéticas e a exposição ambiental com a suscetibilidade à doença. Isso é necessário porque abordagens estatísticas paramétricas tradicionais como a regressão logística têm poder limitado para capturar interações não-lineares de alta ordem, as quais são importantes na etiologia de doenças complexas (MOORE; WILLIAMS, 2002).

O segundo desafio é a seleção de SNPs que devem ser incluídos na análise. Se as interações não-lineares entre os genes explicam uma parte significativa da herdabilidade de doenças comuns, então as combinações de SNPs terão que ser avaliadas a partir de um conjunto inicial de milhares ou milhões de candidatos. Algoritmos de filtragem e/ou algoritmos de busca estocástica ou *wrapper*[14] irão desempenhar um papel fundamental para GWAS (MOORE; ASSELBERGS; WILLIAMS, 2010). Esse fato ocorre porque o custo computacional de avaliar todas as combinações possíveis em um conjunto de dados com milhões de SNPs é, atualmente, alto.

Um terceiro desafio é a interpretação biológica de modelos genéticos não-lineares. Mesmo quando um modelo computacional pode ser usado para identificar genótipos

[13] São trechos de DNA que contém ou são ligados aos genes que determinam uma característica quantitativa.

[14] São métodos de seleção de atributos onde o algoritmo de busca é separado do algoritmo indutor do aprendizado, sendo este último usado somente para a avaliação do conjunto de atributos (KOHAVI; JOHN, 1998).

de SNPs que aumentam a suscetibilidade à doença, as especificidades das relações matemáticas não podem ser traduzidos em estratégias de prevenção e tratamento, sem interpretação dos resultados no contexto da biologia humana (MOORE; ASSELBERGS; WILLIAMS, 2010). Fazer inferências a partir de modelos computacionais etiológicos pode ser o mais importante e o mais difícil desafio de todos (MOORE; RITCHIE, 2004).

Outro ponto crítico que emerge em problemas de seleção de marcadores genéticos é a quantificação da variável referente ao fenótipo em questão, podendo a mesma ser categórica, como exemplo indivíduos que possuem ou não uma determinada doença, ou contínua, como a altura de seres humanos, as quais podem assumir diversos valores fracionários em relação a um padrão de medida. Muitas técnicas de seleção de marcadores do tipo SNP trabalham apenas com variável categórica para o fenótipo como por exemplo, *Troost* (NETO, 2013) e *MIGA-2L* (OLAZAR, 2013), sendo necessária a construção de classes pré-definidas para fenótipos contínuos, o que geralmente pode gerar perda de informação. Métodos de seleção devem ser flexíveis para tratar os dois tipos de mensuração supramencionados, mantendo a variável fenotípica no seu formato original.

Uma das alternativas aos modelos estastísticos frequentistas ou bayesianos é a utilização de ferramentas de Aprendizado de Máquina, pois possuem algumas vantagens, tais como, robustez às premissas sobre distribuições de probabilidades de parâmetros, alta acurácia, flexibilidade em modelar efeitos não-lineares, algoritmos bem desenvolvidos e a capacidade de trabalhar sobre conjuntos de dados em alta dimensão (SZYMCZAK et al., 2009). Além disso, o SVM e o SVR têm demonstrado resultados promissores em problemas de GWAS e, com base nos trabalhos Freitas (2001) e Packard (1990) sobre a seleção de atributos com interação entre os mesmos, acredita-se que o uso de meta-heurísticas como Algoritmos Genéticos (do inglês, *Genetic Algorithms* - GA) podem ser extrapolados para GWAS, dado que neste contexto podem existir interações entre marcadores do tipo SNP. O uso de GA em seleção de variáveis, com ocorrência de interação entre as mesmas, podem gerar melhores resultados que outras metodologias frequentemente usadas em GWAS, as quais são baseadas em *greedy search*[15] como, por exemplo, *Stepwise*. Freitas (2001) faz esse comparativo para conjuntos de dados advindos de contextos distintos de GWAS.

O desenvolvimento desta tese tem por objetivo a construção de um modelo baseado

[15]Um algoritmo baseado em *greedy search* ou algoritmo guloso é um algoritmo que permite a resolução de problemas por meio de heurísticas e faz a escolha localmente ótima em cada fase com a esperança de encontrar um ótimo global (BLACK, 2004).

em Aprendizado de Máquina e Inteligência Computacional resolvido por técnicas computacionais para seleção de marcadores do tipo SNP, maximizando o número de marcadores relevantes associados a um fenótipo de interesse e, concomitantemente, eliminando os redundantes e não-informativos. Essa tese contribuiu para o projeto de pesquisa da Embrapa "Modelos computacionais para estabelecimento de meios e procedimentos metodológicos para análise de dados em bioinformática" (MCBio) descrito detalhadamente em Arbex et al. (2010).

É importante diferenciar estudos de associação em escala genômica (GWAS) e estudos de seleção genômica (GWS). Estudos de associação em escala genômica (GWAS) objetivam encontrar SNPs que marcam regiões no genoma que influenciam determinada característica. Do ponto de vista estatístico, os modelos usados nesses estudos selecionam os SNPs informativos por meio do aumento da previsão do fenótipo. Portanto, o poder preditivo do modelo construído é um meio para selecionar os marcadores causais para a característica em questão, porém, a previsão não é o objetivo primário nesses trabalhos, mas sim a busca pelos genes responsáveis pela variação fenotípica. Por outro lado, existem trabalhos cujo objetivo principal é a previsão de determinado fenótipo como é o caso de trabalhos em GWS. Os métodos usados em GWS tem por objetivo predizer adequadamente o valor genético genômico estimado de animais ou vegetais para selecionar os melhores em relação a um fenótipo com objetivo de obter ganhos de produtividade para essa característica. Desta forma, a finalidade de GWS é buscar modelos que tenham melhor predição para que a seleção de animais seja feita com maior confiança. Cabe ressaltar que alguns estudos de GWS podem utilizar algoritmos de seleção de SNPs, desenvolvidos inicialmente para GWAS, para aumentar o poder preditivo dos modelos analisados.

1.3 Trabalhos Correlatos em GWAS

Técnicas de aprendizado de máquina têm sido amplamente empregadas em trabalhos recentes de GWAS com a abordagem caso-controle em humanos. No estudo conduzido por Wei et al. (2009), o SVM para GWAS foi aplicado em dados de SNPs para diabetes do tipo 1 filtrados inicilamente pelo valor-p do teste de associação qui-quadrado. A acurácia predita (área abaixo da curva ROC) foi analisada em duas amostras independentes e

encontrou-se aproximadamente 0,84.

Em Okser et al. (2010), os autores avaliaram a predição de classes extremas do risco de aterosclerose por meio de um classificador *naives Bayes*, utilizando uma estratificação com base no quantitativo da ultra-sonografia da espessura íntima-média da artéria carótida. A área abaixo da curva ROC encontrada pelo método sugerido foi 0,844 contra 0,761 obtido somente por variáveis clínicas.

Nos estudos de Wei et al. (2009) e Okser et al. (2010), os investigadores encontraram acurácia preditiva muito superior quando incluíram uma quantidade maior de SNPs estatisticamente não-significativos, e, comparativamente um desempenho muito inferior quando incluíram apenas SNPs estatisticamente significativos. Isso demonstra que os SNPs mais significativos pelo valor-p, não servem para predizer o risco de desenvolvimento de determinada doença como foi discutido por Moore, Asselbergs e Williams (2010).

O SVM também foi aplicado em Ban et al. (2010) com o objetivo de descoberta de genes em dados genômicos associados a diabetes tipo 2 em uma amostra de indivíduos coreanos. Neste trabalho, uma taxa de predição de 65,3% foi relacionada com 14 SNPs em 12 genes usando o SVM com *kernel* de função de base radial (do inglês, *Radial Base Function* - RBF), adicionalmente, novas associações entre certas combinações de marcadores e a doença foram obtidas.

Mittag et al. (2012) também utilizam SVM com *kernel* RBF para estudar associações de SNPs com as doenças de Parkinson (baixa hereditariedade 38%) e diabetes tipo 1 (alta hereditariedade 90%), encontrando áreas sob a curva ROC de $0,88$ para diabetes tipo 1 e $0,56$ para doença de Parkinson. Isso mostra que a inclusão de SNPs com frequências entre 1 e 5% e menores que 1%, conforme utilizados no trabalho, podem melhorar a predição do risco de doenças complexas.

No trabalho realizado por Uhmn et al. (2009) usou-se várias metodologias de Aprendizado de Máquina em um estudo de caso-controle baseado em SNPs para prever a suscetibilidade a hepatite crônica, encontrando máxima acurácia entre 67% e 73% dependendo do método usado. Esses métodos foram integrados com vários algoritmos de seleção de atributos para identificar um conjunto de SNPs relevantes para a doença.

Florestas aleatórias (do inglês, *Random Forests* - RF) foram utilizadas para encontrar variantes associadas em quatro genes para esclerose múltipla em Goldstein et al. (2010). A seleção dos SNPs foi baseada no *rank* construído pela importância dos SNPs, calculada

pelo próprio algoritmo de indução do aprendizado da RF.

Diversas técnicas de Aprendizado de Máquina foram empregadas para descobrir associações de SNPs com doenças em dados reais e simulados em Szymczak et al. (2009). Dentre elas pode-se citar: Regressão de Cumeeira (do inglês, *Ridge Regression*), RF, Lasso, Lasso com grupo, Lasso Bayesiano (do inglês, *Bayesian* Lasso), Redes Neurais e SVM. As abordagens foram usadas para diferentes objetivos, como, por exemplo, predição, efeitos principais, interações gene-gene, imputação de genótipo, relação causal, ou até mesmo uma mistura destes objetivos. Os métodos de Aprendizado de Máquina empregados demonstraram vantagens em relação às técnicas estatísticas tradicionais, embora notou-se que a aplicação de técnicas de seleção de variáveis específicas para os dados GWAS são necessários (SZYMCZAK et al., 2009).

Inúmeras vantagens foram encontradas em usar técnicas de Aprendizado de Máquina em detrimento de técnicas estatísticas tradicionais, porém, notou-se que métodos de seleção de atributos específicos são necessários. Wasan et al. (2012) fizeram uma revisão dos métodos mais comuns utilizados no risco de arritmias cardíacas hereditárias tais como *odds ratio*, *hazard ratio*, teste qui-quadrado e regressão logística avaliando seus benefícios e suas desvantagens, além de discutir outros métodos menos tradicionais como árvores de decisão, redes neuronais, SVM e classificadores bayesianos. O autor comenta que tais técnicas realizam seleção de variáveis explicativas, e em alguns casos, demonstram melhores resultados que os métodos clássicos.

Üstünkar et al. (2012) aplicaram uma meta-heurística, denominada *Simulated Anneling* (SA), para a seleção de marcadores SNP com o objetivo de encontrar o menor subconjunto de SNPs informativos para os quais o erro de predição para classificação é minimizado. A função objetivo do algoritmo SA foi a acurácia do classificador *Naive Bayes* baseada numa validação cruzada com 5-*fold* e os fenótipos binários avaliados foram baseados em estudos de caso-controle para as doenças de Alzheimer e artrite reumatóide. Os conjuntos iniciais de SNPs possuíam aproximadamente 550.000 SNPs e, após a primeira seleção baseada em SNPs relevantes biologicamente corrigidos pelo valor-p, cerca de 25.000 SNPs foram selecionados. Esse subconjunto de SNPs foi submetido ao processo de busca SA, determinando ao final do processo 1.300 SNPs aproximadamente relacionados aos dois fenótipos adotados.

1.4 Trabalhos Correlatos para Previsão de Fenótipo

Apesar do presente trabalho estar relacionado com estudos de associação em escala genônima, é importante mostrar que técnicas de Aprendizado de Máquina tais como SVR e RF vem sendo cada vez mais empregadas na previsão de fenótipos contínuos. Os trabalhos avaliados na presente seção se enquadram em seleção genômica e na predição de dosagem de medicamentos para humanos baseada em informações de SNPs.

Morser et al. (2009) utilizaram cinco métodos de regressão para predizer o valor genético de 1.945 touros leiteiros para a porcentagem de proteína e o índice de seleção australiana, ambos fenótipos contínuos, e estimar os efeitos de 7.372 SNPs: regressão fixa usando mínimos quadrados (do inglês, *Fixed Regression using Least Squares* - FR-LS), regressão aleatória BLUP (do inglês, *Random Regression Best Linear Unbiased Prediction* RR-BLUP), regressão Bayesiana (do inglês, *Bayesian Regression* - Bayes-R), por mínimos quadrados parciais de regressão (do inglês, *Partial Least Squares Regression* - PLSR); e máquinas de vetor suporte com regressão (SVR). Os quatro métodos avaliados: RR-BLUP, Bayes-R, PLSR e SVR geraram precisões similares para previsão, e seu uso imediato na seleção de gerações futuras em bovinos leiteiros foi comparável. Os autores também concluíram que a utilização de FR-LS na seleção genômica não é recomendada.

O estudo de Gianola et al. (2011) utilizou Redes Neurais Bayesianas lineares e não-lineares para avaliar o poder preditivo em relação à produção de leite, ao níveis de proteína e de gordura no leite de vacas Jersey. Os pesquisadores encontraram que as correlações de Pearson obtidas pela Redes Neurais linear e não-linear foram, respectivamente, 0,48 e 0,59, indicando que modelos não-lineares podem melhorar a predição de traços quantitativos. Concluem ainda que em uma Rede Neural, quanto maior o número de neurônios nas suas camadas ocultas, muito maior deverá ser sua amostra para possibilitar a extração de padrões durante o aprendizado. O processo de regularização bayesiano é necessário em Redes Neurais aplicadas em GWS, pois o número de variáveis (SNPs) é, geralmente, maior que o número de instâncias disponíveis (neste caso, as vacas Jersey).

O trabalho de Cosgun e Duarte (2011), além de aplicar o SVR, utiliza, também, *Random Forest Regression* e *Boosted Regression Tree* para a previsão da dose de manutenção de varfarina em um estudo de coorte[16] em afro-americanos. Todos os três

[16]São estudos observacionais onde os indivíduos são classificados (ou selecionados) segundo o *status* de exposição, sendo seguidos para avaliar a incidência de doença.

métodos apresentaram melhores resultados do que estudos anteriores nas mesmas bases de dados, sendo que a técnica *Random Forest Regression* demonstrou o melhor desempenho.

1.5 Objetivos

1.5.1 Objetivo Geral

O objetivo principal deste trabalho é propor um método para selecionar um subconjunto de marcadores moleculares do tipo SNP, a partir de um conjunto de dados inicial formado por milhares de marcadores e por um fenótipo de interesse, que contenha a maioria dos SNPs mais informativos biologicamente em relação à característica fenotípica. A novidade do método proposto está em dividir o processo de seleção em três fases: relevância, corte e refinamento. A fase de relevância realiza o ordenamento dos marcadores feito pelo *rank* da *Random Forest*. A etapa de corte, que é a primeira seleção, é composta pela avaliação dos subconjuntos de SNPs gerados a partir do ordenamento da fase de relevância e é baseada no erro de predição do SVM (classificação) ou do SVR (regressão) na validação cruzada com 10-*folds*. Finalmente, a fase de refinamento realiza a segunda seleção de SNPs pela busca da melhor combinação de marcadores pelo Algoritmo Genético com base nos SNPs selecionados na primeira seleção. Deste modo, o método proposto atribui técnicas de Inteligência Computacional robustas em relação aos métodos estatísticos mais comumente usados em GWAS, para cada etapa de seleção, objetivando a minimização da perda de SNPs verdadeiros-positivos e os pontos fracos de cada abordagem quando aplicadas separadamente e, concomitantemente, maximizando o poder de explicação dos marcadores selecionados ao final de todo processo.

A abordagem proposta foi construída para trabalhar com fenótipos que possuem tanto arquiteturas genéticas simples, como as características mendelianas, que são explicadas por genes alelos[17] em um único locus; quanto as complexas, que podem ser influenciadas por múltiplos genes alelos em *loci* distintos no genoma. Além disso, este método deve ser capaz de trabalhar em contextos que possuem SNPs com somente efeitos aditivos, ou somente não-aditivos (com possíveis interações entre SNPs), ou com ambos, onde não se tem especificada a quantidade de SNPs em interação. Adicionalmente, tal método deve

[17] Alelos são as formas alternativas de um gene (PIERCE, 2010). Por exemplo, um gene para a cor de pelagem em gatos pode existir em um alelo que codifica pelagem preta ou um alelo que codifica pelagem laranja (PIERCE, 2010).

ser flexível para tratar tanto com fenótipos contínuos (problema de regressão), quanto com binários (problemas de classificação).

O objetivo de selecionar o subconjunto de marcadores do tipo SNP mais importantes para a explicação do fenótipo, vem de encontro à possibilidade de melhorias na detecção do risco de desenvolvimento de doenças em humanos, animais e vegetais; na identificação de animais e vegetais que tenham melhor desempenho em características de cunho sócio-econômico; produção de *chips* de genotipagem personalizados de baixa densidade para um fenótipo específico. Com isso, o desenvolvimento de métodos mais eficientes em GWAS podem gerar melhorias em diversos segmentos da sociedade.

1.5.2 Objetivos Específicos

Os objetivos específicos desse estudo são:

- comparar o método proposto com as abordagens mais comuns usadas em GWAS para selecionar marcadores moleculares do tipo SNP, que geralmente são baseadas na escolha dos marcadores com o menor valor-p baseado na regressão linear simples, no caso de fenótipo contínuo, ou com o menor valor-p em testes de associação (teste do Qui-Quadrado) no caso de fenótipos binários;

- comparar a seleção do método proposto com a seleção feita pelo *Bayesian Lasso* (BLASSO), método que é baseado na variância explicada por cada marcador em relação à variância fenotípica;

- testar a abordagem sugerida em conjunto de dados sintéticos, gerados a partir de simuladores distintos para avaliar as possibilidades de aplicação do método em cenários de complexidades distintas.

- mostrar que o problema do crime de inversão [18] não é cometido nesse trabalho, pois o método para simular os conjuntos de dados é distinto do usado para selecionar os marcadores SNP;

- avaliar o método sugerido em um conjunto de dados reais fornecida pela Embrapa Gado de Leite, com as possíveis implicações em relação aos SNPs selecionados;

[18] Crime de inversão ocorre quando os mesmos (ou quase os mesmos) ingredientes teóricos são empregues para sintetizar, bem como para inverter dados em um problema inverso (WIRGIN, 2004). O problema inverso pode ser definido como dado um conjunto de entrada e saída, determinar o conjunto de parâmetros que estão de acordo com a relação entre a entrada e a saída.

- verificar a efetividade do método no conjunto de dados reais por meio da proximidade dos SNPs selecionados com QTLs ou genes analisados em bases de dados específicas e em estudos anteriores, que estão relacionados com a produção ou a composição do leite, ou até mesmo com a ocorrência de mastite.

1.6 Estrutura do Texto

Este estudo baseou-se em trabalhos no estado da arte em GWAS. Além disso, não foram encontrados outros trabalhos com esta abordagem específica perante este tema.

O presente capítulo objetivou caracterizar o problema e mostrar a importância do desenvolvimento de soluções adequadas para o mesmo. Além do mais, melhorias em outras áreas correlacionadas com GWAS são comentadas para mostrar as possibilidades de uso do método apresentado.

O Capítulo 2 apresenta os conceitos biológicos usados em GWAS para fornecer o conhecimento mínimo do problema tratado neste estudo. Os SNPs são conceituados e os possíveis desdobramentos de sua ocorrência são discutidos. As características quantitativas são definidas e exemplicadas no intuito de mostrar a complexidade envolvida na sua determinação. Os conceitos de herdabilidade e QTL são apresentados e discutidos para o entendimento da caracterização de fenótipos a partir da influência do genótipo. Finalmente, os possíveis tipos de ações gênicas são definidos e ilustrados para indicar que os modelos de seleção de SNPs devem ser flexíveis para identificar corretamente os sinais gerados pelos marcadores informativos.

O Capítulo 3 descreve as etapas de GWAS e os filtros usados para o controle de qualidade dos dados genômicos. O desequilíbrio de ligação é definido e algumas métricas usadas para medi-lo são apresentadas, bem como suas propriedades envolvendo vantagens e desvantagens. Os blocos LD e os *tag* SNPs são mostrados e sua conexão com métodos de seleção de SNPs é discutida.

O Capítulo 4 introduz Técnicas de Inteligência computacional usadas neste estudo tais como Árvores de Decisão e Regressão, Métodos *Ensembles*, *Random Forests*, *Support Vector Machine*, *Support Vector Regression* e Algoritmos Genéticos. Além disso, métricas de referência utilizadas como validação cruzada e área abaixo da curva ROC são apresentadas. As vantagens e desvantagens dessas abordagens também são discutidas.

O Capítulo 5 apresenta os métodos de seleção mais usados em GWAS como o valor-p da regressão linear ou o valor-p do teste de associação Qui-Quadrado, os quais são classificados como paramétricos. O método paramétrico de seleção baseado no Lasso Bayesiano (Blasso) também é discutido. Trabalhos correlatos que usaram métodos de seleção de atributos não-paramétricos de inteligência computacional em GWAS também são apresentados.

O Capítulo 6 apresenta o método proposto em suas duas versões. O algoritmo completo é apresentado tanto para fenótipos contínuos (regressão) quanto para binários (classificação). Todas as etapas são apresentadas detalhadamente, além das justificativas para a construção das mesmas. As vantagens e desvantagens do método são discutidas ao final com objetivo de determinar o real potencial e possíveis limitações do método.

O Capítulo 7 aborda a descrição dos dados experimentais. O processo de geração dos dados simulados é apresentado detalhadamente, além da descrição pormenorizada ser realizada para o conjunto de dados reais fornecido pela Embrapa.

O Capítulo 8 descreve como foram realizados os experimentos computacionais. Todos os parâmetros do método proposto são determinados e todos os passos são descritos para uma melhor compreensão da solução fornecida.

O Capítulo 9 apresenta as conclusões. Tentou-se relacionar cada conclusão com os objetivos geral e específicos apresentados no presente capítulo.

O Capítulo 10 trata dos trabalhos futuros. As possíveis melhorias do SMS são discutidas, além do uso do SMS em conjuntos de dados simulados e reais com diferentes complexidades.

Finalizando, as referências citadas são listadas. Como anexos, são apresentados os artigos publicados em periódicos internacionais e o capítulo de livro construído a partir do *workshop* promovido pela Embrapa. O registro e a descrição do método proposto também estão incluídos no anexo.

2 Conceitos Biológicos

Neste capítulo, serão apresentados os polimorfismos de base única bem como as consequências de sua ocorrência. O conceito de característica quantitativa será explorado para mostrar a complexidade envolvida em estudos de associação em escala genômica. A herdabilidade será apresentada com o objetivo de quantificar os vários tipos de características quantitativas em relação às variações genéticas e ambientais. As regiões no genoma que contém genes denominadas QTL também serão apresentadas. As possíveis ações gênicas existentes sobre o fenótipo serão apresentadas e exemplificadas.

2.1 Polimorfismos de base única

A Figura 2.1 mostra dois segmentos de DNA do mesmo gene em dois cromossomos homólogos (um cromossomo do pai e outro da mãe) não necessariamente idênticos. As bases nitrogenadas complementares GC e AT representam duas formas distintas para o mesmo gene denominadas alelos **A** e **a** respectivamente. Essa variante **Aa** será um SNP (pronuncia-se "snip") caso apresente frequência maior ou igual a 1% na população, caso contrário, essa variante é considerada uma mutação (BROOKES, 1999). Assim, uma forma de representar essa informação é dizer que esse indivíduo possui o genótipo **Aa** para esse *locus* no gene em questão.

A Figura 2.2 exemplifica uma amostra com três indivíduos e as sequências de bases nitrogenadas de uma única fita do DNA de pares de cromossomos homólogos. Como exemplo, pela complementariedade entre as bases nitrogenadas, sabe-se que A (adenina) se liga a T (timina) ou vice-versa, e que C (citosina) se liga a G (guanina), logo, basta ter a informação de uma fita de cada cromossomo. Na Figura 2.2, para o SNP1, o indivíduo 1 apresenta os genótipos TC (fica subentendido que T se liga a A e C se liga a G); para o indivíduo 2, os genótipos TT (fica subentendido que T se liga a A) e para o indivíduo 3, os genótipos CC (fica subentendido que C se liga a G).

Segundo Arbex (2009), se o polimorfismo estiver nas células reprodutoras de um indivíduo, então pelo menos um de seus descendentes podem herdar essa modificação e, após muitas gerações, o SNP pode, eventualmente, se estabelecer na população. Todavia,

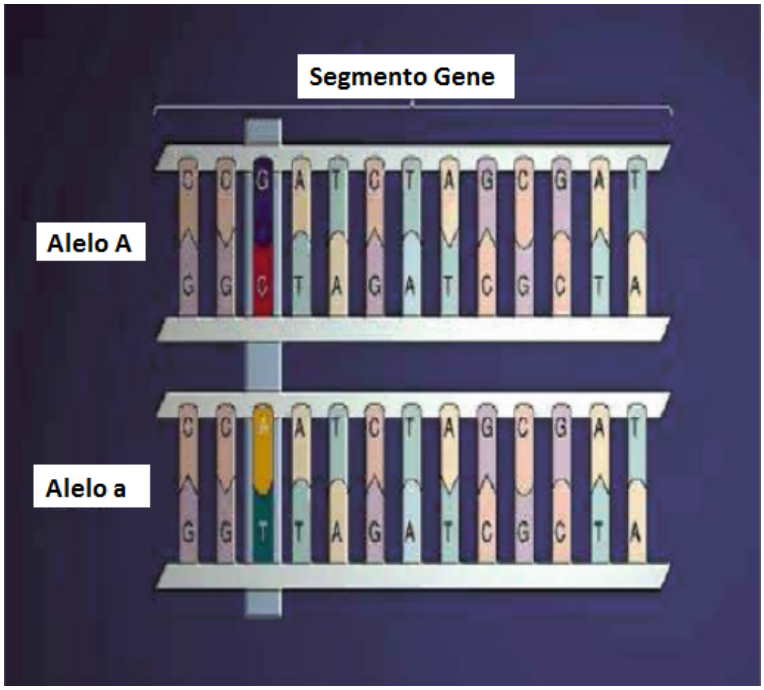


Figura 2.1 Ilustração de um SNP de um gene em um par de cromossomos autossômicos. O terceiro par de bases de cada cromossomo mostra a variação (GC ou AT). Adaptado de Laird e Lange (2011).

existem apenas dois alelos, o da sequência original e a versão polimórfica (ARBEX, 2009). A probabilidade do surgimento de um terceiro alelo na mesma posição e estabelecimento do mesmo na população a partir da reprodução de seus portadores é muito baixa (ARBEX, 2009). Por isso, os SNPs, algumas vezes, são chamados de "marcadores bi-alélico" (BROOKES, 1999).

De acordo com Ban et al. (2010), um SNP em uma sequência que codifica uma proteína pode induzir mudanças em aminoácidos, resultando em alterações funcionais da proteína. Alguns SNPs em uma região promotora[1] podem afetar a regulação transcricional, e um SNP em uma região de íntron pode afetar o *splicing*[2] ou a expressão de um gene (BAN et al., 2010). Por outro lado, SNPs que ocorrem fora de regiões codificantes podem não desempenhar um papel determinante na doença, mas mesmo SNPs que ocorrem em

[1]É uma sequência de DNA que o aparelho de transcrição reconhece e se liga (PIERCE, 2013). Ele indica qual dos dois filamentos de DNA deve ser lido como molde e a direção de transcrição (PIERCE, 2013).

[2]Neste processo (cujo nome significa "ato de cortar", em português), regiões específicas do RNA mensageiro (os íntrons) são recortadas e eliminadas. Os íntrons eliminados são segmentos não-codificantes, porque não levam nenhuma mensagem para produção de proteínas. Depois que eles são eliminados, os segmentos resultantes (os éxons) unem-se entre si, formando a molécula de RNA mensageiro funcional, com a mensagem madura, ou a mensagem propriamente dita (ALBERTS et al., 2010). Este processo é importante pois somente após ter passado por ele é que o RNA mensageiro se torna ativo na codificação da mensagem que levará à produção de uma proteína específica (ALBERTS et al., 2010).

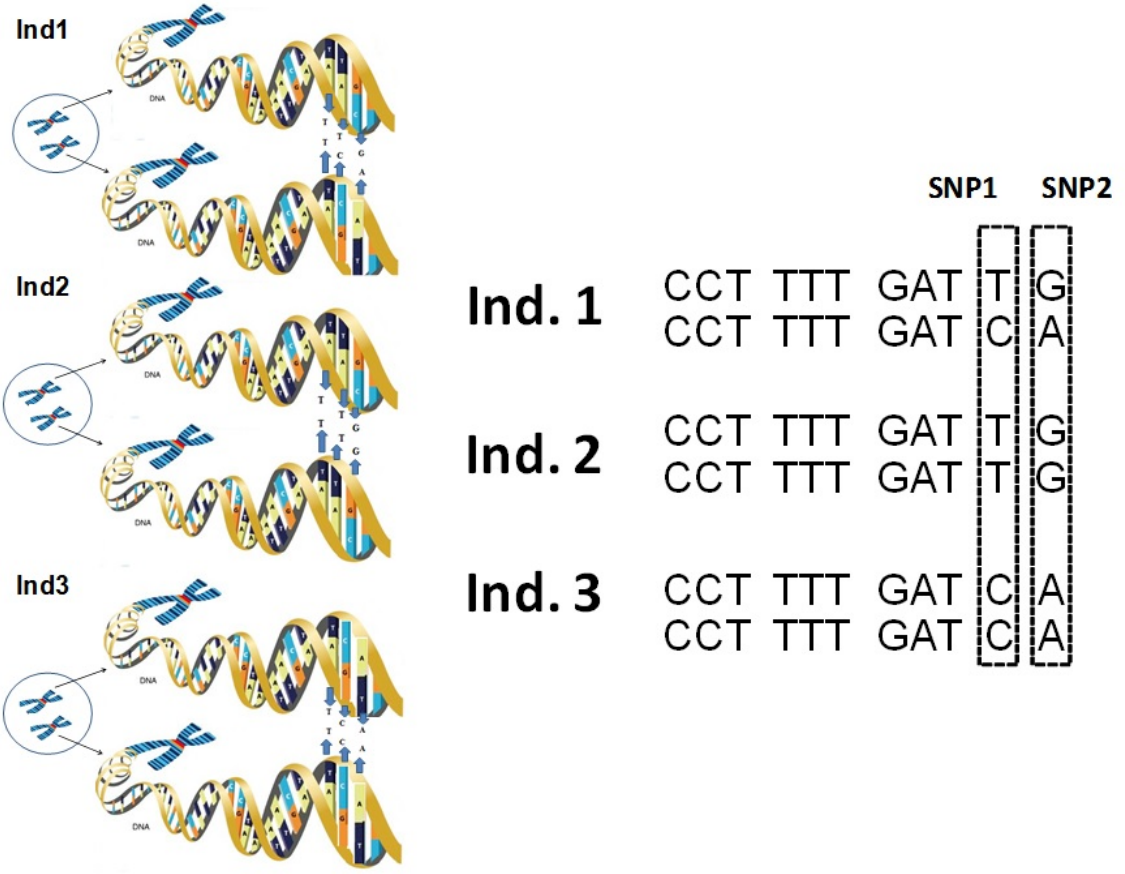


Figura 2.2 Exemplo de dois SNPs em uma amostra de três indivíduos.
Adaptado de Silva (2013).

uma região codificante podem não ter qualquer efeito biológico por causa da flexibilidade existente na codificação de sequências (LAIRD; LANGE, 2011). Como mostrado na Figura 2.3, os aminoácidos são codificados por códons[3] que são três pares de sequências de base. A maioria dos códons tem uma relação de muitos-para-um em relação ao número de aminoácidos, isto é, várias sequências básicas de três pares podem codificar o mesmo aminoácido (LAIRD; LANGE, 2011). Por exemplo, se a terceira base do códon no TCT para a serina é alterado para qualquer uma das outras três bases, por exemplo, TCA, ainda será codificada uma serina. Tais variantes são denominadas de SNP silencioso ou sinônimo porque não causam mudança alguma em seu produto, mas são úteis como marcadores genéticos, pois podem sinalizar regiões que contenham a verdadeira mutação causal (LAIRD; LANGE, 2011).

Como exemplo de SNPs não-sinônimos, a anemia falciforme é causada por um única

[3]É uma sequência de três bases nitrogenadas do RNA mensageiro que codificam um determinado aminoácido ou que indicam o ponto de início ou fim de tradução da cadeia de RNA mensageiro (PIERCE, 2013).

mudança no par de bases no gene da hemoglobina no cromossomo 11 em humanos (LAIRD; LANGE, 2011). A Figura 2.3 mostra a sequência codificada do gene normal da hemoglobina (A) e do gene da hemoglobina falciforme (S). As duas sequências diferem por uma mudança de uma forma regular na sequência que codifica a hemoglobina glutamina. A base nitrogenada **A** na sequência GAG (em vermelho) é trocada por **T** na sequência GTG (em vermelho), gerando uma variante (SNP). O alelo falciforme (chamado S para doente) altera a seqüência, por isso, que a valina (Val) é codificada em vez de glutamina (Glu). Indivíduos com SS desenvolvem a anemia falciforme, enquanto que indivíduos com AS e AA não são afetados pela doença.

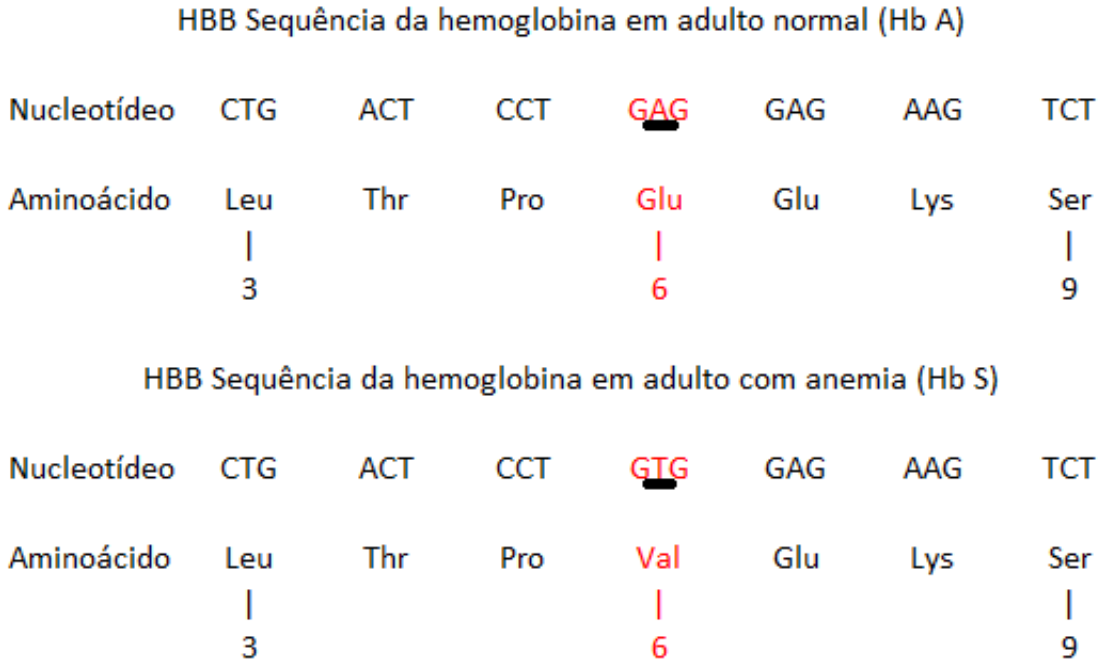


Figura 2.3 Uma variante no gene da hemoglobina causando anemia falciforme. Adaptado de Laird e Lange (2011).

Uma das vantagens do uso de SNPs na predição de fenótipos em seleção genômica é não necessitar da realização de análise de ligação, uma vez que devido a grande saturação do genoma com marcadores SNPs, assume-se que estes estão diretamente em LD com o QTL (SILVA, 2013). Além disso, as posições de todos os marcadores são conhecidas, pois são obtidas diretamente do processo de genotipagem (SILVA, 2013). Isto permite "extrapolar"os resultados (efeitos de marcadores) obtidos de um grupo de indivíduos para outro grupo de indivíduos relacionados de alguma forma com o primeiro, por exemplo, entre gerações (SILVA, 2013). Assim, diferentemente da análise clássica de QTL

via microssatélites[4], a qual apresenta utilidade apenas dentro da população estudada, pois QTLs identificados em uma família F2 só podem ser usados nesta população, as informações genéticas sobre SNPs podem ser exploradas em outras populações (SILVA, 2013).

Outra vantagem dos SNPs é que características quantitativas são governadas por um grande número de genes com pequenos efeitos (natureza poligênica), os SNPs contemplam esta hipótese, pois tem-se um grande número de marcadores teoricamente com pequenos efeitos (SILVA, 2013). Com isso, não é necessário assumir, como é feito nas metodologias baseadas em microssatélites, que os poucos QTLs identificados explicam grande parte da variação da característica (SILVA, 2013). Portanto, do exposto anteriormente, os SNPs são elementos essenciais na identificação de variantes causais para doenças ou características de interesse econômico em GWAS ou para predição de fenótipos em Seleção Genômica.

2.2 Características Quantitativas

Uma característica quantitativa varia continuamente em uma população e são determinadas por múltiplos genes e fatores ambientais, ou seja, são características multifatoriais (PIERCE, 2013). Para contrapor a ideia de característica quantitativa, Mendel estudou a altura das plantas de ervilha, que pode ser descrita medindo-se o tamanho do caule da planta (PIERCE, 2013). Porém, as plantas que Mendel analisou exibiam apenas dois fenótipos diferentes (algumas eram altas e outras baixas), e essas diferenças eram determinadas por alelos em um único *locus* (PIERCE, 2013). Portanto, a natureza dessa característica era descontínua e mendeliana (não quantitativa), pois possuía somente duas classes para altura, além de ser determinada pela variação em somente um *locus*.

Algumas características não são contínuas, mas são determinadas por muitos fatores genéticos e ambientais, como as características merísticas, que são medidas por números inteiros. Um exemplo é o número de filhotes em uma ninhada de camundongos, pois a mesma pode ter 4, 5 ou 6 filhotes, mas a determinação subjacente da característica ainda pode ser quantitativa (PIERCE, 2013). Outro tipo de característica quantitativa

[4]Os microssatélites são curtas sequências de DNA que existem em múltiplas cópias repetidas em tandem (PIERCE, 2013).

descrito por Pierce (2013) é a característica com limiar, a qual indica presença ou ausência no indivíduo considerado. Como exemplo, a presença de algumas doenças pode ser considerada característica com limiar tais como diabetes tipos 1 e 2, câncer de mama, doença de Crohn, Alzheimer entre outras. Embora as características com limiar demonstrem apenas dois fenótipos, elas são consideradas quantitativas porque são determinadas por múltiplos fatores ambientais e genéticos (PIERCE, 2013).

Segundo Pierce (2013), a expressão da característica de uma suscetibilidade subjacente, denominada de propensão ou risco, varia continuamente, logo, somente quando o risco é maior que o limiar, uma característica específica é expressa. A Figura 2.4 exemplifica essa situação para uma doença, onde o indivíduo está doente se a sua suscetibilidade é maior que o limiar, e normal caso contrário. A maioria dos trabalhos em GWAS buscam identificar SNPs associados às características quantitativas.

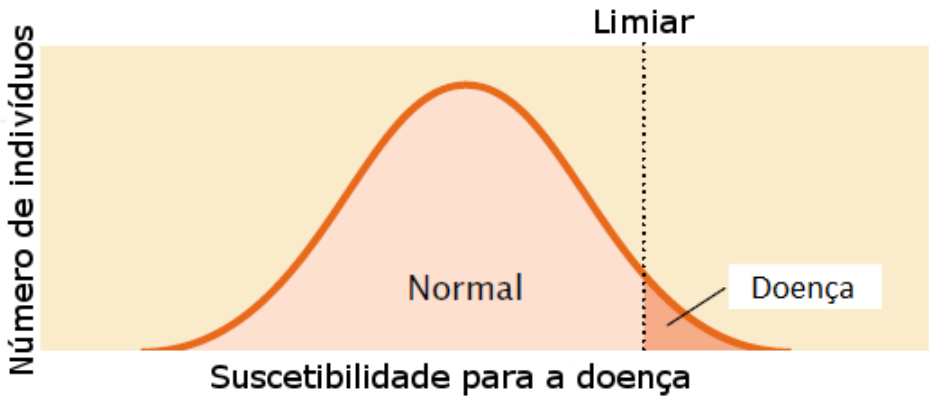


Figura 2.4 Características com limiar somente para dois possíveis fenótipos.

A característica está presente ou ausente - mas elas são quantitativas porque a suscetibilidade subjacente varia continuamente. Quando a suscetibilidade excede o limiar, a característica é expressa (PIERCE, 2013).

2.3 Herdabilidade

As características quantitativas são poligênicas, mas também são influenciadas por fatores ambientais (PIERCE, 2013). Para quantificar a influência do genótipo e do ambiente é útil saber o quanto da variação de uma característica é devido às diferenças genéticas e o quanto é referente às diferenças ambientais.

Uma forma de obter a variação de uma característica, medida por uma variável contínua, é obter uma amostra representativa de indivíduos e medir o fenótipo dos

mesmos e, então, calcular a média e a variância fenotípicas. A variância do fenótipo é designada por V_P. Algumas das diferenças entre fenótipos são devido às diferenças ambientais, simbolizada por V_E, e às diferenças no genótipo, denominada V_G. Uma terceira fonte de variação surge quando o efeito de um gene depende do ambiente específico no qual é encontrado (PIERCE, 2013). Essa variância devido à interação gene-ambiente é representada por V_{GE}. Logo, a Expressão 2.1 mostra a relação entre V_P, V_E, V_G e V_{GE}.

$$V_P = V_G + V_E + V_{GE} \tag{2.1}$$

Outros componentes da variância genética podem ser desdobrados em variância genética aditiva (V_A), variância genética de dominância (V_D) e variância de interação gênica (V_I). Com isso, a Expressão 2.2 explicita a variância fenotípica em função das variâncias referentes aos efeitos aditivos, dominantes e de interação entre genes, além do fator ambiental e da interação gene-ambiente.

$$V_P = V_A + V_D + V_I + V_E + V_{GE} \tag{2.2}$$

A proporção da variação fenotípica total que é devida às diferenças genéticas é conhecida como herdabilidade (PIERCE, 2013). Existem dois tipos de herdabilidade: a herdabilidade em sentido amplo (H^2) e a herdabilidade no sentido estrito (h^2). A H^2 avalia o quanto da variância fenotípica é devido à variância do genótipo, incluindo as variâncias aditivas, de dominância e de interação. Enquanto que h^2 detecta a proporção da variância do fenótipo resultante da variância genética aditiva. A variância genética aditiva determina primariamente a semelhança entre genitores e prole (PIERCE, 2013). Como exemplo, considere o caso extremo em que toda variância fenotípica é devido à variância genética aditiva, então os fenótipos da prole serão exatamente intermediários aos dos genitores; mas, se alguns genes são dominantes, então a prole pode ser fenotipicamente diferente de ambos os genitores (por exemplo, $Aa \times Aa$ pode gerar uma prole aa) (PIERCE, 2013). As Expressões 2.3 e 2.4 mostram as fórmulas para o cálculo de H^2 e h^2.

$$H^2 = \frac{V_G}{V_P} \tag{2.3}$$

$$h^2 = \frac{V_A}{V_P} \tag{2.4}$$

O conhecimento da herdabilidade h^2 possui grande valor prático, pois ele permite prever estatisticamente os fenótipos da prole baseado no fenótipo dos genitores (PIERCE, 2013). Entretanto, conforme Pierce (2013), a herdabilidade h^2 possui limitações, que são resumidas nos itens a seguir:

1. a herdabilidade não indica o grau em que uma característica é geneticamente determinada;
2. um indivíduo não tem herdabilidade, ou seja, é uma medida populacional;
3. não existe herdabilidade universal para uma característica, isto é, cada população possui sua herdabilidade;
4. mesmo quando a herdabilidade é alta, os fatores ambientais podem influenciar a característica;
5. as herdabilidades não indicam nada sobre a natureza das diferenças populacionais em uma característica.

2.4 *Quantitative Trait Loci* - QTL

Segundo Silva (2002), o modelo clássico de herança, fundamentado na aditividade dos genes, mostrou ser relativamente robusto, pois bons resultados foram obtidos com seu uso. Todavia, em cenários onde os efeitos maiores sobre o fenótipo são devidos a um ou a poucos pares de genes, este modelo não tem apresentado a mesma eficiência (SILVA, 2002).

Os genes de efeito maior sobre um fenótipo são chamados de *quantitative trait loci* (*locus* da característica quantitativa - QTL) (FALCONER; MACKAY; FRANKHAM, 1996). Muitas das vezes, esses genes não podem ser localizados individualmente, mas, é possível identificar a região do genoma onde estes *loci* podem estar presentes e estimar a fração da variação total do fenótipo referente a eles (SILVA, 2002).

O princípio da identificação de QTL ligados a marcadores, por meio do uso de modelos fixos, é baseado na genotipagem de indivíduos para o *locus* do marcador e na medição de seus fenótipos para a característica de interesse (SILVA, 2002). Conforme Martinez (1998), a maioria dos estudos realizados para detecção de QTLs basearam-se

em cruzamentos planejados, apesar de algumas análises serem feitas com populações de animais já existentes, como por exemplo, gado de leite.

2.5 Ação Gênica Aditiva

Cada par de gene possui efeito próprio e independente dos outros que se encontram presentes no genótipo do indivíduo (VALENTE et al., 2001). Assim, a ação total do genótipo sobre o fenótipo será igual a soma dos efeitos de cada par de gene e a simples substituição de um alelo por outro em um gene impacta o resultado total dos efeitos gênicos (VALENTE et al., 2001). Sob essa hipótese, a relação entre o genótipo e o fenótipo é linear. Considere os dois alelos A e a de um gene que age de forma aditiva sobre um determinado fenótipo. Dessa forma, o efeito de substituição do alelo a pelo A aumenta o fenótipo em 5 unidades conforme mostra a Figura 2.5, adotando-se a codifição $aa = 0$ (ausência do alelo A), $Aa = 1$ (presença de um alelo A) e $AA = 2$ (presença de dois alelos A).

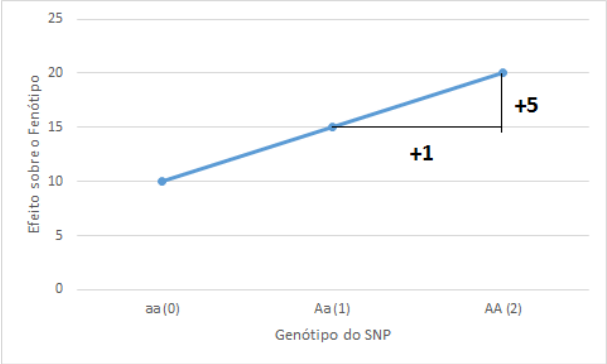


Figura 2.5 Exemplo do efeito aditivo de um gene (*locus*) sobre o fenótipo.

Considere que um indivíduo possua três genes que apresentam um efeito de 2 gramas (g) para ganho de peso para os alelos A, B e C; e para os alelos a, b e c , apenas 1 grama. Assim, um indivíduo cujo genótipo seja ABC terá fenótipo igual a $2g + 2g + 2g = 6g$. Para outro indivíduo com genótipo Abc, o fenótipo será $2g + 1g + 1g = 4g$. Além dos efeitos aditivos de alguns genes, existem genes que exercem ações gênicas não-aditivas sobre o fenótipo.

2.6 Ação Gênica Não-aditiva

2.6.1 Interação no mesmo locus (intralocus)

Dominância representa a situação que o fenótipo do heterozigoto é o mesmo que o de um homozigoto (PIERCE, 2013). Como exemplo, considere que A^1A^1 codifica flores vermelhas e que A^2A^2 codifica flores brancas. Se o heterozigoto A^1A^2 é vermelho, o alelo A^1 é dominante em relação ao alelo A^2. Se o heterozigoto A^1A^2 é branco, o alelo A^2 é dominante em relação ao alelo A^1. Naturalmente, essa situação incorpora uma não-linearidade na relação genótipo-fenótipo no mesmo *locus*.

Dominância incompleta é a situação que o fenótipo do heterozigoto é intermediário (está dentro de um intervalo) entre os fenótipos dos dois homozigotos (PIERCE, 2013). A Figura 2.6 exibe a situação que o fenótipo do heterozigoto A^1A^2 se situa entre os fenótipos dos homozigotos A^1A^1 (flor vermelha) e A^2A^2 (flor branca). Caso a cor seja quantificada em uma escala numérica contínua, o fenômeno da codominância incompleta pode gerar indivíduos com fenótipos apresentando um amplo espectro de variação, o que produz uma relação não-linear entre o genótipo e o fenótipo *intralocus*.

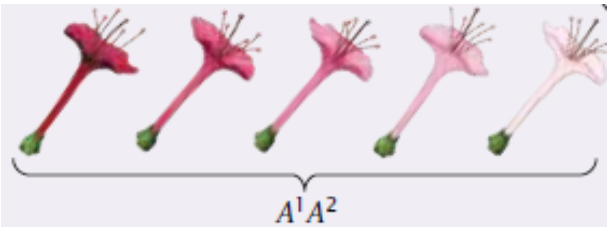


Figura 2.6 Exemplo de dominância incompleta.
Adaptado de Pierce (2013).

Codominância é quando o fenótipo do heterozigoto inclui os fenótipos de ambos homozigotos (PIERCE, 2013). Um exemplo de codominância ocorre nos tipos sanguíneos MN (PIERCE, 2013). No *locus MN*, existem dois alelos: L^M e L^N, os quais codificam os antígenos M e N respectivamente.

Sobredominância é uma forma de dominância em que o efeito do heterozigoto (Aa) é superior a qualquer dos dois homozigotos (AA e aa) (VALENTE et al., 2001). Por exemplo, para o fenótipo, aa tem efeito nulo, AA tem efeito 2 e Aa possui efeito 2,5.

2.6.2 Interação entre genes (interlocus)

Em geral, os genes exibem distribuição independente[5], mas não atuam independentemente em sua expressão fenotípica, ou seja, os efeitos dos genes em um *locus* dependem da presença de genes em outros *loci* (PIERCE, 2013). Esse tipo de interação entre genes em *loci* distintos (ou seja, genes não alelos) sobre uma caracterítica é denominado de interação gênica (PIERCE, 2013). Destaca-se que esse tipo de interação é distinto da interação entre os alelos de um gene como ocorre na dominância, pois este tipo de interação ocorre em um determinado *locus* de um gene.

Um exemplo de interação gênica é dados pela cor do pimentão *Capsicum annuum* que é determinada pela interação de dois *loci*, denotados por Y e C (PIERCE, 2013). As possibilidades de genótipo e fenótipo estão ilustradas na Tabela 2.1.

Tabela 2.1 Possibilidades de genótipos e fenótipos para a cor de pimentões a partir da interação gênica dos *loci* Y e C. Adaptado de Pierce (2013).

Genótipo[a]	Fenótipo (cor)
$Y\_\ C\_$	vermelho
$Y\_\ cc$	pêssego
$yyC\_$	laranja
$yycc$	creme

[a] O primeiro símbolo $\_$ representa Y ou y, e o segundo C ou c.

Uma das possibilidades para ocorrência de interação gênica é quando *loci* diferentes influenciam etapas diferentes em uma via bioquímica comum porque o produto de uma enzima afeta o substrato de outra enzima (PIERCE, 2013). A Figura 2.7 mostra que a cor nos pimentões de *Capsicum annuum* resulta de quantidades relativas de carotenoides vermelhos e amarelos. O *locus* Y codifica uma enzima (a primeira etapa na via da Figura 2.7) e o *locus* C, codifica a enzima da última etapa da via.

O termo epistasia foi cunhado por Bateson (1909), sendo um fenômeno de interação gênica, no qual o efeito de um gene é mascarado pelo efeito de outro gene (PIERCE, 2013). Essa situação é equivalente à dominância, porém ela ocorre em *loci* distintos. O gene que mascara é denominado de gene epistático e o gene cujo efeito é mascarado é um

[5] O princípio da segregação independente diz que os genes que codificam características distintas separam-se independentemente uns dos outros quando são formados gametas, devido à separação independente dos pares de cromossomos homólogos na meiose (PIERCE, 2013). Os genes que ficam próximos no mesmo cromossomo não se segregam independentemente (PIERCE, 2013).

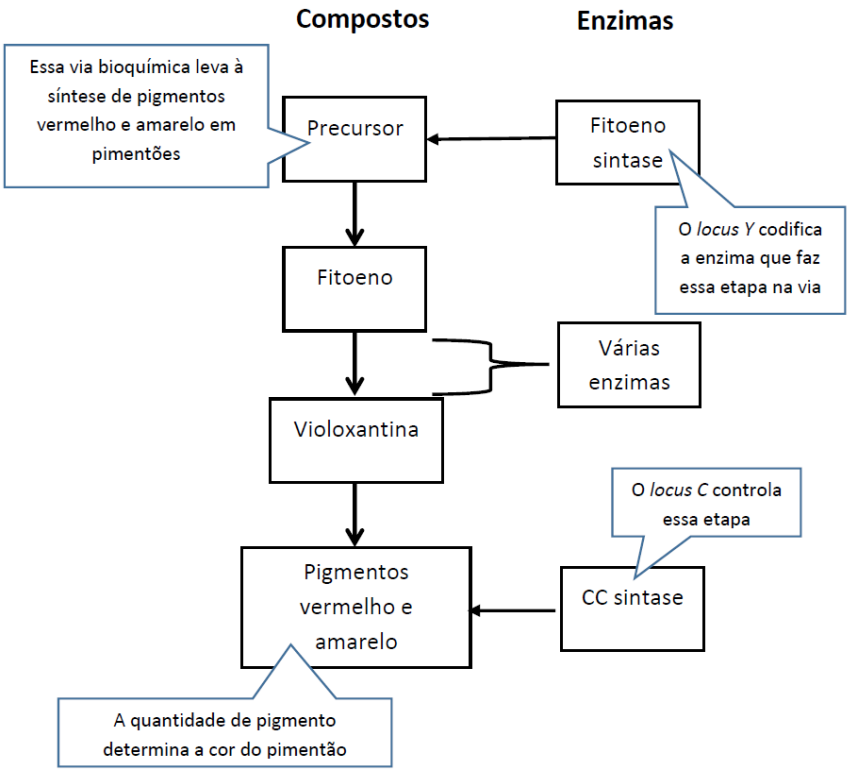


Figura 2.7 Uma via bioquímica de várias etapas sintetiza os pigmentos carotenoides responsáveis pela variação de cor em pimentões.
Adaptado de Pierce (2013).

gene hipostático (PIERCE, 2013). Segundo Pierce (2013), os genes epistáticos podem ser recessivos ou dominantes em seus efeitos.

Cordell (2002) argumenta que existem diversas situações de interações entre SNPs que o conceito de epistasia sugerido por Bateson (1909) não contemplou. Fisher (1919) propôs uma definição estatística para epistasia, a qual se refere a desvios da aditividade nos efeitos dos alelos em diferentes *loci* com respeito à contribuição para o fenótipo. Conforme Rothman, Greenland e Walker (1980) e Frankel e Schork (1996), a escolha da escala de medida para o fenótipo tem papel fundamental, pois fatores que são aditivos para uma escala, podem exibir interação para outra escala.

No caso de características quantitativas, a epistasia descreve qualquer interação entre dois ou mais *loci*, tais que o fenótipo de qualquer genótipo não pode ser predito simplesmente pela soma dos efeitos de *loci* individuais (CARLBORG; HALEY, 2004). Para outros exemplos de epistasia em características quantitativas descritas de forma contínua ver Carlborg e Haley (2004). Uma discussão aprofundada sobre epistasia e modelos que simulam epistasia pode ser encontrada em Moore e Williams (2015).

2.7 Resumo do Capítulo

Diversos tipos de características quantitativas são abordadas em GWAS dado seu perfil poligênico e multifatorial. Os efeitos gênicos podem ser aditivos ou não-aditivos promovendo uma complexidade a mais na busca pelos SNPs informativos. Outra dificuldade é selecionar SNPs associados às características quantitativas com baixa herdabilidade. O conjunto de obstáculos a serem superados pelos métodos de seleção de SNPs mostra o quanto esse problema é difícil e relevante. Consequentemente, qualquer melhoria em algum desses problemas mencionados, pode propiciar uma melhor compreensão do processo biológico envolvido entre o genótipo e o fenótipo.

3 Estudos de Associação em Escala Genômica

Este capítulo apresenta as etapas de um estudo de associação em escala genômica, as medidas usadas para o cálculo do desequilíbrio de ligação e os principais filtros usados nos dados referentes ao genótipo em GWAS. A correção de Bonferroni para múltiplos testes de hipóteses também será discutida bem como suas vantagens e desvantagens.

3.1 Etapas de GWAS

Existem diversos *workflows*[1] construídos para descrever as etapas de estudos de associação em escala genômica (GWAS), porém, no presente trabalho, serão considerados as fases propostas por Kingsmore et al. (2008) e Olazar (2013) para doenças em humanos, destacadas a seguir:

1. realização de um bom planejamento do estudo a ser realizado, onde seleciona-se um grande número de indivíduos com a doença (casos) e sem a doença (controles);

2. coleta do genótipo das amostras de casos e controles (amostras maiores que 1.000) a partir da genotipagem do DNA de cada indivíduo, considerando variáveis de confundimento como raça, etnia e sexo com o objetivo de estratificar a amostra a fim de maximizar os sinais dos SNPs informativos;

3. usar aproximadamente 1 milhão de SNPs aleatórios ou 25.000 SNPs não-sinônimos[2];

4. realizar controle de qualidade sobre os dados brutos gerados pela genotipagem, para verficação e correção de erros do processo de genotipagem;

5. derivação dos blocos haplótipos[3];

[1] Fluxo de trabalho (do inglês, *workflow*) é descrito por um grafo direcionado acíclico no qual cada tarefa computacional é representada por um nó e a dependência entre as tarefas de controle é representado por um arco dirigido entre os nós correspondentes (BALA; CHANA, 2011).

[2] SNPs que levam a uma alteração na sequência de aminoácidos que promove uma alteração na proteína codificada pelos genes e que, portanto, pode afetar sua função (KINGSMORE et al., 2008).

[3] Uma combinação de alelos em *loci* ligados (em uma única cromátide) que são transmitidos muitas mais vezes juntos do que aleatoriamente (KINGSMORE et al., 2008).

6. realizar múltiplos testes estatísticos χ^2 ou similares para verificar a associação entre os SNPs e a doença;

7. considerar valores-p menores que 10^{-7}, como indicado por Kingsmore et al. (2008), ou realizar algum tipo de correção como por exemplo, taxa de detecção de falsos SNPs ou correção de Bonferroni;

8. Mapeamento fino do sinal de associação com a genotipagem adicional de SNPs na região, mapeamento refinado de LD na região de associação, derivação empírica de haplótipos e exame do efeito da estratificação, se disponível;

9. Confirmação dos sinais de associação positiva a partir da replicação dos resultados em amostras independentes de uma população (amostras maiores que 1.000 indivíduos), realizando a genotipagem de SNPs candidatos nomeados (menos que 20 SNPs) e os testes estatísticos χ^2 ou similares;

10. validação biológica de associação pela identificação de variantes para o aumento de risco, exame da consequência funcional da variante e determinação do mecanismo de aumento do risco.

Além das etapas anteriores, segundo Olazar (2013) uma méta-análise pode ser realizada para incrementar o poder de detecção de variantes raras utilizando uma amostra maior que a do estudo original. Conforme Cantor, Lange e Sinsheimer (2010), meta-análise é uma abordagem estatística válida e bem estabelecida para combinar evidências entre qualquer número de estudos independentes, sendo que cada um deles é delineado para examinar a mesma hipótese de pesquisa. Ao invés de usar os dados originais a partir desses estudos, o que pode ser computacionalmente custoso e operacionalemente difícil, a meta-análise combina seus resultados (CANTOR; LANGE; SINSHEIMER, 2010).

Outro ponto a ser comentado é que para GWAS em animais, as amostras disponíveis nem sempre são maiores que 1.000, como em Wong e Bernardo (2008) que trabalharam com população de 50 indivíduos, indicando uma diminuição do poder estatístico de detecção dos SNPs informativos nesses estudos como discutido por Gondro, Werf e Hayes (2013). Outra dificuldade em animais é a disponibilidade de amostras para replicação usadas para comprovar os SNPs selecionados.

3.2 Desequilíbrio de Ligação

A hipótese de interesse em GWAS é a busca por genes candidatos que podem ser a causa de uma determinada característica fenotípica, que pode tanto ser uma doença quanto um traço benéfico, como por exemplo, a produção de leite em bovinos. Deste modo, o SNP considerado pode não ser a verdadeira causa da doença, mas o mesmo pode ser altamente associado com a variante funcional, que é a origem da doença. Para denominar essa associação não-aleatória entre o marcador e a mutação funcional, diz-se que os dois estão em desequilíbrio de ligação (do inglês, *Linkage Disequilibrium* - LD). Caso contrário, se o marcador tem pouca associação com a mutação funcional, diz-se que o SNP e a mutação estão em equilíbrio de ligação (do inglês, *Linkage Equilibrium* - LE).

Pela Figura 3.2, percebe-se essa associação pela proximidade entre os SNPs, pois quanto mais próximos fisicamente eles estão, maior a probabilidade deles serem herdados juntos após a recombinação gênica [4]. (RADDING, 1982). Dentro de uma família, a ligação ocorre quando dois marcadores genéticos (pontos em um cromossomo) permanecem ligados em um cromossomo em vez de ser quebrada por eventos de recombinação durante a meiose, identificados como linhas vermelhas (RADDING, 1982). Em uma população, longos segmentos cromossômicos fundadores da geração inicial são sequencialmente reduzido de tamanho por eventos de recombinação contíguos (RADDING, 1982). Ao longo do tempo, um par de marcadores (ou pontos) em um cromossomo movem do desequilíbrio de ligação para o equilíbrio de ligação, quando os eventos de recombinação eventualmente ocorrem entre cada ponto possível no cromossomo (RADDING, 1982).

O LD é relacionado ao conceito de ligação cromossômica, onde dois marcadores em um cromossomo permanecem fisicamente juntos através de gerações de uma família (BUSH; MOORE, 2012). Eventos de recombinação dentro de uma família de geração para geração quebram segmentos dos cromossomos (BUSH; MOORE, 2012). A Figura 3.1 indica a relação entre LD numa população e a ligação cromossômica em uma família. A Figura 3.2 explicita o SNP causal e o SNP genotipado em uma região de alto desequilíbrio de ligação, ou seja, grande parte da informação do SNP causal está no SNP genotipado.

O LD pode ser interpretado de maneira direta através dos SNPs 1 e 2 da Figura 3.3, porque sempre ocorrem os pares de alelos A no SNP1 e G no SNP2, e os alelos T no

[4]Recombinação é a distribuição de alelos em novas combinações. O *crossing over* é a base para a recombinação, criando novas combinações de alelos em uma cromátide como descrito em Pierce (2013).

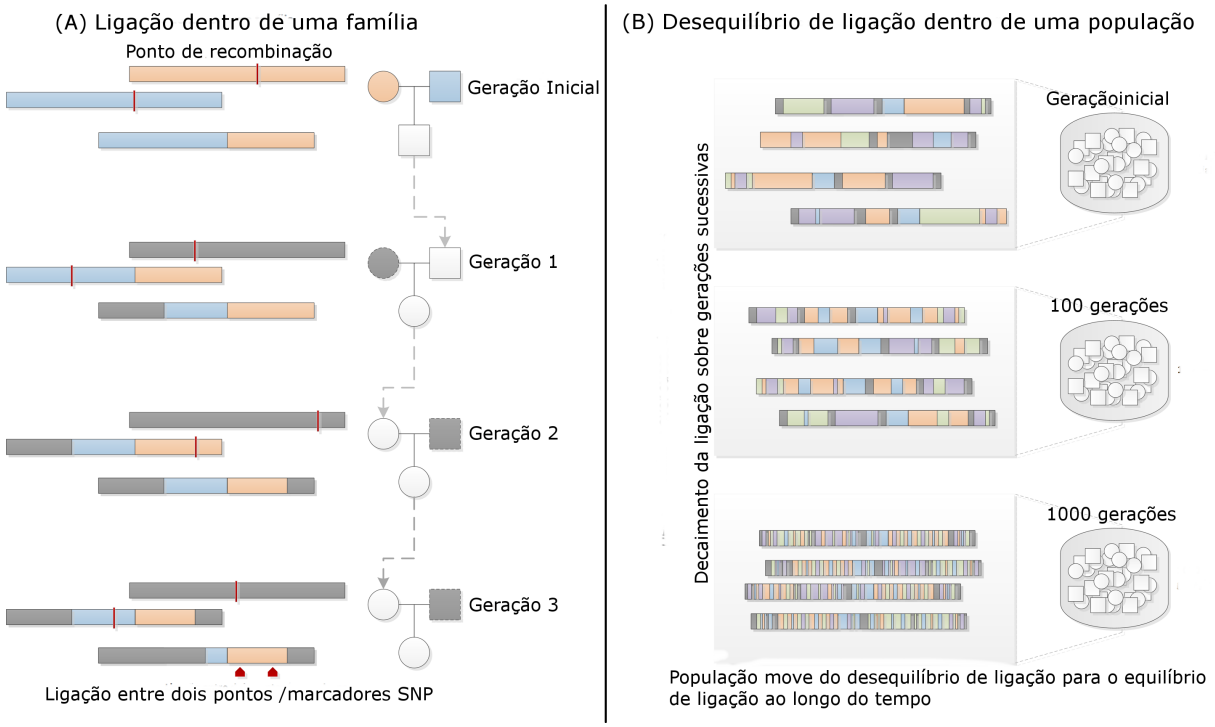


Figura 3.1 Relação entre a ligação cromossômica em uma família e o LD numa população ao longo de gerações.

(A) A ligação que ocorre entre dois marcadores dentro de uma família quando permanecem ligados em um cromossomo em vez de serem separados por eventos de recombinação durante a meiose, mostrados como linhas vermelhas. (B) Marcadores movem do desequilíbrio de ligação para o equilíbrio de ligação ao longo do tempo em uma população. Adaptado de Bush e Moore (2012).

SNP1 e A no SNP2, sendo que, neste caso, a associação entre os SNPs 1 e 2 é perfeita. A mesma análise pode ser feita para os pares SNP3 com SNP5 e SNP4 com SNP6.

Uma metodologia satisfatória para selecionar somente os marcadores informativos deveria excluir os SNPs correlacionados (em desequilíbrio de ligação), eliminando a redundância do conjunto inicial. Entretanto, como os níveis de correlação são distintos, podem existir SNPs correlacionados que marcam regiões distintas do genoma. Tal metodologia de seleção deve considerar essas situações para não eliminar SNPs informativos.

3.2.1 Medidas para LD: D' e r^2

Existem várias métricas para avaliar o LD, cada uma delas com suas vantagens e desvantagens. Para uma discussão mais aprofundada veja (DEVLIN B E RISCH, 1995), o Capítulo 8 de Thomas (2004) e o Capítulo 9 de Ziegler e Koenig (2007). As medidas

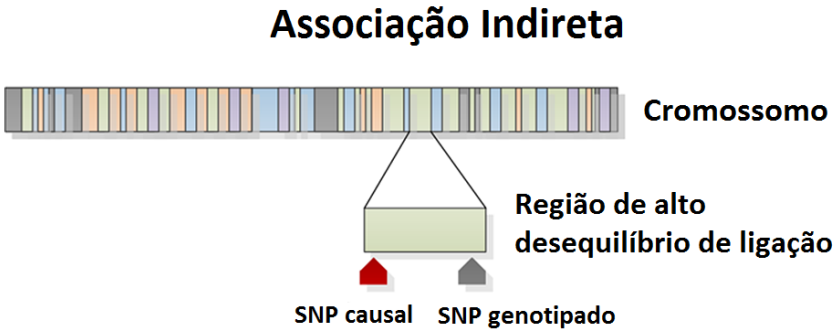


Figura 3.2 LD entre o SNP causal e o SNP genotipado em uma região de alto desequilíbrio de ligação.
Adaptado de Bush e Moore (2012).

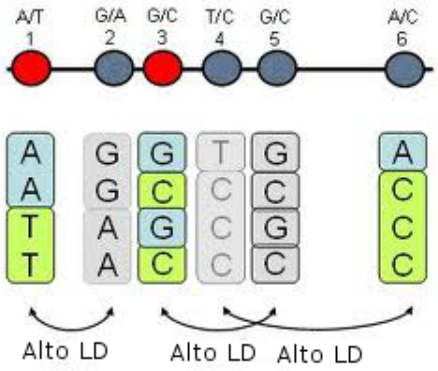


Figura 3.3 LD entre pares de SNPs.
Adaptado de Pierce (2013).

de LD mais usadas são D' e r^2. Para demonstrar o cálculo de D, toma-se a Tabela 3.1 para representar a hipótese de independência entre os *loci* 1 e 2. A Tabela 3.2 considera a distribuição alélica sob a hipótese de dependência indicada pelo LD.

Tabela 3.1 Distribuição alélica esperada sob a hipótese de independência. Adaptado de Foulkes (2009).

Locus		Locus 2		Total parcial
		B	b	
Locus 1	A	$n_{11} = Np_Ap_B$	$n_{12} = Np_Ap_b$	$n_{1.}$
	a	$n_{21} = Np_ap_B$	$n_{22} = Np_ap_b$	$n_{2.}$
Total parcial		$n_{.1}$	$n_{.2}$	$N = 2n$

Com a introdução do escalar D, consegue-se medir o LD entre os dois SNPs. Se D é igual a zero, a Tabela 3.2 é reduzida para a Tabela 3.1. Se D é próximo de zero, os valores observados serão próximos aos valores sob a hipótese de independência. Caso D

Tabela 3.2 Distribuição alélica esperada sob a hipótese de LD. Adaptado de Foulkes (2009).

Locus		Locus 2		Total parcial
		B	b	
Locus 1	A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1.}$
	a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2.}$
Total parcial		$n_{.1}$	$n_{.2}$	$N = 2n$

seja relativamente alto, então os valores n_{11} e n_{22} observados serão superiores aos seus respectivos esperados sob independência, e os números n_{12} e n_{21} observados serão inferiores aos esperados. Pode-se expressar D em termos da probabilidade conjunta de A e B e o produto das probabilidades alélicas individuais como expresso na Equação 3.1.

$$D = p_{AB} - p_A p_B \tag{3.1}$$

Na prática, estima-se D, juntando as correspondentes estimativas das probabilidades marginais e conjuntas (FOULKES, 2009). É simples mostrar que $\widehat{p_A} = \frac{n_{1.}}{N}$ e $\widehat{p_B} = \frac{n_{.1}}{N}$ são as estimativas de p_A e p_B, respectivamente. A estimativa de p_{AB} não é tão simples, no entanto, uma vez que não é observado o número de homólogos na nossa amostra com os alelos A e B. Ou seja, o número de cópias do haplótipo[5] AB é incerto, resultante do fato de se dispor de dados com base na população de indivíduos não aparentados (FOULKES, 2009). Para estimar p_{AB} neste cenário, maximiza-se o logaritmo da verossimilhança condicional aos dados observados. Mas esse procedimento não possui solução analítica fechada, portanto necessita de métodos numéricos apropriados.

Como os números na Tabela 3.2 representam quantidades, então D fica restrito a valores não-negativos que estão relacionados com p_A, p_a, p_B e p_b e dessa forma, pode-se reescalá-lo conforme as Equações 3.2 e 3.3 (FOULKES, 2009). Essa nova medida derivada de D é chamada de D' (pronuncia-se D *prime*).

$$D' = \frac{|D|}{D_{max}} \tag{3.2}$$

Onde,

$$D_{max} = \begin{cases} min(p_A p_b, p_a p_B) & \text{se } D > 0; \\ min(p_A p_B, p_a p_b) & \text{se } D < 0. \end{cases} \tag{3.3}$$

[5] É o conjunto específico de SNP e outras variantes genéticas observadas em um único cromossomo ou parte de um cromossomo (PIERCE, 2013).

Outra medida bastante usada para LD é o r^2. Ela é baseada na estatística χ^2 para o teste de associação entre linhas e colunas de uma tabela de contigência. A fórmula para seu cálculo é dada pela Equação 3.4.

$$r^2 = \frac{\chi_1^2}{N} \tag{3.4}$$

Efetuando devidas manipulaçõees algébricas é possível mostrar que r^2 pode ser escrito em função de D com fator de escala diferente do usado em D' como é indicado pela Equação 3.5.

$$r^2 = \frac{\chi_1^2}{N} = \frac{D^2}{p_A p_B p_a p_b} \tag{3.5}$$

Segundo Hartl e Clark (2006), D' é uma medida de desequilíbrio de ligação que é principalmente influenciada pela quantidade de recombinação, enquanto r^2 também captura informação sobre quando e onde na genealogia dos haplótipos as mutações aconteceram. Essa diferença explica por que D' e r^2 são medidas complementares e também por que r^2 pode assumir uma faixa de valores para qualquer valor de D' (HARTL; CLARK, 2006). As duas métricas de LD, D' e r^2, possuem uma propriedade na qual elas obscurecem a direção do desequilíbrio de ligação, pois o sinal original (positivo ou negativo) de D é perdido (HARTL; CLARK, 2006).

3.2.2 Blocos LD e tag SNPs

Na subseção anterior, foram mostradas duas formas para o cálculo do LD entre um par de marcadores. Mas, a abordagem mais geral para tratar a redundância de informação dada pelo alto LD entre vários pares de SNPs, em uma determinada região do DNA, é a construção de blocos de SNPs. Uma possível forma de montar tais blocos é separar os grupos de marcadores identificando regiões *hotspots*[6], as quais tem grande chance de sofrer recombinação. Para computar o LD desse bloco, basta calcular a média do LD entre todos os pares que compõem o mesmo. Feito isso, escolhe-se um marcador representante, denominado *tag* SNP, para esse bloco e utiliza-o juntamente com informações fenotípicas na tentativa de encontrar o subconjunto de *tag* SNPs que estão associados ao fenótipo.

[6]São regiões onde os eventos de recombinação ocorrem com maior frequência. A taxa de pico dentro dessas regiões pode ser centenas ou milhares de vezes maior do que na região circundante (JEFFREYS; KAUPPI; NEUMANN, 2001).

Com esse procedimento, haveria uma redução de dimensão para problemas em GWAS, pois grande parte da redundância estaria sendo eliminada pelo *tag* SNP. A Figura 3.4 mostra uma possível configuração de *tag* SNPs em blocos LD separados por *hotspots*.

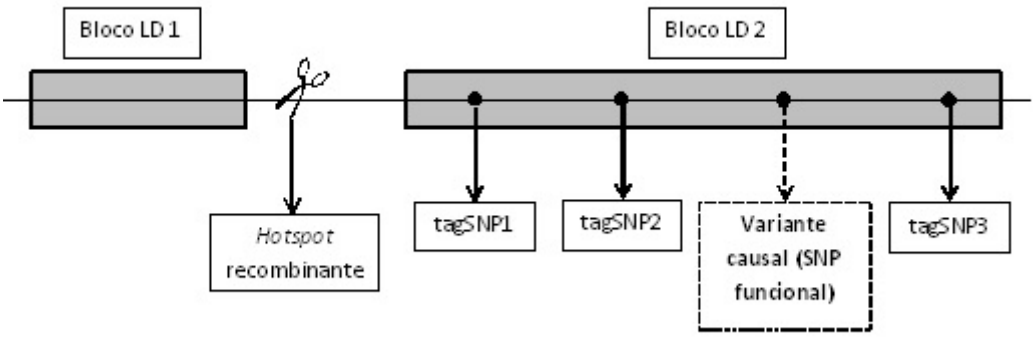


Figura 3.4 Blocos LD, *tag* SNPs e regiões *hotspots*.
Adaptado de Foulkes (2009).

O entendimento e o uso do LD entre marcadores pode melhorar o desempenho de métodos ou técnicas usadas em GWAS e fornecer melhor compreensão sobre a estrutura de associação não-aleatória entre marcadores encontrados ao final do processo de seleção. Como consequência, um método de seleção de SNPs deve selecionar os marcadores mais informativos com baixo ou nenhum LD, pois, desta forma, a redudância será reduzida ou até mesmo eliminada.

De acordo com Foulkes (2009), os blocos LD diferem substancialmente entre raças e grupos étnicos e tendem a ser menores para negros não-hispânicos do que para brancos e hispânicos. Como resultado, um conjunto de *tag* SNPs podem capturar informações de variantes causais de uma doença em um grupo, mas não em outro (FOULKES, 2009). Logo, considerar esse fenômeno e a aplicação de abordagens apropriadas é crucial para GWAS baseados em populações que incluam múltiplas raças e grupos étnicos (FOULKES, 2009).

Existem diversos procedimentos para a construção de blocos LD baseados nas medidas r^2 ou D'. Por exemplo, os *softwares* Haploview (BARRETT et al., 2005) e Plink (PURCELL et al., 2007) utilizam a mesma técnica para criar os blocos LD, sendo esses algoritmos descritos em Gabriel et al. (2002) e Wang et al. (2002). Entretanto, é importante ressaltar que é possível construir outros blocos LD por meio da alteração de

parâmetros específicos.

3.3 Princípio de Hardy-Weinberg

O princípio de Hardy-Weinberg diz que em uma população, atendidos determinados pressupostos, as frequências alélicas permanecerão constantes ao longo das gerações. Independentemente de um gene ser raro ou frequente, sua frequência permanecerá a mesma com relação aos outros desde que essas condições sejam mantidas. Intuitivamente, poder-se-ia supor que alelos raros se tornariam cada vez mais raros e que alelos frequentes aumentariam cada vez mais sua frequência, simplesmente por já serem raros ou comuns, mas o princípio de Hardy-Weinberg demonstra matematicamente que isso não ocorre (LAIRD; LANGE, 2011).

A primeira implicação do princípio de Hardy-Weinberg é que apenas a reprodução não causa evolução, sendo necessária a presença de outros fatores tais como seleção natural, mutação, migração ou o acaso para que as populações evoluam (PIERCE, 2013). De acordo com Pierce (2013), uma segunda implicação é que, quando uma população está em equilíbrio de Hardy-Weinberg, as frequências genotípicas são determinadas pelas frequências alélicas. Supondo um *locus* com dois alelos, a frequência do heterozigoto é maior quando as frequências alélicas estão entre 0,33 e 0,66, atingindo o máximo quando as frequências alélicas estão em 0,50 (Figura 3.5). Para o caso mais simples de um único *locus* com dois alelos A e a com frequências alélicas p e q, respectivamente, o princípio de Hardy-Weinberg prediz que a frequência genotípica para o homozigoto AA será p^2, para o heterozigoto Aa será $2pq$ e o outro homozigotos aa será de q^2.

Os pressupostos originais para o princípio de Hardy-Weinberg são que a população considerada é idealizada como sendo infinita, no sentido de eliminar-se a deriva genética[7]). Além do mais, a população está sob um regime de reprodução sexuada; os indivíduos acasalam-se aleatoriamente (sem seleção sexual ou desvio de aleatoriedade por dispersão

[7]Deriva genética, deriva gênica, deriva alélica, derivação genética ou ainda oscilação genética é um mecanismo microevolutivo que modifica aleatoriamente as frequências alélicas ao longo do tempo (RIDLEY, 2006). A deriva genética é um processo estocástico, não é possível prever a direção da mudança na frequência de um alelo causada pela deriva (RIDLEY, 2006). Esse mecanismo resulta em perda de variação genética e na fixação de alelos em diferentes loci. Os alelos fixados pela deriva podem ser neutros, deletérios ou vantajosos. Nesses dois últimos casos, a trajetória da frequência alélica ao longo do tempo será determinada pela interação entre a deriva e a seleção natural (FREEMAN; HERRON, 2009). O efeito da deriva é maior quanto menor o tamanho da população, podendo aparecer em diferentes momentos da história evolutiva da população (FREEMAN; HERRON, 2009)

geográfica); existem somente diplóides ou poliplóides; o número de fêmeas igual ao número de machos, todos os casais são férteis e têm o mesmo número de prole e não sofre de seleção natural, mutações e migração (ausência de fluxo gênico) (LAIRD; LANGE, 2011).

Desse modo, quando SNPs possuem frequências alélicas observadas próximas às esperadas pelo princípio de Hardy-Weinberg, diz-se que eles estão em equilíbrio de Hardy-Weinberg (do inglês, *Hardy-Weinberg Equilibrium* - HWE), entretanto, caso alguma premissa não seja satisfeita, pode-se gerar algum desvio e, neste caso, os marcadores estão em desequilíbrio de Hardy-Weinberg. O desvio em relações às frequências de equilíbrio é testado usando o teste de aderência qui-quadrado que será abordado com maiores detalhes no próximo tópico sobre pré-processamento de dados genômicos. Esse teste pode ser usado para avaliar se há erros de genotipagem, se ocorre endogamia, ou em alguns casos, existe estratificação na população avaliada. As curvas de probabilidades esperadas para marcadores que estão em HWE podem ser avaliadas pela Figura 3.5.

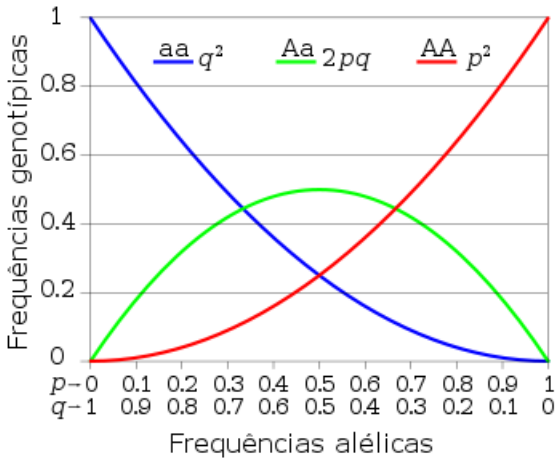


Figura 3.5 Curvas de frequência genotípica para os homozigotos AA e aa e para o heterozigoto Aa em HWE.
Onde p e q indicam as frequências alélicas de A e a respectivamente. Adaptado de Pierce (2013).

3.4 Pré-processamento de Dados Genômicos (Controle de Qualidade)

Nesta seção serão apresentadas as medidas mais usadas para controle de qualidade dos dados referentes aos marcadores moleculares usados em GWAS. Os fitros *call rate*, MAF e teste para equilíbrio de Hardy-Weinberg serão apresentados e exemplificados.

3.4.1 Call rate

Segundo Silva (2013), a *call rate* (CR) é uma medida de qualidade utilizada para eliminar SNPs com grande quantidade de "valores perdidos" (*missing genotypes*). Esta métrica é calculada proporcionalmente em relação ao número de observações válidas (*nonmissing genotypes*) e geralmente opta-se por trabalhar com SNPs cuja *call rate* seja maior que 95% $(0,95)$. A Expressão 3.6 mostra como se calcula a CR.

$$CR = 1 - \frac{x}{y} \tag{3.6}$$

Onde x representa o número de valores faltantes e y é o número de não-faltantes. Como exemplo, considere para um determinado *locus* de um marcador SNP em uma amostra de 500 indivíduos, sendo que desse total somente 63 estão ausentes por erro de genotipagem. Assim, a CR é calculada pela Equação 3.7 da seguinte forma:

$$CR = 1 - \frac{63}{437} = 0,86 \tag{3.7}$$

3.4.2 MAF (Minor Allele Frequency)

É uma medida relacionada com a variação dos alelos na população, alelos pouco variáveis são pouco informativos e não apresentam relevância genética na população (SILVA, 2013). Geralmente utiliza-se MAF $\geq 5\%$ (ou 0,05). A Equação 3.8 demonstra como é efetuado o cálculo da MAF a partir da proporção do alelo A ($f(A)$) e da proporção do alelo a ($f(a)$), ou seja, entre os valores $f(A)$ e ($f(a)$) toma-se o menor.

$$MAF = min(f(A), f(a)) \tag{3.8}$$

sendo:

$$f(A) = \frac{2 \times \text{número de AA+número de Aa}}{2 \times \text{número total de indivíduos}}$$

e,

$$f(a) = \frac{2 \times \text{número de aa+número de Aa}}{2 \times \text{número total de indivíduos}}$$

Suponha uma população com 437 pessoas e que em um *locus* tem-se 53 homozigotos dominantes AA, 196 heterozigotos Aa e 188 homozigotos recessivos aa. Daí, a MAF é igual a 0,35 e seu cálculo é mostrado pela Equação 3.9.

$$MAF = min\left(\frac{2 \times 53 + 196}{2 \times 437}; \frac{2 \times 188 + 196}{2 \times 437}\right) = min(0,35; 0,65) = 0,35 \tag{3.9}$$

3.4.3 Correção de Bonferroni

Existem dois tipos de correções para múltiplos testes de hipóteses: correção de um único passo e correção em dois passos (FOULKES, 2009). O primeiro é o ajuste de um único passo (*single-step adjustment*), em que um critério simples é usado para assegurar a significância de todos os testes estatísticos ou valores-p correspondentes (FOULKES, 2009). O segundo tipo é o *step-down adjustment*, o qual envolve a ordenação estatística dos valores-p dos testes e, em seguida, utiliza um critério potencialmente diferente para cada um dos valores ordenados (FOULKES, 2009).

Suponha que realizaram-se m testes de hipóteses dados por $H_0^1, H_0^2, ..., H_0^m$, e cada teste é controlado com um nível de significância α. Isto significa que, para um teste simples, a probabilidade de rejeitar a hipótese nula incorretamente, a taxa de erro do tipo I, é menor ou igual a α. A Equação 3.10 formaliza esse fato para todo $i = 1, 2, ..., m$.

$$Pr(\text{rejeitar } H_0^i \mid H_0^i \text{ é verdadeira}) \leq \alpha \tag{3.10}$$

Note que assume-se que esses testes são independentes uns dos outros, então a probabilidade de rejeição de um teste não depende da resposta de outros testes. Seja V o número de hipóteses nulas verdadeiras que são declaradas significantes, então a probabilidade de rejeitar incorretamente ao menos uma hipótese nula é dada pela Equação 3.11.

$$\begin{aligned} Pr(V \geq 1 \mid H_0^C \text{ é verdadeira}) &= 1 - Pr(V = 0 \mid H_0^C \text{ é verdadeira}) \\ &= 1 - \prod_{i=1}^{m} [Pr(\text{não rejeitar } H_0^i \mid H_0^i \text{ é verdadeira})] \\ &= 1 - \prod_{i=1}^{m} [1 - Pr(\text{rejeitar } H_0^i \mid H_0^i \text{ é verdadeira})] \\ &\leq 1 - \prod_{i=1}^{m} (1 - \alpha) = 1 - (1 - \alpha)^m \end{aligned} \tag{3.11}$$

Observe que se $m = 1$, então a Equação 3.10 se reduz a um teste simples. Porém, se $m = 2$, dois testes independentes são realizados, cada um com $\alpha = 0,05$, então a probabilidade de cometer um erro do tipo I é menor ou igual a $1 - (1 - 0,05)^2 = 0,0975$. Realizando 10 testes cada um ao nível α, a probabilidade de ao menos um erro do tipo I ocorrer dado que todas as hipóteses nulas são verdadeiras é menor ou igual a $1 - (1 - 0,05)^{10} = 0,401$ e assim por diante.

A correção de Bonferroni é um dos procedimentos mais simples para aplicar em múltiplas comparações. Basta usar $\alpha' = \dfrac{\alpha}{m}$ no lugar de α para o nível de significância de cada teste, onde m é o número de testes a ser realizado. Por exemplo, se queremos conduzir $m = 10$ testes de hipóteses e queremos controlá-los a um nível global de $\alpha = 0,05$, então $\alpha' = \dfrac{0,05}{10} = 0,005$. Neste caso, a Equação 3.11 reduz à Equação 3.12, onde $H_0^C = [H_0^1, H_0^2, ..., H_0^m]$.

$$Pr(V \geq 1 \mid H_0^C \text{é verdadeira}) \leq 1 - (1 - 0,005)^{10} = 1 - 0,0951 = 0,049 \tag{3.12}$$

Assim, se controlamos cada um dos m testes no nível $\alpha' = \dfrac{\alpha}{m}$, então, controlamos globalmente ao nível α, a probabilidade de ao menos um erro do tipo I ocorrer dado que todas as hipóteses nulas são verdadeiras.

Segundo Gondro, Werf e Hayes (2013), a correção de Bonferroni não leva em conta que os "testes" são feitos no mesmo cromossomo, não sendo independentes, além de que os marcadores podem estar em desequilíbrio de ligação uns com os outros, bem como o QTL. Como resultado, a correção de Bonferroni tende a ser muito conservadora, ou requer alguma decisão a ser tomada sobre a forma como muitas regiões independentes do

genoma foram testadas.

3.4.4 Teste para Equilíbrio de Hardy-Weinberg

É utilizado para verificar se as frequências genotípicas observadas estão de acordo com as esperadas conforme o equilíbrio de Hardy-Weinberg (do inglês, *Hardy-Weinberg Equilibrium* - HWE), caso não estejam, podem haver problemas em exercer a seleção considerando os *locus* que desviam do HWE, pois estes podem estar altamente influenciados pelo tamanho da população, mutação, migração e seleção (SILVA, 2013).

Tabela 3.3 Estatísticas dos fenótipos avaliados

Frequência	AA	Aa	aa
Frequência observada (O)	53	196	188
Frequência esperada (E)	$Nf(A)^2$	$2Nf(A)f(a)$	$Nf(a)^2$
Valor de E	$437 \times 0,35^2 = 52,7$	$2 \times 437 \times 0,35 \times 0,65 = 197,64$	$437 \times 0,65^2 = 187,17$

De acordo com a Tabela 3.3, realiza-se o teste de aderência qui-quadrado com o intuito de verificar se a frequência observada é próxima da esperada. Para isso é necessário calcular o valor crítico da distribuição qui-quadrado como observado na Equação 3.13.

$$\chi_c^2 = \sum_{i=1}^{3} \frac{(O-E)^2}{E} = \frac{(53-52,17)^2}{52,17} + \frac{(196-197,64)^2}{197,64} + \frac{(188-187,17)^2}{187,17} = 0,0304 \tag{3.13}$$

Se $\chi_c^2 \geq \chi_\alpha^2$, então rejeita-se H_0 e, caso contrário, aceita-se H_0 a um nível de significância α adotado *a priori*.

Neste exemplo, o valor-p associado ao valor crítico é valor-p $= 1 - P(0,0304; df = 1) = 0,8615857$, onde $P(0,0304; df = 1)$ é a probabilidade de se ter um valor menor ou igual a 0,0304 a partir de uma distribuição qui-quadrado com 1 grau de liberdade (do inglês, *degree of freedom* - *df*).

Uma importante observação na aplicação simultânea do teste de HWE em m marcadores SNP é que deve-se realizar a correção do nível global de significância, pois do contrário, o erro do tipo I será aumentado rapidamente como foi demonstrado na subseção anterior. Logo, o modo mais prático para se realizar esse ajuste é por meio da correção de Bonferroni, discutida anteriormente, a qual considera somente os marcadores com valor-p menor que $\frac{\alpha}{m}$. Mas esse método considera que o crescimento do erro do tipo I

aumenta de forma linear com o número de SNPs e na realidade, esse aumento é não-linear sendo, na verdade, dominado pela função linear, o que superestima o erro gerado pela não correção. Ou seja, essa correção é muito restritiva, podendo eliminar verdadeiros-positivos associados com o fenótipo em questão.

3.5 Resumo do Capítulo

O desequilíbrio de ligação é importante para agrupar os SNPs em haplótipos (blocos LD), permitindo a diminuição da redundância existente. As medidas *call rate* e MAF e o teste para equilíbrio de Hardy-Weinberg podem ser usadas como filtros para controle de qualidade, diminuindo o nível de ruído nos dados genômicos (SNPs).

4 Técnicas de Inteligência Computacional

No presente capítulo, todas as ferramentas de Inteligência Computacional usadas no método de seleção de atributos desenvolvido neste trabalho serão descritas de forma detalhada, além da discussão de suas vantagens e desvantagens. Algumas medidas usadas em classificação e regressão também serão apresentadas, pois as mesmas serão usadas no método proposto.

4.1 Avaliadores de Métodos para Classificação e Regressão

4.1.1 Validação Cruzada

A validação cruzada é um método empregado para avaliação e posterior comparação entre modelos de classificação ou de regressão por meio da estimação da acurácia dos mesmos, a qual é definida por uma medida especificada a *priori*. Esse método divide o conjunto de instâncias ou observações com n elementos em k partes iguais quando o n é divisível por k, ou em $k-1$ partes iguais e a k-ésima parte com o número de elementos igual ao resto da divisão entre n e k. Nessa abordagem, cada observação é usada um mesmo número de vezes para treinamento e exatamente uma única vez para teste.

Na Figura 4.1 é exemplificado o uso de k-*fold* com $k = 4$. Primeiramente, o conjunto de dados inicial é dividido em 4 partes, onde a primeira parte em vermelho é separada para teste e as outras 3 partes em azul são usadas para o treinamento do classificador. Esse procedimento é replicado k vezes e o erro de predição é calculado a partir dos k conjuntos de teste. Cada subgrupo é usado somente uma vez no teste e 3 vezes no treinamento, isto é, todas as observações são usadas ora para treinamento, ora para teste. Com isso, elimina-se o possível viés existente na partição do conjunto em apenas dois blocos, como é feito no método *Holdout*, pois a escolha da divisão em duas partes pode subestimar ou superestimar o classificador ou regressor avaliado. Portanto, como concluiu Kohavi

(1995), o melhor método para usar em seleção de modelos é a validação cruzada com 10-*fold*, mesmo se o poder computacional permite usar mais *folds*.

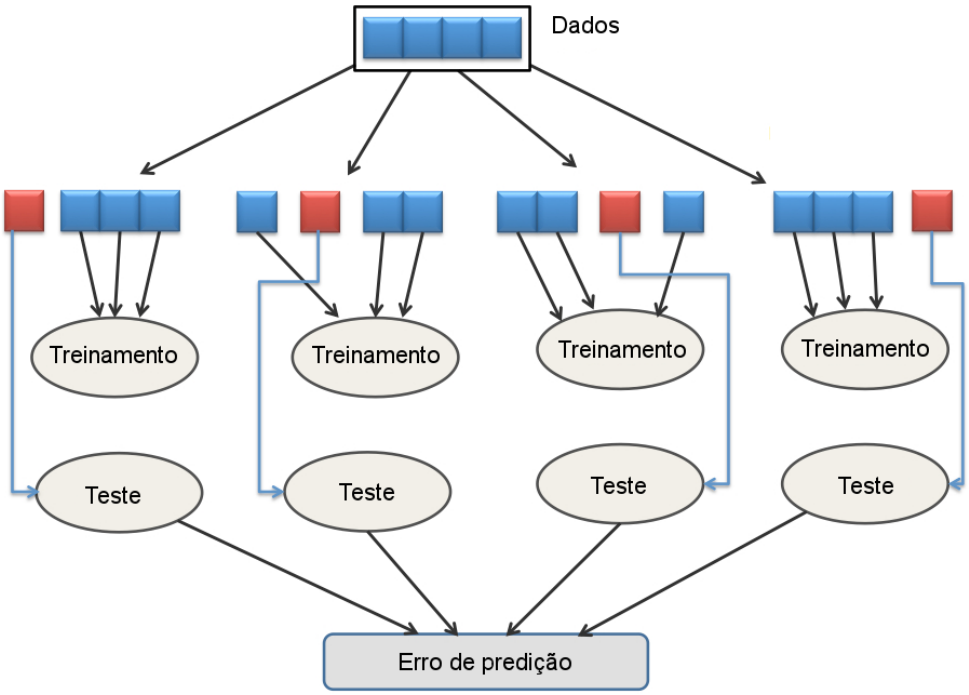


Figura 4.1 Exemplo de *k-fold* com $k = 4$.
Adaptado de Castellano (2014).

No presente estudo, quatro diferentes medidas estatísticas foram empregadas para julgar o desempenho dos resultados de uma regressão a saber: erro médio quadrático (do inglês, *mean square error* - MSE), erro percentual médio absoluto (do inglês, *mean absolute percentage error* - MAPE), coeficiente de correlação de Pearson (r) e coeficiente de correlação de Spearman (ρ), que são definidos matematicamente pelas Equações 4.1, 4.2, 4.3 e 4.4, respectivamente. Para problemas de classificação, foi usada a área abaixo da curva ROC (AUC) que será explicada na seção posterior.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{4.1}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{X_i - Y_i}{Y_i}\right|\right) \tag{4.2}$$

$$r = \left(\frac{\sum_{i=1}^{n}\left[(X_i - \bar{X})(Y_i - \bar{Y})\right]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}\right) \tag{4.3}$$

$$\rho = \left(\frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \right) \tag{4.4}$$

onde n é o número de amostras, X_i é o valor observado, Y_i é o valor predito, $\bar{X}$ é a média dos valores observados, $\bar{Y}$ é a média dos valores preditos, x_i e y_i são os postos[1] das variáveis com seus valores brutos dados por X_i e Y_i, e k é o número de variáveis de entrada.

O método de avaliação para todos modelos de regressão e classificação usados no presente trabalho foi o 10-*fold* para as quatro medidas descritas anteriormente. Além disso, as dez partições geradas são as mesmas para todos os modelos de regressão e de classificação em todas as etapas que se faz uso da validação cruzada. Esse procedimento permite que se compare diretamente as predições geradas por cada modelo para cada partição ou para a média da medida adotada (correlação para regressão ou AUC para classificação) para as dez partições.

4.1.2 Área abaixo da Curva ROC

Considere um problema de classificação com somente duas classes, onde uma delas é designada como positiva (+), e a outra como negativa (-). Posto isso, consegue-se construir a matriz de confusão mostrada pela Tabela 4.1, onde VP é o número de verdadeiros positivos (número de exemplos da classe positiva classificados corretamente), VN é o número de verdadeiros negativos (número de exemplos da classe negativa classificados corretamente), FP é o número de falsos positivos (número de exemplos da classe negativa que foram classificados incorretamente na classe positiva) e FN é o número de falsos negativos (número de exemplos da classe positiva que foram classificados incorretamente na classe negativa). Cabe ressaltar que essas medidas são baseadas em instâncias do conjunto de teste, pois o objetivo principal é a maximização do poder de generalização do classificador em análise.

Com base na matriz de confusão, várias medidas de desempenho para classificadores podem ser desdobradas tais como a taxa de falsos positivos (TFP) (Expressão 4.5) e taxa de verdadeiros positivos (TVP) (Expressão 4.6). O símbolo $\hat{f}$ significa a predição do

[1] São as posições em ordem crescente para os valores da variável em questão.

Tabela 4.1 Matriz de confusão para um problema com duas classes. Adaptado de Faceli et al. (2011).

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

classificador f para todas instâncias avaliadas (classe predita).

$$TFP(\hat{f}) = \frac{FP}{FP + VN} \tag{4.5}$$

$$TVP(\hat{f}) = \frac{VP}{VP + FN} \tag{4.6}$$

Outra forma de visualização, organização e avaliação de classificadores binários é por meio do uso das curvas ROC (do inglês, *Receiving Operating Characteristics*). O gráfico ROC é um gráfico bidimensional onde os eixos X e Y representam respectivamente a taxa de falsos positivos (TFP) e a taxa de verdadeiros positivos (TVP) (FACELI et al., 2011). Logo, o desempenho de um classificador para um conjunto de dados pode ser plotado nesse gráfico como um ponto no espaço com duas dimensões.

Como exemplo, a Figura 4.2 exibe um gráfico ROC com os pontos A, B, C, D e E. A linha diagonal tracejada na Figura 4.2 simboliza classificadores que fazem predições aleatórias, por conseguinte, os classificadores A, B e D são superiores ao classificador C, o qual pode ser considerado aleatório, mas o classificador E é inferior ao C. Alguns pontos no gráfico ROC merecem destaque para melhor compreensão das análises subsequentes. O ponto D, cujas coordenadas são (0,1), representa um classificador perfeito, pois sua TFP é zero e sua TVP é um, sendo denominado "*céu ROC*" (FACELI et al., 2011). De forma contrária, o ponto (1,0) simboliza o "*inferno ROC*", pois a TFP é um e a TVP é zero. O ponto (1,1) indica classificações sempre positivas, e o ponto (0,0) sempre negativas. Um ponto no gráfico ROC é melhor que outro se o primeiro está a noroeste do segundo, ou seja, o primeiro ponto tem uma TVP maior e uma TFP menor do que as respectivas medidas do segundo ponto (FAWCETT, 2006).

Conforme Faceli et al. (2011), apesar de existirem mecanismos para comparação de classificadores baseados em seus pontos no gráfico ROC, o procedimento mais usual é construir uma curva ROC. Assim, a partir de um conjunto de pontos no gráfico

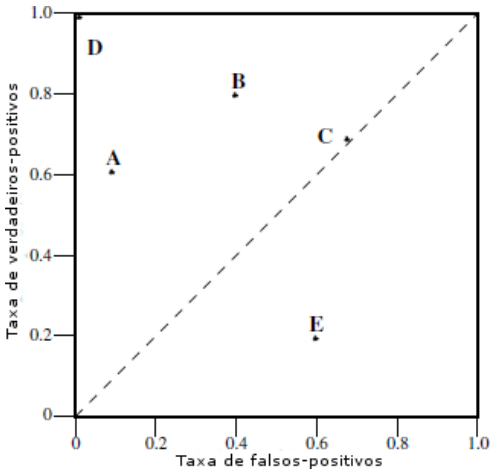


Figura 4.2 Um gráfico ROC com cinco classificadores discretos.
Adaptado de Fawcett (2006).

ROC associados a um classificador, pode-se construir uma curva. Para decidir entre dois classificadores com curvas ROC que não se interceptam, basta escolher a que está mais próxima do ponto (0,1). A Figura 4.3 indica as áreas abaixo das curvas ROC para dois classificadores denominados A e B, onde as curvas ROC dos mesmos se interceptam, gerando dúvida quanto ao melhor classificador quando avaliados para cada ponto bidimensional no gráfico ROC, pois existem regiões onde um classificador é melhor que o outro. Daí, para eliminar essa inconclusão, criou-se um única medida associada à curva ROC, denominada área abaixo da curva ROC (do inglês, *Area Under ROC Curve* - AUC), daí, o classificador que possui a maior área abaixo da curva ROC é considerado melhor. No caso da Figura 4.3, a partir do critério da AUC, o classificador B é preferível ao A. Os valores de AUC variam entre 0 e 1 e quanto mais próximo de 1, melhor é o classificador. Uma possibilidade é utilizar a área abaixo da curva ROC em conjunto com a validação cruzada com k-*fold* para avaliar um classificador ou para comparar classificadores. Uma possível medida de avaliação dos classificadores seria a média das áreas abaixo da curva ROC dos k-*fold* utilizados.

4.2 Métodos de Aprendizado Supervisionado

4.2.1 Árvores de Decisão e de Regressão

Duas técnicas muito difundidas nas áreas de Mineração de Dados (do inglês, *Data Mining*) e Aprendizado de Máquina (do inglês, *Machine Learning*) são as Árvores de Decisão para

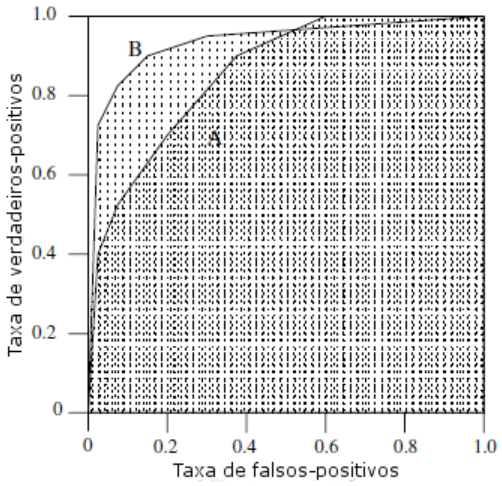


Figura 4.3 Gráficos da área abaixo da curva ROC para dois classificadores. Adaptado de Fawcett (2006).

problemas de classificação e as Árvores de Regressão, para regressão, sendo aplicadas nos mais diversos campos tais como: Biologia, Finanças, Química entre outros. A ideia central dessas técnicas para resolver problemas complexos de predição e/ou de explicação é dividir os mesmos em subproblemas mais simples, aos quais recursivamente é aplicado o mesmo procedimento (FACELI et al., 2011). A partir desse ponto, conforme Faceli et al. (2011), as soluções dos subproblemas são combinadas em formato de árvore para gerar uma solução para o problema inicial. A Figura 4.4 elucida a ideia básica da estratégia *dividir para conquistar* de uma árvore de decisão.

No caso de Árvores de Decisão (problemas de classificação), a ideia básica de funcionamento do algoritmo é construir uma árvore[2] partindo de um nó inicial, denominado nó raiz, que representa a variável mais "importante" para a previsão da variável explicada y, inserindo sucessivamente, em ordem decrescente de "importância", outros nós (variáveis explicativas) na árvore. O processo termina quando algum critério de parada é atendido, como, por exemplo, quando um determinado subgrupo tenha menos que quatro indivíduos.

Segundo Foulkes (2009), formalmente, seja uma variável explicada $\mathbf{y} = (y_1, y_2, ..., y_n)$ e um conjunto potencial $\mathbf{X} = \{x_1, x_2, ..., x_p\}$ com p variáveis explicativas sendo cada variável x_i um vetor da forma $x_i = (x_{1i}, x_{2i}, ..., x_{ni})^T$, onde $i = 1, 2, ..., n$ indica os indivíduos na amostra considerada. Seja Ω o conjunto de todos os indivíduos na amostra de treino.

[2] Árvore é um grafo acíclico direcionado em que cada nó ou é um nó de divisão ou um nó folha.

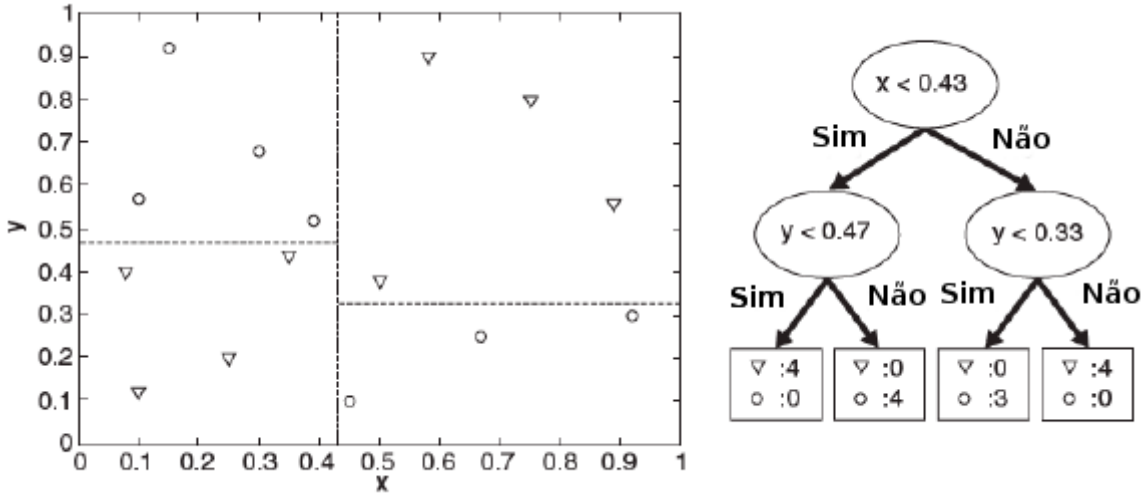


Figura 4.4 Na direita, uma árvore de decisão, e na esquerda, os correspondentes limites de decisão no plano cartesiano.
Adaptado de Tan, Steinbach e Kumar (2006).

Para simplificar o raciocínio, suponha que todas variáveis explicativas sejam binárias podendo assumir apenas valores 0 e 1, e que $x_{(1)}$ á a variável preditora mais importante em relação a y_i. Primeiramente, indivíduos são divididos em dois grupos Ω_1 e Ω_2 baseado no valor de corte de $x_{(1)}$. Ou seja, pode-se definir dois conjuntos $\Omega_1 = \{i : x_{i(1)} = 0\}$ e $\Omega_2 = \{i : x_{i(1)} = 1\}$ para os indivíduos designados por $i = 1, 2, ..., n$. O próximo passo do algoritmo é identificar novamente a variável preditora mais importante $x_{(2)}$ para y_i, excluindo $x_{(1)}$ do conjunto restante de variáveis. Esse procedimento de escolha da variável preditora mais importante é continuado até o critério de parada ser estabelecido. Como exemplo, na Expressão 4.7 estão identificados os subgrupos para as variáveis $x_{(2)}$ e $x_{(3)}$ e a estrutura final dessa árvore pode ser vista na Figura 4.5 .

$$\begin{aligned}
\Omega_{1,1} &= \{i : i \in \Omega_1 \quad \text{e} \quad x_{i(2)} = 0\} = \{i : x_{i(1)} = 0 \quad \text{e} \quad x_{i(2)} = 0\} \\
\Omega_{1,2} &= \{i : i \in \Omega_1 \quad \text{e} \quad x_{i(2)} = 1\} = \{i : x_{i(1)} = 0 \quad \text{e} \quad x_{i(2)} = 1\} \\
\Omega_{2,1} &= \{i : i \in \Omega_2 \quad \text{e} \quad x_{i(3)} = 0\} = \{i : x_{i(1)} = 1 \quad \text{e} \quad x_{i(3)} = 0\} \\
\Omega_{2,2} &= \{i : i \in \Omega_2 \quad \text{e} \quad x_{i(3)} = 1\} = \{i : x_{i(1)} = 1 \quad \text{e} \quad x_{i(3)} = 1\}
\end{aligned} \tag{4.7}$$

Para facilitar a visualização da estrutura de uma árvore de decisão induzida a partir de uma base de dados com características próximas a uma base real usada em GWAS, o

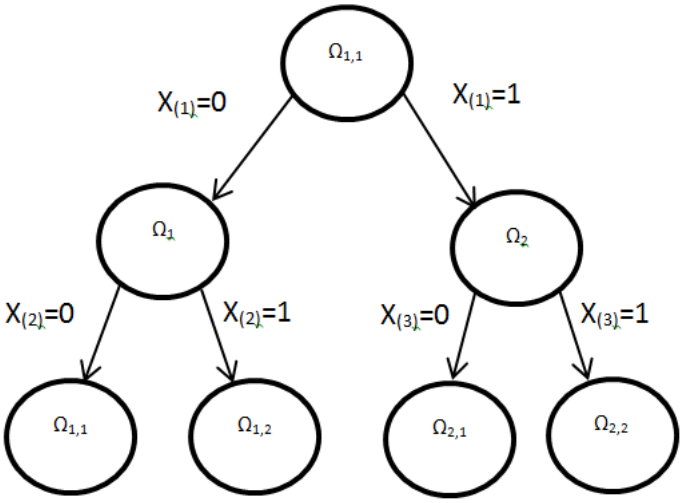


Figura 4.5 Estrutura da árvore de decisão.
Adaptado de Foulkes (2009).

genótipo e o fenótipo foram simulados a partir do pacote *scrime* do software R (TEAM, 2013), sendo o número de indivíduos igual a 1.000 e o número total de marcadores igual a 50. Nesta simulação, os SNPs 1,2,3,4 e 5 foram escolhidos para representarem os QTLs (SNPs causais). Conforme a Figura 4.6, nota-se que o SNP2, nó raiz, foi a variável mais importante no conjunto de 50 marcadores simulados, em segundo lugar, os SNPs 1 e 5, e, finalmente, os SNPs 3 e 4. Pela Figura 4.7, observa-se que o SNP2, nó raiz, foi a variável mais importante no conjunto de 50 marcadores simulados com fenótipo contínuo.

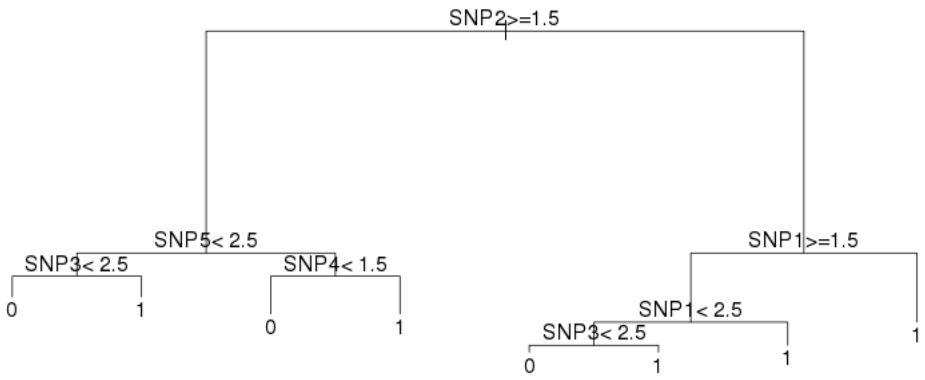


Figura 4.6 Exemplo de árvore de decisão aplicada ao conjunto de 50 SNPs com fenótipo dicotômico.

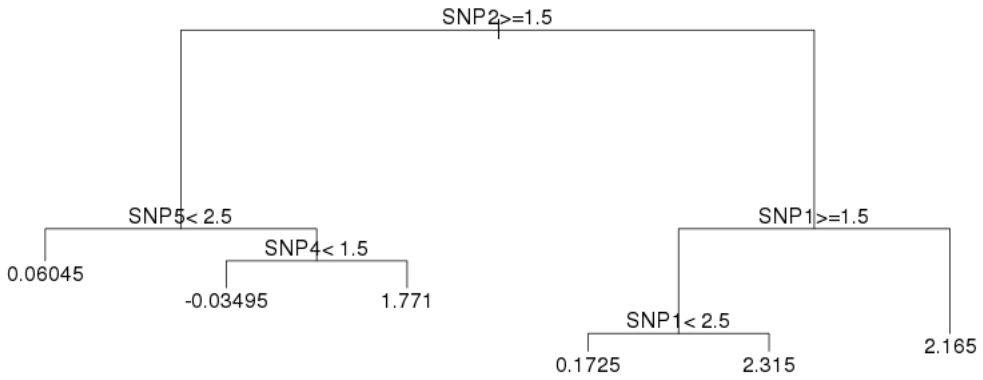


Figura 4.7 Exemplo de árvore de regressão aplicada ao conjunto de 50 SNPs com fenótipo contínuo.

O problema de encontrar a árvore com menor número de nós é um problema NP-completo (Rivest,1987). Por isso, são usadas diversas heurísticas que usam a estratégia de olhar um passo a frente para tornar o tempo linear em função do número de atributos.

Existem diversas medidas usadas para selecionar a melhor divisão nos nós de uma árvore. As mais utilizadas são a medida de entropia, o índice de Gini e o erro de classificação. As fórmulas dessas métricas são mostradas nas Expressões 4.8, 4.9 e 4.10. O símbolo $p(i|t)$ denota a fração de registros pertencentes à classe i em um determinado nó t.

$$Entropia(t) = \left[-\sum_{i=1}^{c} p(i|t) \log_2 p(i|t) \right] \tag{4.8}$$

$$Gini(t) = \left[1 - \sum_{i=1}^{c} [p(i|t)]^2 \right] \tag{4.9}$$

$$Erro(t) = \left[1 - \max_i [p(i|t)] \right] \tag{4.10}$$

Como comenta Faceli et al. (2011), as vantagens das árvores de decisão são apresentadas a seguir:

a) Flexibilidade

O modelo de árvore não possui a premissa quanto à distribuição dos dados, ou seja, essa técnica é não paramétrica (FACELI et al., 2011).

b) Robustez

Árvores são invariantes a qualquer transformação estritamente monótona nas variáveis da base de dados inicial (FRIEDMAN, 2001). Como exemplo, utilizar x_j, $log(x_j)$, e^{x_j} ou x_j^a como a j-ésima variável de entrada produz ávores com a mesma estrutura (FRIEDMAN, 2001). Por conseguinte, não existe a necessidade de utilizar estas transformações nos dados de entrada no intuito de melhorar a predição (FRIEDMAN, 2001). Outra consequência dessa invariância é a pequena sensibilidade à distribuições com "caudas longas" e *outliers* (FRIEDMAN, 2001).

c) Seleção de atributos

Durante o processo de construção da árvore, ocorre naturalmente a escolha das variáveis a serem usadas no modelo. Portanto, a árvore gerada é robusta na presença de variáveis redundantes e irrelevantes na base de dados inicial (FACELI et al., 2011).

d) Interpretabilidade

A estrutura de uma árvore de decisão é naturalmente traduzida em um conjunto de regras. Isso facilita o entendimento da relação entre as variáveis explicativas e a variável explicada, por mais complexa que seja essa relação (FACELI et al., 2011).

e) Eficiência

Os algoritmos usados para construção de árvores de decisão são lineares em relação ao tamanho da base de dados de treinamento, indicando boa eficiência no quesito de tempo computacional (FACELI et al., 2011).

Em contrapartida, como comenta Faceli et al. (2011), as desvantagens são enumeradas a seguir:

a) Replicação

Uma subárvore pode ser replicada várias vezes em níveis distintos em uma árvore de decisão e isto torna a árvore mais complexa do que necessário e reduz a

interpretabilidade do modelo (TAN; STEINBACH; KUMAR, 2006). Pagallo e Haussler (1990) argumentam que esse tipo de problema é inerente à representação de árvore.

b) Valores ausentes

Como a árvore toma decisão sobre os valores de determinada variável escolhida para realizar a quebra, então valores ausentes causam problemas relativos a que ramo seguir. Devido a esse problema, Friedman, Kohavi e Yun (1996) comentam que aproximadamente metade do código no CART e 80% dos esforços de programação foram desenvolvidos pela ausência de valores.

c) Atributos contínuos

Quando existem variáveis explicativas contínuas, torna-se necessário uma ordenação para cada nó de decisão (FACELI et al., 2011). Catlett (1991) argumenta que esta operação consuma 70% do tempo necessário para ajustar uma árvore de decisão em grandes conjuntos de dados com multiplos atributos contínuos.

d) Instabilidade

Pequenas variações no conjunto de treinamento podem gerar grandes variações na árvore final (BREIMAN et al., 1984; BREIMAN, 1996b; KOHAVI; KUNZ, 1997). Isso ocorre se para a escolha da variável usada na quebra há similaridade entre duas ou mais variáveis. Dessa forma, uma pequena alteração no conjunto de treino, muda toda subárvore abaixo daquele nó de decisão, pois é escolhida uma variável em detrimento de outra. Além disso, quanto mais próximo de um nó folha o nó em questão está, menos dados são usados para a quebra, logo, as inferências baseadas neste nó são menos confiáveis.

4.2.2 Métodos Ensemble

A agregação de vários classificadores básicos para melhorar a predição é uma técnica denominada métodos de grupos (do inglês, métodos *ensemble*). Um método de grupo constrói um conjunto de classificadores básicos a partir dos dados de treinamento e executa a classificação recebendo um voto sobre as previsões feitas por cada um dos classificadores básicos (TAN; STEINBACH; KUMAR, 2006).

Suponha que criou-se um grupo de 25 classificadores binários, sendo que cada um possui uma taxa de erro de $\epsilon = 0,35$. O classificador de grupo prevê o rótulo de classe de exemplo de teste recebendo o voto da maioria sobre as previsões realizadas pelos classificadores básicos. Se os classificadores básicos forem idênticos, então o grupo classificará erroneamente os mesmos exemplos previstos incorretamente pelos classificadores básicos. Logo, a taxa de erro permanece em 0,35. De forma contrária, se os classificadores básicos forem independentes, isto é, seus erros não estiverem correlacionados, então o grupo faz uma previsão errada se mais da metade dos classificadores de grupo preverem incorretamente. Assim, a taxa de erro do classificador de grupo é igual ao somatório da probabilidade de exatamente 13 classificadores errarem a classificação mais a probabilidade de exatamente 14 classificadores e assim sucessivamente, até a probabilidade de todos os 25 errarem a classificação. O somatório total, referente ao classificador de grupo, torna-se igual a 0,06, conforme mostra a Expressão 4.11. Ressalta-se que o número de termos da Expressão 4.11 é 13 $(25 - 13 + 1)$, ou seja, mais da metade dos classicadores básicos.

$$e_{grupo} = \sum_{13}^{25} \binom{25}{i} \epsilon^i (1-\epsilon)^{25-i} = 0,06. \tag{4.11}$$

A partir da Figura 4.8 pode-se concluir que para taxas de erros menores ou iguais a 0,5, um classificador de grupo formado por classificadores totalmente independentes melhoram a taxa de acerto para a classe predita. Todavia, para taxas de erros superiores a 0,5, a relação se inverte.

Como concluiu Tan, Steinbach e Kumar (2006), o exemplo ilustrado anteriormente mostra que duas condições são necessárias para que um classificador de grupo seja melhor do que um classificador único: os classificadores básicos devem ser independentes entre si e os classificadores básicos devem ser melhores do que um classificador que faça suposições aleatórias. Na prática, a independência total entre os classificadores básicos é difícil de ser garantida, entretanto, melhorias nas precisões de classificação têm sido observadas nos métodos de grupo que possuem classificadores ligeiramente correlacionados (TAN; STEINBACH; KUMAR, 2006).

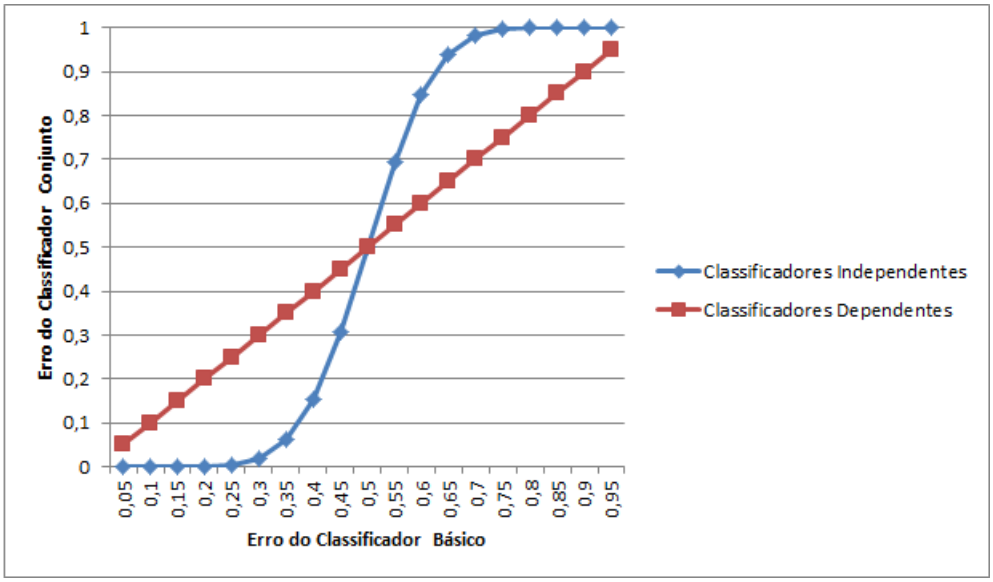


Figura 4.8 Comparação entre erros de classificadores básicos e erros do classificador de grupo.

O classificador de grupo é composto por 25 classificadores básicos, ora totalmente dependentes, ora totalmente independentes. Adaptado de Tan, Steinbach e Kumar (2006).

4.2.3 Random Forests

Para contornar a instabilidade observada nos classificadores gerados pelos algoritmos (CART, C4.5, dentro outros) para árvores de decisão e de regressão, Breiman propôs um método *ensemble* denominado *Bagging* (***B**ootstrap **Agg**regati**ng***) Breiman (1996a). *Bagging* é um procedimento simples onde sucessivas amostras de *bootstrap* de dados são selecionadas, $(\mathbf{x}^b, y^b)$, e uma previsão, $\hat{f}(\mathbf{x}^b)$, é derivada a partir de cada uma destas amostras. A previsão final, $\frac{1}{B}\sum_{b=1}^{B}\hat{f}(\mathbf{x}^b)$, é feita pela média das B previsões ou tomando o voto majoritário, $arg(max_k(\hat{f}_{bag}(\mathbf{x})))$, para classificação (GOLDSTEIN; POLLEY; BRIGGS, 2011).

Conforme Goldstein, Polley e Briggs (2011), um dos primeiros passos no entendimento de um preditor é analisar como suas predições contribuem para o viés e para a variância. Com isso, seja dado o vetor de saída y, o vetor de entrada $\mathbf{x}$ e a função $y = f(\mathbf{x}) + \epsilon$, onde $E(\epsilon) = 0$ e $Var(\epsilon) = \sigma_\epsilon^2$. Para um dado conjunto de treinamento T, a predição é $\hat{f}(\mathbf{x}|T)$. A decomposição para o erro de generalização sob a função de perda baseada no

erro quadrático com saída contínua é dada pela Expressão 4.12.

$$E_T\left[y-\hat{f}(\mathbf{x}|T)\right]^2 = \underbrace{\sigma_\epsilon^2}_{\text{ruído}} + \overbrace{\left[f(\mathbf{x})-E_T\hat{f}(\mathbf{x}|T)\right]^2}^{\text{viés}} + \underbrace{E_T\left[\hat{f}(\mathbf{x}|T)-E_T\hat{f}(\mathbf{x}|T)\right]^2}_{\text{variância}} \quad (4.12)$$

Onde a esperança é calculada sobre os conjuntos aleatórios de treinamento T. O primeiro termo é a variância de y referido como ruído e representa o erro irredutível. Os próximos dois termos representam o erro redutível. O primeiro deles é o viés, sendo a diferença sistemática entre a predição e a observação. O termo final é a variância, que mede a aleatoriedade da predição. Esse termo é importante para notar que a variância é independente da verdadeira saída y e da verdadeira função $f(\mathbf{x})$. Na classificação com saída 0-1, minimiza-se $P(\hat{f}(\mathbf{x}) \neq y)$ com $y \in \{0,1\}$. A função de perda erro de classificação é mostrada pela Expressão 4.13.

$$l(y,\hat{f}(\mathbf{x})) = \begin{cases} 1 & \text{se} \quad y \neq \hat{f}(\mathbf{x}) \\ 0 & \text{se} \quad y = \hat{f}(\mathbf{x}) \end{cases} \quad (4.13)$$

A ideia central da *Random Forest* (RF) é unir várias árvores classificadoras ou regressoras com pequeno viés e alta variância, para criar um preditor global com pequeno viés e pequena variância. Essa propriedade obtida pela agregação de várias árvores é conseguida pela técnica de amostragem denominada *Bagging*. Conforme Breiman (1996a), *Bagging* de preditores é um método para a geração de múltiplas versões de um preditor com objetivo de usá-los para conseguir uma predição agregada. As múltiplas versões são formadas por réplicas de *bootstrap* do conjunto de treinamento (BREIMAN, 1996a) e cada nó da árvore é escolhido a partir de um subconjunto de variáveis explicativas amostrado aleatoriamente do conjunto que engloba todos os atributos da base de dados inicial. Nenhum tipo de poda é feito na construção das árvores. Os vetores aleatórios baseados no conjunto de treinamento original são gerados a partir de uma distribuição de probabilidades fixa, diferentemente da abordagem adaptativa usada no *AdaBoost*, onde a distribuição de probabilidades é variada para enfocar exemplos que sejam difíceis de classificar (TAN; STEINBACH; KUMAR, 2006).

Testes em conjunto de dados reais e simulados utilizando árvores de classificação e de regressão e seleção de subconjuntos em regressão linear mostram que o *Bagging* pode dar

ganhos substanciais na precisão. Como argumenta Breiman (1996a), o elemento crucial é a instabilidade do método de predição em relação à mudanças no conjunto de treinamento, pois se perturbar o mesmo, podem ocorrer significativas alterações no preditor construído, logo o *Bagging* pode melhorar a precisão. Detalhadamente, na agregação de *bootstrap* ou *Bagging* são criadas amostras aleatórias de instâncias com reposição para treinamento a partir de um conjunto de dados original. Cada amostra possui o mesmo tamanho do conjunto inicial. Como a amostragem é realizada com reposição, algumas observações podem aparecer diversas vezes no mesmo conjunto de treinamento, enquanto que outras podem ser omitidas do mesmo. Com isso, a probabilidade de uma instância ser selecionada é $\frac{1}{n}$, logo, a probabilidade de uma observação não ser selecionada é $1 - \frac{1}{n}$. Em n repetiçõees, a probabilidade de uma observação não ser selecionada é $\left(1 - \frac{1}{n}\right)^n$, portanto, a probabilidade de uma instância ser selecionada em n repetições é $1 - \left(1 - \frac{1}{n}\right)^n$. Assim, fazendo as devidas manipulaçõees algébricas e usando o limite fundamental [3], tem-se que $\left[\lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e} \approx 0,632\right]$. Daí, segue-se que quando o tamanho do conjunto inicial n é suficientemente grande, em média, uma amostra *bootstrap* de treinamento conterá aproximadamente 63% do conjunto original, consequentemente, uma amostra de teste possuirá aproximadamente 37% do conjunto inicial. Esse conjunto de teste é denominado amostra *out-of-bag* (OOB). Breiman (1996c) comenta que uma das vantagens do *Bagging* é apresentar um meio computacionalmente eficiente para estimar o erro de generalização (ou erro de predição) em um conjunto de teste independente. A Figura 4.9 ilustra uma instância de teste que compõe as amostras OOB sendo aplicada em cada árvore da floresta. Breiman (1996c) demonstrou que a amostra OOB pode ser usada para estimar a medida de erro denominada de taxa de erro OOB.

Definição 1 (*Random Forest*)**.** *Uma RF é um classificador consistindo de uma coleção de árvores de classificação* $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ *onde* $\{\Theta_k\}$ *são vetores aleatórios identicamente distribuídos e cada árvore atribui um único voto para a classe mais frequente referente ao vetor de entrada* $\mathbf{x}$*. (BREIMAN, 2001)*

Há um teorema em Breiman (2001) que mostra que o limite superior para o erro de generalização de RF converge para a Expressão 4.14, quando o número de árvores for

[3] $\left[\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e \approx 2,718.\right]$

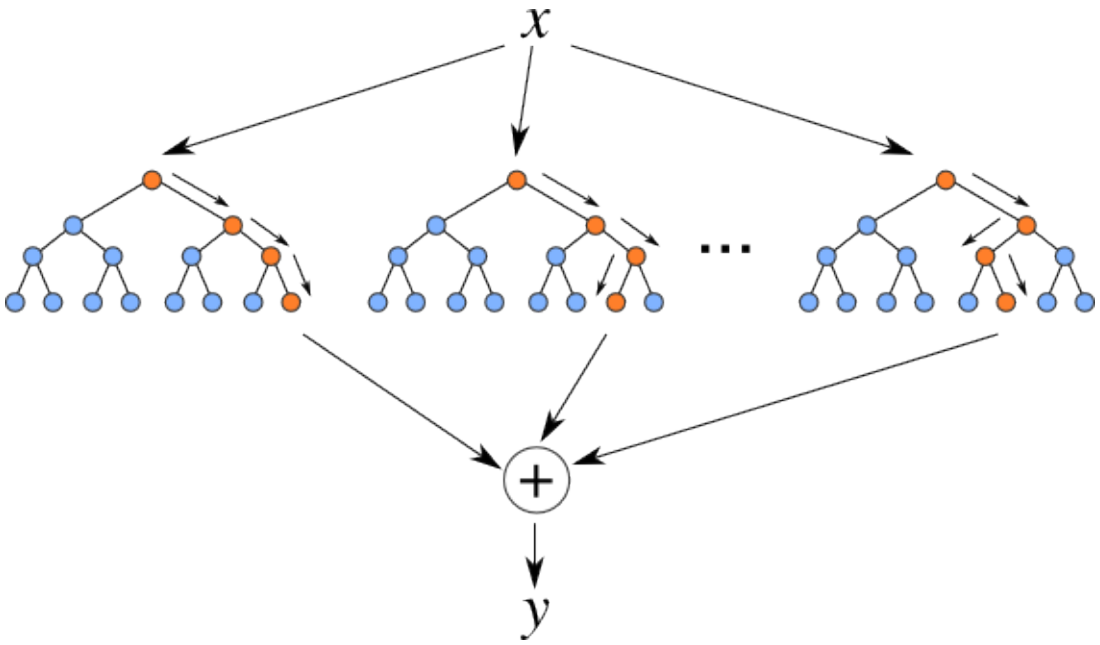


Figura 4.9 Exemplo da predição de uma instância x aplicada à cada árvore da RF. Extraído de Goldstein, Polley e Briggs (2011).

suficientemente grande.

$$\text{Erro de generalização} \leq \frac{\overline{\rho}(1-s^2)}{s^2} \tag{4.14}$$

Onde $\overline{\rho}$ é a correlação média entre as árvores e s é a quantidade que mede a "força" dos classificadores formados por árvores. A Expressão 4.14 é bastante intuitiva, pois à medida que as árvores se tornam mais correlacionadas ou a "força" do conjunto diminui, o limite do erro de generalização tende a aumentar. A aleatoriedade possibilita reduzir a correlação entre árvores de maneira que o erro de generalização possa ser reduzido. A "força" é definida como o valor esperado da margem do classificador e é mostrada na Expressão 4.15

$$s = E_{\mathbf{X},Y} mr(\mathbf{X}, Y) \tag{4.15}$$

A "força" de um conjunto de classificadores se refere ao desempenho médio dos classificadores, onde o desempenho é medido probabilisticamente em termos da margem do classificador (TAN; STEINBACH; KUMAR, 2006). A Expressão 4.16 indica como é

calculada a margem de um classificador.

$$mr(\mathbf{X}, Y) = P_{\Theta}(h(\mathbf{X}, \Theta) = Y) - max_{j \neq Y}(P_{\Theta}(h(\mathbf{X}, \Theta) = j)) \tag{4.16}$$

Onde $h(\mathbf{X}, \Theta)$ é a classe prevista para $\mathbf{X}$ de acordo com um classificador construído a partir do vetor aleatório Θ. Quanto maior for a margem, mais provável será a previsão correta para um determinado exemplo $\mathbf{X}$.

O algoritmo de RF começa pela seleção de uma amostra aleatória de inicialização dos dados. O subconjunto aleatório de variáveis é selecionado e procura até encontrar a variável ideal para ser o nó de divisão. Este processo é repetido para cada nó até que uma árvore seja formada, mas não podada como no algoritmo para árvores de decisão denominado CART. Os dados que não fazem parte da amostra de *bootstrap* são aplicados em cada árvore para derivar a taxa de erro denominada *out-of-bag* (OOB) e medidas de importância de variável (VI). Este processo é repetido até que uma floresta é completa com o número de árvores especificado inicialmente. Após treinar k árvores de decisão (classificadores), uma observação de teste é atribuída à classe que recebe o maior número de votos. Em problemas de regressão, a média dos valores preditos por cada árvore é a predição da RF. O Algoritmo 1 que implementa o procedimento completo de construção para uma RF foi proposto formalmente por Breiman (2001) e está demonstrado pela Figura 4.10.

Algoritmo 1 Random Forest ($ntree$, $mtry$)

```
para b ← 1 ... ntree faça
    selecionar uma amostra bootstrap;
    repita
        selecionar aleatoriamente mtry variáveis;
        encontrar o resultado das melhores partições;
    até formar a árvore;
    predizer Y para OOB;
    predizer Y para X permutado;
fim
calcular o erro OOB;
calcular a importância das variáveis;
```

Como citado por Breiman (2001), o erro de generalização (ou erro de predição) de uma RF depende da força das árvores individuais na floresta e da correlação entre elas. Para melhorar a precisão, a aleatoriedade foi injetada na RF para minimizar a correlação $\overline{\rho}$ mantendo ao mesmo tempo a "força" designada por s (BREIMAN, 2001). As florestas estudadas aqui consistem de usar aleatoriamente entradas ou combinações de entradas selecionadas em cada nó para crescer cada árvore. As florestas resultantes apresentam

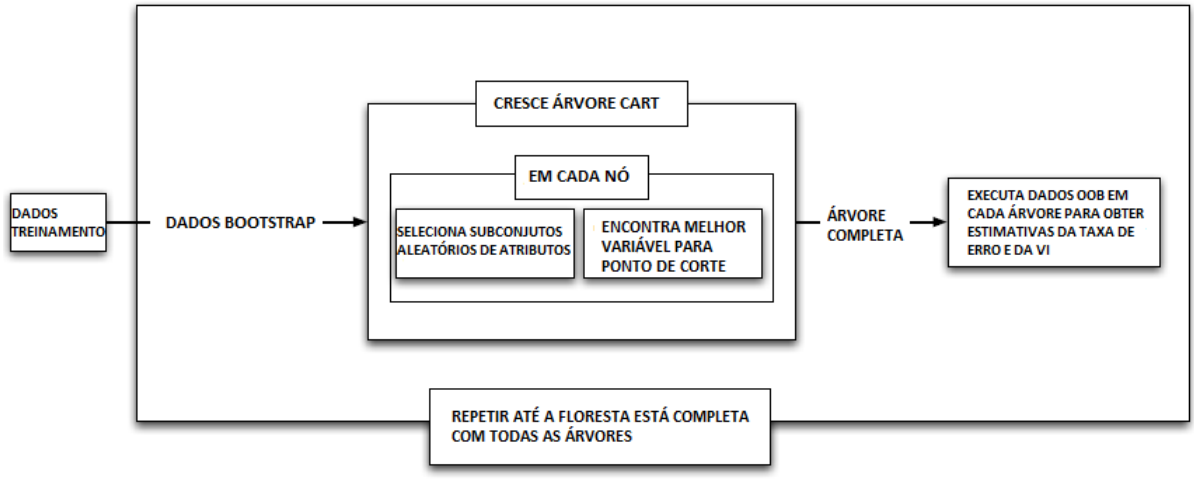


Figura 4.10 Visão geral do algoritmo para construção de uma RF.
Extraído de Goldstein, Polley e Briggs (2011).

precisão que comparam favoravelmente com *AdaBoost*. Segundo Breiman (2001), esta classe de procedimentos tem características desejáveis:

(i) sua precisão é tão boa quanto *AdaBoost* e às vezes melhor;

(ii) é relativamente robusto para valores aberrantes (*outliers*) e ruído;

(iii) é mais rápido do que o *Bagging* ou *Boosting*;

(iv) dá estimativas internas úteis de erro, de força, de correlação e importância variável;

(v) é simples e facilmente paralelizado.

As medidas mais comuns para o cômputo da importância da variável (do inglês, *Variable Importance* - VI) são: importância de permutação (pVI) e importância de gini (gVI).

A importância de permutação é o aumento no erro de classificação para instância OOB i após a variável j ser permutada na árvore k. Com o objetivo de definir formalmente pVI, necessita-se especificar as seguintes medidas:

- s_{ijk} é o número de árvores que usam a variável j em algum nó e erram na observação i.

- r_{ijk} é o número de árvores que não usam a variável j em algum nó e erram na observação i.

- ps_{ijk} é o número de árvores que usam a variável j e erram na observação i quando a variável j é permutada.

- pr_{ijk} é o número de árvores que não usam a variável j em algum nó e erram na observação i quando a variável j é permutada.

Assim, pVI_{ijk} é definida pela Expressão 4.17.

$$pVI_{ijk} = (ps_{ijk} + pr_{ijk}) - (s_{ijk} + r_{ijk}) = ps_{ijk} - s_{ijk} \text{ desde que } pr_{ijk} = r_{ijk} \tag{4.17}$$

Daí, calculam-se pVI_{ij}, pVI_{jk} e pVI_j pela Expressão 4.19.

$$\begin{aligned} pVI_{ij} &= \frac{1}{ntree}\sum_{k=i}^{ntree}(ps_{ijk} - s_{ijk}) \\ pVI_{jk} &= \frac{1}{np}\sum_{i=i}^{np}(ps_{ijk} - s_{ijk}) \\ pVI_{j} &= \frac{1}{np \times ntree}\sum_{i=i}^{np}\sum_{k=i}^{ntree}(ps_{ijk} - s_{ijk}) \end{aligned} \tag{4.18}$$

Os parâmetros np e $ntree$ são o número de instâncias OOB e o número de árvores da floresta respectivamente. Como analisa Goldstein, Polley e Briggs (2011), pVI tem propriedades interessantes, pois como esse indicador é calculado para as amostras OOB, ele pode ser visto como a qualidade preditiva da variável. Então se uma variável não tem importância para predição, $E(pVI) = 0$, ou seja, a permutação não aumentaria e nem diminuiria o erro de classificação.

Diferentemente do pVI, o gVI é somente aplicado para classificação. O índice de gini (GI) é o critério usado para crescer as árvores na RF para classificação. Sua fórmula é dada pela Expressão 4.19 para problemas de classificação com somente duas classes.

$$GI = 2p(1-p) \tag{4.19}$$

Onde p é a proporção na segunda classe. A quebra que minimiza GI é a preferida. Seja n o índice para um nó em uma dada árvore, assim, a Expressão 4.20 define as medidas gVI_{jkn} (importância da variável j no nó n na árvore k), gVI_{jk} (importância da variável

j na árvore k) e gVI_j (importância da variável j).

$$
\begin{aligned}
gVI_{jkn} &= (GI_{pai} - GI_{\text{filho esquerda}} + GI_{\text{filho direita}})np_{kn} \\
gVI_{jk} &= \frac{1}{np}\sum_{n_j \in Tree_k}^{N} gVI_{jkn} (\text{somando sobre os nós que contêm a variável } j \text{ na árvore } k) \\
gVI_j &= \frac{1}{ntree}\sum_{k=1}^{ntree} gVI_{jk}
\end{aligned} \tag{4.20}
$$

Quanto maior o valor de gVI, melhor a variável foi em dividir os dados (GOLDSTEIN; POLLEY; BRIGGS, 2011). Diferentemente do pVI, o gVI não possui a noção de qualidade preditiva no conjunto de teste OOB. Em vez disso, gVI_{jkn} pode ser avaliada como um teste χ^2, subordinada ao que já ocorreu na árvore (o nó raiz não é condicional a nenhuma variável)(GOLDSTEIN; POLLEY; BRIGGS, 2011).

Outra propriedade é que $gVI_j \geq 0$, com a igualdade se a variável j não aparece em qualquer árvore (GOLDSTEIN; POLLEY; BRIGGS, 2011). Desde gVI é calculado com base nos dados dentro da amostra de treinamento não tem uma interpretação no nível da população como pVI(GOLDSTEIN; POLLEY; BRIGGS, 2011). O gVI considera apenas a relação entre a variável e o modelo (GOLDSTEIN; POLLEY; BRIGGS, 2011).

Goldstein, Polley e Briggs (2011) argumentam que a medida pVI é mais comumente usada que gVI, porém, quando o erro na amostra OOB é aproximadamente 50%, ou seja, a qualidade de predição é baixa, gVI pode ser preferível. Como pVI é calculado com base no aumento do erro de classificação após permutar a variável j, se a taxa de erro de classificação base já é baixa, há pouca probabilidade para a permutação tornar a predição pior (GOLDSTEIN; POLLEY; BRIGGS, 2011). Isso faz com que o pVI seja uniformemente baixa. Por outro lado, uma vez que gVI é calculado em relação à árvore crescida, não sofre deste problema.

Em comparação com as árvores de decisão e de regressão, o que se perde com as RF é a estrutura simples e interpretável e o que se ganha com elas é um aumento de precisão devido à agregação de classificadores pelo *Bagging* (FACELI et al., 2011). Mas, se o objetivo for a seleção de atributos, não há necessidade primária de uma estrutura que permita interpretação explícita, porém, que consiga capturar vários tipos de relações não-lineares.

4.2.4 Máquina de Vetores Suporte (Support Vector Machine-SVM)

O SVM é uma técnica de aprendizado supervisionado que analisa padrões entre os dados de entrada, caraterizados por variáveis numéricas contínuas ou discretas, com os dados de saída, designados por um atributo dicotômico (problema de classificação). Esse modelo foi desenvolvido por Cortes e Vapnik (1995) e é baseado na ideia de encontrar o hiperplano ótimo que separa as duas classes por meio da maximização da margem de separação das classes consideradas. A Figura 4.11 mostra um exemplo de classificação perfeita baseado no SVM linear.

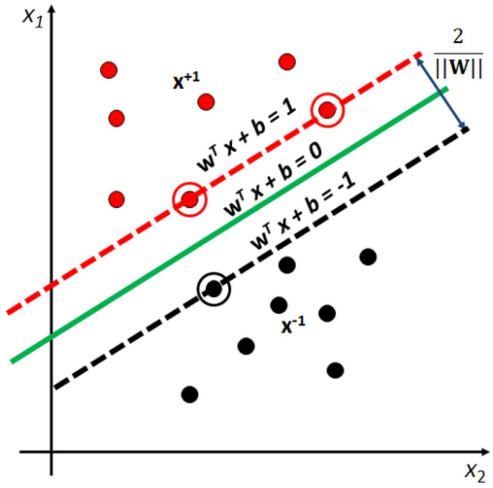


Figura 4.11 Classificação perfeita pelo hiperplano ótimo do SVM de margens rígidas. Extraído de Liu et al. (2013).

Como o modelo matemático do SVM linear é um problema de programação quadrática, então a solução ótima encontrada é um ótimo global, ou seja, repetindo o algoritmo um número n de vezes, chega-se ao mesmo resultado, diferentemente do que ocorre com redes neurais, as quais podem convergir para diferentes soluções ótimas locais em cada execução.

Posteriormente, o SVM foi adaptado para permitir a flexibilização na classificação, de forma que fosse possível a inclusão de elementos, originalmente de uma classe, classificados na outra classe. Nessa situação, tem-se o SVM linear com margens flexíveis (Figura 4.12).

Conforme a referência Cristianini e Shawe-Taylor (2000), aplicações complexas no mundo real requerem hipóteses mais expressivas que espaços de funções lineares. Deste modo, outra forma de avaliar este problema é que frequentemente a variável explicada não pode ser expressa como uma simples combinação linear dos atributos

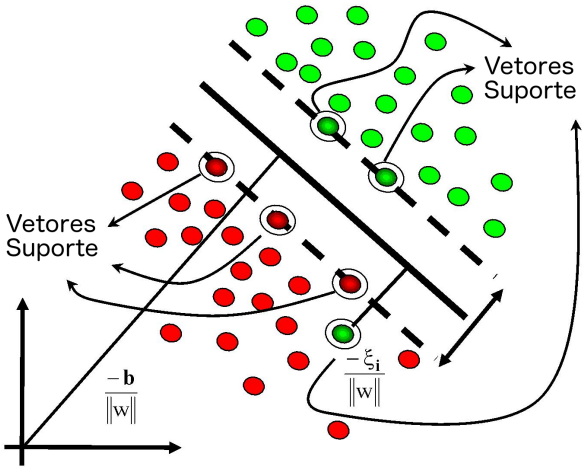


Figura 4.12 Classificação imperfeita pelo hiperplano ótimo do SVM de margens flexíveis. Extraído de Chandrasekhar e Raghuveer (2013).

considerados, mas, geralmente, requer que mais características abstratas dos dados sejam exploradas(CRISTIANINI; SHAWE-TAYLOR, 2000).

Por outro lado, para generalizar o SVM em problemas de classificação não-linear e com margens flexíveis, as funções *kernel* (Definição 2) foram inseridas no mesmo. Assim, as variáveis iniciais de entrada são mapeadas pela função *kernel* para o espaço de características, com dimensão superior ao espaço de entrada inicial. Tal fato aumenta a probabilidade de separação linear no espaço de características. Essa explicação é ilustrada na Figura 4.13 para facilitar o entendimento desse processo.

Definição 2. *[Função kernel] Um kernel é uma função K tal que para todo $\mathbf{x}, \mathbf{z} \in X$ satisfaz $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ onde ϕ é uma função de X para um espaço de características com produto interno F, onde $\phi : \mathbf{x} \mapsto \phi(\mathbf{x}) \in F$.*

Para esclarecer o conceito de função *kernel*, segue um exemplo extraído de Shawe-Taylor e Cristianini (2004): considere um espaço de entrada $X \subseteq \mathbb{R}^2$ juntamente com a função de característica $\phi : \mathbf{x} = (x_1, x_2) \mapsto \phi(\mathbf{x}) = \phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in F = \mathbb{R}^3$.

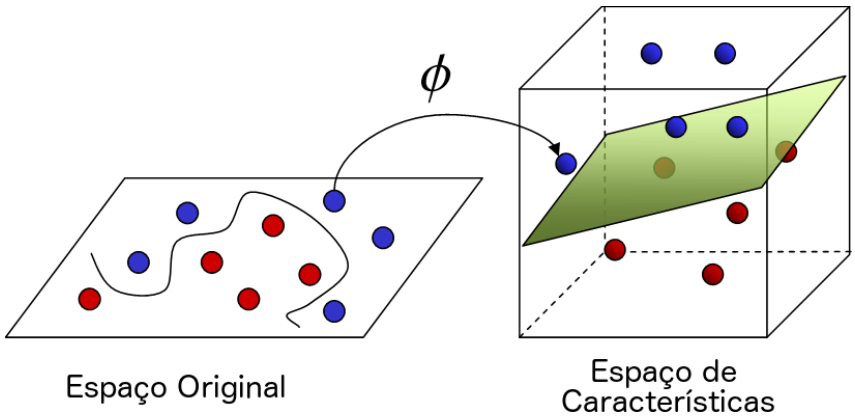


Figura 4.13 Classificação perfeita pelo hiperplano ótimo do SVM com *kernel* não-linear. Extraído de Karargyris e Bourbakis (2011).

A hipótese do espaço de funções lineares em F poderia ser:

$$\begin{aligned}\langle\phi(\mathbf{x}),\phi(\mathbf{z})\rangle &= \\ \langle(x_1^2,x_2^2,\sqrt{2}x_1x_2),(z_1^2,z_2^2,\sqrt{2}z_1z_2)\rangle &= \\ x_1^2z_1^2+x_2^2z_2^2+2x_1x_2z_1z_2 &= \\ (x_1z_1+x_2z_2)^2 &= \\ \langle\mathbf{x},\mathbf{z}\rangle^2 &\end{aligned}$$

Logo, a função $K(\mathbf{x},\mathbf{z})=\langle\mathbf{x},\mathbf{z}\rangle^2$ é uma função *kernel* com seu espaço de característica $F=\mathbb{R}^3$. De acordo com Shawe-Taylor e Cristianini (2004), isto significa que pode-se computar o produto interno entre as projeções de dois pontos no espaço de características sem explicitamente calcular suas coordenadas.

Com a adaptação adequada, o SVM pode tratar problemas de regressão, permitindo resolver um leque maior de problemas associados ao mapeamento entre variáveis de entrada explicativas e variáveis de saída explicadas contínuas, o que será abordado com mais detalhes a seguir.

4.2.5 Máquina de Vetores Suporte com Regressão (Support Vector Regression-SVR)

A primeira versão do SVM com regressão foi proposta em 1997 por Drucker et al. (1997), e foi denominada como Regressão com Máquina de Vetores Suporte (SVR - *Support Vector Regression*). Dentre as vantagens do SVR, vale citar que este método não pressupõe

linearidade do modelo, desde que se adote função *kernel* não-linear, não necessita de normalidade dos resíduos e adapta-se facilmente a dados de alta dimensionalidade (número de instâncias menor que o número de atributos).

Seja o conjunto de treinamento $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\} \subset (\mathcal{X} \times \mathbb{R})$ composto de n instâncias e d atributos, onde $\mathbf{x_i} \in \mathcal{X}$(sendo $\mathcal{X} \subset \mathbb{R}^d$) e $y_i \in \mathbb{R}$ para todo $i \in \{1, 2, ..., n\}$. O conjunto $\mathcal{X}$ denota o espaço com os padrões de entrada e y_i é a variável de saída a ser predita, que neste caso é contínua para todo i. Conforme a referência Vapnik (1995), na regressão baseada no SVR, o objetivo é encontrar uma função f que tem desvio máximo ε dos alvos y_i efetivamente obtida para todos os dados de treino, e ao mesmo tempo é tão plana quanto possível. Em outras palavras, não se preocupam com os erros, enquanto eles são menores que ε, mas não é permitido qualquer desvio maior do que este.

Inicialmente, serão consideradas somente as funções lineares, as quais podem ser descritas pela Expressão 4.21 e que mapeiam linearmente as variáveis do espaço de entrada $\mathbb{R}^d$ na variável do espaço de saída $\mathbb{R}$.

$$f(x) = \langle w, x \rangle + b \tag{4.21}$$

A notação $\langle w, x \rangle = w_1 x_1 + w_2 x_2 + ... + w_d x_d$ denota o produto interno em $\mathcal{X}$. E, matematicamente, quando busca-se por uma função não-linear mais plana quanto possível, pretende-se reduzir sua complexidade, o que pode ser obtido pela minimização da norma do vetor w, isto é, $\|w\|^2 = \langle w, w \rangle$. Com $w, x \in \mathbb{R}^d$ e $b \in \mathbb{R}$, onde w e b significam, respectivamente, a inclinação e o intercepto do hiperplano a serem estimados a partir da otimização do modelo matemático constituído pela função objetivo, indicada pela Expressão 4.22, e pelo conjunto de restrições ilustrado em 4.23.

$$\text{Minimizar } Z(w, b) = \frac{1}{2}\|w\|^2 \tag{4.22}$$

sujeita às restrições:

$$\begin{cases} y_i - \langle w, x_i \rangle - b & \leq \quad \varepsilon \\ \langle w, x_i \rangle + b - y_i & \leq \quad \varepsilon \end{cases} \tag{4.23}$$

Salienta-se ainda, segundo Smola e Schölkopf (2004), que uma premissa no modelo expresso pela função objetivo 4.22 e pelas restrições 4.23 é que existe a função f que

aproxima todos os pares (x_i, y_i) com uma precisão ε, ou seja, o problema de otimização convexo é viável. Entretanto, algumas vezes, tal problema pode ser inviável, ou até mesmo, pode-se permitir alguns erros superiores à margem ε. Deste modo, para flexibilizar o modelo anterior a aceitar erros superiores ao desvio ε, são introduzidas as variáveis de folga ξ_i e ξ_i^*. Com isso, obtém-se a formulação proposta pela referência Vapnik (1995) denotada pelas Expressões 4.24 e 4.25.

$$\text{Minimizar } Z(w, b, \xi_i, \xi_i^*) = \left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right] \tag{4.24}$$

sujeita às restrições:

$$\begin{cases} y_i - \langle w, x_i\rangle - b & \leq \quad \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i & \leq \quad \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq \quad 0 \end{cases} \tag{4.25}$$

Outra forma de escrever a Expressão 4.24 é mostrada em 4.26.

$$\text{Minimizar}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} L_\varepsilon(f(x_i), y_i)\right] \tag{4.26}$$

Assim, a função de perda ε-insensível é definida como indicado em 4.27.

$$L_\varepsilon(f(x_i), y_i) = \begin{cases} 0 & \text{se } |y_i - f(x_i)| \leq \varepsilon; \\ |y_i - f(x_i)| - \varepsilon & \text{se } |y_i - f(x_i)| > \varepsilon. \end{cases} \tag{4.27}$$

Por outro lado, de acordo com a Expressão 4.26, o termo $\frac{1}{2}\|w\|^2$ indica a complexidade do modelo e o termo $L_\varepsilon(f(x_i), y_i)$ traduz a função de perda ε-insensível que penaliza somente os valores fora do tubo, ou seja, com erros maiores que ε. Já o parâmetro C é chamado de constante de regularização e traduz o equilíbrio entre a complexidade de f e a quantidade de desvios maiores do que ε que podem ser tolerados ((ÜNSTü; MELSSEN; BUYDENS, 2006)). Assim, quanto menor o tubo (menor ε), mais complexa é a função f e, de forma contrária, quanto maior o tubo (maior ε), menos complexidade é necessária para f. A função de perda ε-insensível com SVR linear é mostrada na Figura 4.14.

Conforme a referência Ünstü, Melssen e Buydens (2006), com a introdução de variáveis de folga ξ_i e ξ_i^* e devidas manipulações algébricas, as Expressões 4.22 e 4.23 se transformam na função objetivo 4.24 e nas restrições 4.25. Tal formulação é chamada de primal, pois a

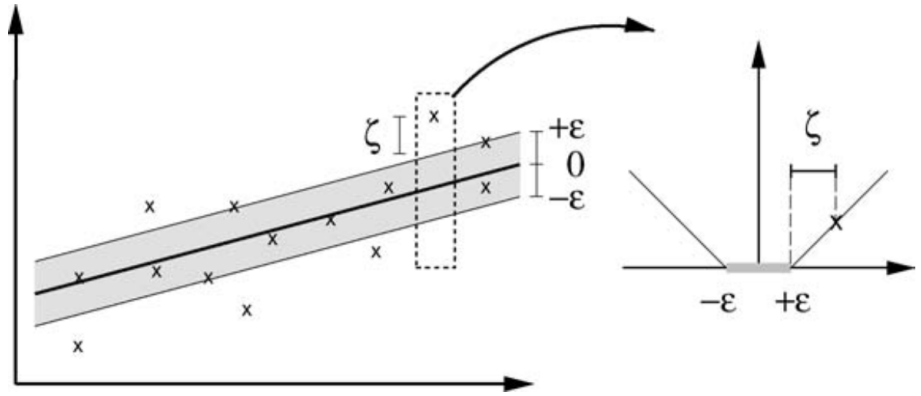


Figura 4.14 A função de perda com margem flexível com SVR linear.
Adaptado de Smola e Schölkopf (2004).

regressão é baseada no espaço original dos dados. Já as variáveis de folga têm por objetivo possibilitar a ocorrência de vetores fora do tubo, sendo os mesmos chamados vetores suporte, pois são somente eles que contribuem para a regressão. Desta forma, todos os outros vetores dentro do tubo podem ser removidos após a construção do modelo. Essa propriedade permite que o SVR modele relações em que o número de variáveis dependentes seja superior ao número de instâncias na amostra de treinamento.

No caso dos padrões de entrada x_i não possuírem relação linear com a variável dependente designada pelos valores y_i, a função f do modelo primal é reformulada para o modelo dual como mostra a Equação 4.28. Com isso, o espaço original é mapeado para um novo espaço, denominado espaço de características, por meio da função ϕ e do produto interno $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, sendo K chamada de função *kernel*. Esta função traduz a relação subjacente entre os dados de entrada e os dados de saída.

$$f(x) = \left[\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x_j) \rangle \right] + b \tag{4.28}$$

As variáveis duais α_i e α_i^* representam os multiplicadores de Lagrange que satisfazem as desigualdades e que podem ser obtidos pela Expressão 4.29 e pela Equação 4.30.

$$\begin{aligned} \text{Maximizar } Q(\alpha_i, \alpha_i^*) = \Bigg\{ & -\frac{1}{2} \left[\sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) \right] \\ & - \varepsilon \left[\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \right] + \left[\sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \right] \Bigg\} \end{aligned} \tag{4.29}$$

sujeita às restrições:

$$\begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) &= 0 \\ \alpha_i, \alpha_i^* &\in [0, C] \end{cases} \tag{4.30}$$

Isto é chamado de expansão de vetores suporte, ou seja, w pode ser determinado por uma combinação linear dos padrões de treinamento x_i. Com essa observação, conclui-se que a representação da complexidade de uma função por vetores suporte é independente da dimensionalidade do espaço de entrada $\mathcal{X}$, mas depende somente do número de vetores suporte (SMOLA; SCHÖLKOPF, 2004).

Uma primeira forma para tratar do SVR não linear seria realizar um pré-processamento dos dados de entrada para o espaço de características a partir da função $\phi : \mathcal{X} \mapsto F$, e, em seguida, aplicar o SVR linear padrão nos dados transformados. Com isso, a linearidade da regressão é obtida no espaço de características e não no espaço original como pode ser notado na Figura 4.15. Entretanto, esses dois passos em problemas com uma grande quantidade de dados de treinamento torna o SVR computacionalmente inviável. Para superar essa dificuldade, são escolhidos funções *kernel* que podem ser escritas em função dos dados de treinamento, logo, o cálculo do produto interno entre os vetores transformados do espaço de característica não é mais necessário. Esse cômputo é feito implicitamente pela função *kernel* baseando-se nos dados de treinamento, o que é feito em um único passo.

A função *kernel* linear é calculada a partir da Equação 4.31, sendo C e ε os únicos parâmetros dessa função *kernel*. De maneira geral, esse *kernel* é usado como modelo controle para comparação com outros tipos de *kernels*.

$$K(x_i, x_j) = \langle x_i, x_j \rangle \tag{4.31}$$

A função *kernel* baseada na função de base radial (RBF - *Radial Base Function*) é um *kernel* de propósito geral quando não se tem conhecimento *a priori* sobre os dados (KARATZOGLOU; SMOLA; HORNIK, 2004). Esse *kernel* é computado pela Equação 4.32 e possui o parâmetro γ que deve ser escolhido adequadamente *a priori* da

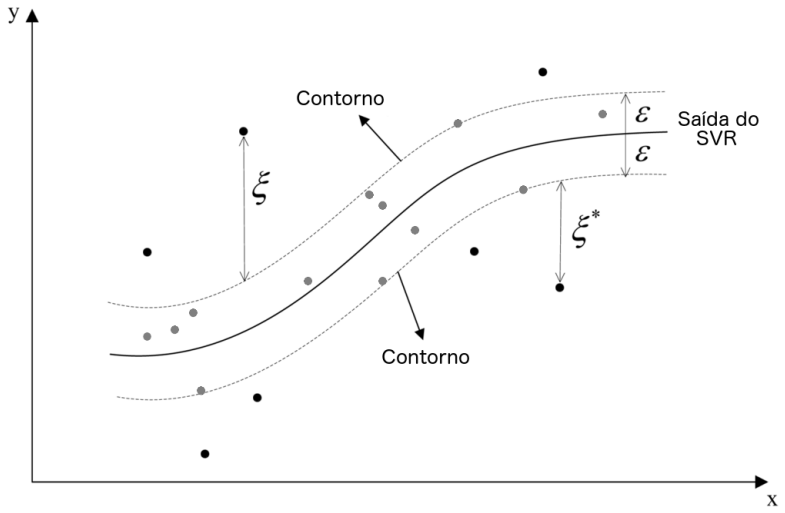


Figura 4.15 Regressão com *kernel* não linear com função de perda ε-insensível, onde os círculos em preto são os vetores suportes.
Adaptado de Ma, Song e Xiao (2012).

mesma forma que os parâmetros C e ϵ.

$$K(x_i, x_j) = exp(\gamma \|x_i - x_j\|^2) \tag{4.32}$$

Outra função *kernel* é o Pearson VII Universal *kernel* (PUK) explicitado pela Expressão 4.33. Ünstü, Melssen e Buydens (2006) mostrou que o PUK é alterado facilmente, adaptando seus dois parâmetros σ e ω, entre as formas das funções Gaussiana e Lorentziana, e até mesmo a outras funções.

$$K(x_i, x_j) = \frac{1}{\left[1 + \frac{2\sqrt{\|x_i - x_j\|^2}\sqrt{2^{1/\omega} - 1}}{\sigma}\right]} \tag{4.33}$$

Como sugerido pela referência Shawe-Taylor e Cristianini (2004), os passos para a execução da abordagem do SVR baseada em *kernel* são:

(i) os dados são incorporados em um espaço vetorial chamado espaço de características;

(ii) as relações lineares são procuradas entre as imagens dos vetores do espaço original, as quais pertencem ao espaço de características;

(iii) os algoritmos são implementados de tal forma que as coordenadas dos pontos incorporados não são necessários, apenas o resultado do produto interno entre todos os pares;

(iv) os produtos internos emparelhados podem ser calculados eficientemente a partir dos dados originais usando uma função *kernel*.

Caso os parâmetros do SVR/SVM sejam otimizados, tem-se uma efetividade no ajuste do modelo aos dados e na generalização das predições, como é discutido em Ünstü, Melssen e Buydens (2006).

4.2.6 Algoritmos Genéticos

Algoritmos Genéticos (do inglês, *Genetic Algorithm* - GA) são algoritmos de busca baseados em processos genéticos e seleção natural (GOLDBERG, 1989). Ele simula a evolução através de três operadores genéticos: seleção, *crossover* e mutação. Uma população inicial é gerada com tamanho pré-estabelecido e, em seguida, é realizada a seleção de indivíduos que participarão no processo de reprodução. Após a seleção, é aplicado o *crossover* e, sucessivamente, a reprodução é realizada, criando uma nova população. Subsequentemente, a mutação é aplicada em alguns dos descendentes gerados. Os operadores de *crossover* e mutação são baseados em probabilidades, logo, nem toda população cruzará e/ou sofrerá mutação.

A Figura 4.16 indica o funcionamento geral de um algoritmo genético. Inicialmente, é necessário definir os tipos de variáveis e a codificação das mesmas. Com isso, constrói-se a função de aptidão, que é a função objetivo a ser otimizada, além de atribuir um mérito pelo desempenho de cada indivíduo do GA. Os operadores genéticos, *crossover* e mutação, são aplicados estocasticamente em cada passo do processo evolutivo, a partir das respectivas probabilidades definidas pelo usuário. Por último, um critério de parada para o GA deve ser definido, que pode ser o número máximo de gerações, ou o número máximo gerações sem melhoria da aptidão do melhor indivíduo, ou um erro máximo estipulado.

Algoritmos Genéticos podem ser usados para seleção de variáveis, otimizando alguma medida relativa à relevância dessas variáveis, tanto em problemas de classificação como em problemas de regressão. Sua grande vantagem é que ele executa uma pesquisa "inteligente" no sentido de caminhar na direção de "bons" candidatos, garantindo uma melhoria constante, mas mantendo alguma variabilidade para aumentar a possibilidade de encontrar a solução ótima e não parar em mínimos ou máximos locais.

O GA parte de uma amostra do espaço de soluções e caminha para soluções informativas, o que elimina a necessidade de avaliar todas as soluções candidatas dentro

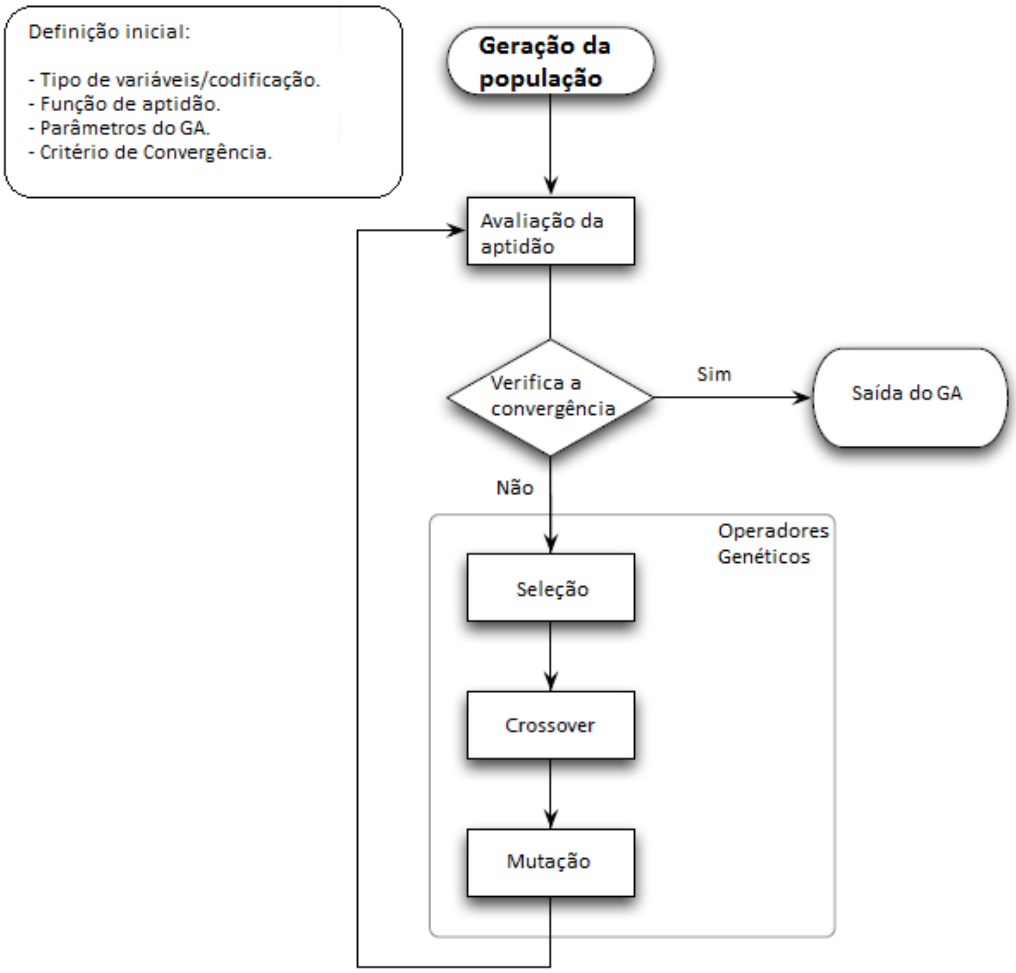


Figura 4.16 Fluxograma do GA.
Adaptado de Scrucca (2012).

do domínio especificado. Como exemplo, suponha que o objetivo é selecionar o "melhor" subconjunto de variáveis explicativas para um modelo preditivo baseado em uma base de dados inicial contendo 300 variáveis explicativas. Desta forma, o número total de soluções avaliadas por um algoritmo baseado em busca exaustiva será o número de combinações com uma variável mais o número de combinações com duas variáveis mais o número de combinações com três variáveis e, assim sucessivamente, até o número de combinações com 300 variáveis perfazendo um total de $2^{300} - 1 \approx 2,03704 \times 10^{90}$ [4] soluções avaliadas, o que é computacionalmente proibitivo.

[4]Para qualquer n natural tem-se que $C_0^n + C_1^n + C_2^n + ... + C_n^n = 2^n$ onde C_k^n significa o número de combinações de n elementos tomados k a k, ou seja, é o número total de subconjuntos de k elementos construídos a partir de um conjunto com n elementos.

4.3 Resumo do Capítulo

A técnica de validação de cruzada usada em problemas de classificação e regressão foi apresentada como principal artifício para generalizar os resultados do processo de indução do aprendizado pelo SVM/SVR, o que diminui a captura de atributos espúrios para a variável explicada. As árvores de decisão e de regressão foram explicadas, bem como suas vantagens e desvantagens. A técnica da RF foi apresentada e discutida, mostrando que a mesma pode ser usada tanto para predição quanto para seleção de atributos a partir das medidas de importância de variável que desdobram naturalmente de sua estrututra. As formulações matemáticas do SVM e o do SVR foram feitas e a possibilidade de captura de não-linearidades nos atributos foi destacada pelo uso de *kernels* não-lineares. Em relação aos Algoritmos Genéticos, discutiu-se que os mesmos são processos robustos de buscas não-exaustivas que, em geral, conseguem encontrar os mínimos e máximos globais de diversas funções multivariadas, além de serem adaptáveis para a seleção de atributos.

5 Métodos para Seleção de Atributos

Neste capítulo, os métodos de seleção de atributos serão apresentados e categorizados no intuito de contextualizar o método de seleção de SNPs desenvolvido neste estudo. A classificação desses métodos serão feitas do ponto de vista de Aprendizado de Máquina como métodos baseados em filtros, métodos *wrapper* e métodos embutidos. Outras discriminações usadas para os modelos utilizados em GWAS são referentes à relação estatística entre genótipo e fenótipo e às distribuições de probabilidade do genótipo, do fenótipo e do erro.

5.1 Introdução

Em muitas tarefas de classificação, o número total de possíveis atributos que podem ser empregues é relativamente alto (STAŃCZYK; JAIN, 2015). Caso todos sejam usados, resultaria no problema de alta dimensionalidade, o que dificulta o processamento, ou até mesmo, tornando-o impraticável. Além de que, a presença de muitas variáveis é uma desvantagem para a maioria dos indutores mesmo quando elas são relevantes para a tarefa de classificação, para não mencionar variáveis irrelevantes ou redundantes que podem obscurecer outros padrões (JOHN et al., 1994).

De maneira geral, os métodos para seleção de atributos podem ser agrupados em três grandes classes: métodos baseados em filtro, métodos *wrapper* (cuja tradução livre é invólucro) e métodos embutidos (do inglês, *embedded*). O método baseado em filtro tenta avaliar o mérito dos atributos sem considerar qualquer indutor em particular (KOHAVI; JOHN, 1998). Métodos *wrapper* avaliam subconjuntos a partir de algum algoritmo de indução do aprendizado, entretanto, esse algoritmo indutor não é usado durante a busca pelos subconjuntos, sendo considerado uma caixa-preta (KOHAVI; JOHN, 1998). Esses métodos permitem, diferentemente das abordagens de filtro, detectar possíveis interações entre as variáveis (PHUONG; LIN; ALTMAN, 2005). Os métodos embutidos usam seu próprio algoritmo de aprendizado para selecionar os atributos tais como as árvores de decisão e as RF discutidas no capítulo anterior.

Os métodos de seleção baseados em filtros trabalham de forma independente do

classificador envolvido no reconhecimento de padrões, independentemente das suas especificidades e de seus parâmetros (GUYON; ELISSEEFF, 2003). Essas abordagens podem ser tratadas como procedimentos de pré-processamento. Elas exploram informações contidas no conjunto de dados de entrada buscando atributos que gerem ganho de informação, entropia ou consistência (DASH; LIU, 2003). Além disso, são particularmente efetivos em tempo computacional e robustos em relação ao *overfitting* (HAMON, 2013). A natureza geral dos filtros torna-os aplicáveis em todos os casos, no entanto, o fato deles ignorarem totalmente o desempenho de um sistema de classificação que emprega o conjunto de variáveis selecionadas causa resultados tipicamente piores do que outras abordagens e isto é considerado como uma desvantagem (STAŃCZYK; JAIN, 2015).

No método *wrapper*, argumenta-se que a melhor avaliação de alguns subconjuntos de variáveis candidatas é obtida, verificando sua utilidade na classificação, pois a precisão da previsão estimada é geralmente considerada o mais importante indicador de relevância para os atributos (KOHAVI; JOHN, 1997). Nessa abordagem um algoritmo de aprendizagem é usado para medir a qualidade de subconjuntos de variáveis sem incorporação de conhecimento sobre a estrutura específica da função de classificação ou de regressão, e, portanto, podem ser combinados com qualquer máquina de aprendizagem (LAL et al., 2006). Uma vez que o processo de pesquisa e seleção é ajustado à característica específica do indutor, que pode mostrar um viés, resultando num aumento para o desempenho do classificador escolhido, mas piores resultados para outro, especialmente quando eles variam significativamente em propriedades (STAŃCZYK; JAIN, 2015). Em outras palavras, *wrappers* tendem a construir conjuntos de atributos que são personalizados, feito sob medida para alguma tarefa especial e/ou algum sistema particular (STAŃCZYK; JAIN, 2015). Outra desvantagem desta abordagem é de custos computacionais necessários (STAŃCZYK; JAIN, 2015). A execução do algoritmo de aprendizado para muitos subconjuntos pode se tornar inviável, não só quando há um número muito elevado de atributos a considerar, mas também nos casos em que o processo de aprendizado é complexo e consome tempo mesmo para um número pequeno de variáveis (STAŃCZYK; JAIN, 2015). Por exemplo redes neurais artificiais lidam muito melhor com mais instâncias do que para situações quando seu número é baixo (STAŃCZYK; JAIN, 2015).

Métodos embutidos diferem de outros métodos no modo como a seleção de variáveis e o aprendizagem interagem (LAL et al., 2006). Como exemplo, Weston et al. (2000) mede a importância de um atributo usando um limite que é válido somente para SVM, portanto, não é possível usar esse método com, por exemplo, árvores de decisão.

Há também combinações de abordagens, onde, por exemplo, em primeiro lugar um filtro é utilizado, em seguida, um *wrapper*, ou quando um *wrapper* é utilizado como um filtro (STAŃCZYK; JAIN, 2015). Também é possível aplicar algum algoritmo para obter uma ordenação de atributos inicial, a qual será base para seleção ou redução de atributos que poderá ser executada posteriormente (STAŃCZYK; JAIN, 2015).

5.2 Métodos Paramétricos em GWAS

5.2.1 Métodos baseados no valor-p

5.2.1.1 Regressão Linear Simples

A regressão linear simples é um método para construir a relação linear estatística entre duas variáveis, ou seja, considerando a presença de ruídos nos dados, diferentemente da função linear perfeita definida matematicamente (NETER et al., 1996). Segundo Gujarati (2006), a análise de regressão se ocupa da dependência de uma variável, a(s) variável(is) dependente(s), em relação a uma variável, a variável independente ou explanatória, com vistas a estimar ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas. O modelo de regressão linear simples baseado na população é dado pela Expressão 5.1, onde x é a variável independente (explanatória ou explicativa ou preditora), y é a variável dependente (explanada ou explicada ou predita), ϵ é o termo de erro estocástico, β_0 e β_1 são os coeficientes linear e angular da reta ajustada para população.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{5.1}$$

Devido às diversas restrições de custo, tempo, acesso entre outras, a maioria dos estudos que utilizam regressão não são populacionais, mas baseados em amostras, logo, a Expressão 5.1, torna-se o modelo de regressão amostral indicado pela Expressão 5.2, com

i variando de 1 a n (tamanho da amostra).

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \tag{5.2}$$

A ideia central do modelo de regressão linear simples é encontrar estimativas amostrais para os coeficientes β_0 e β_1 denotadas por $\hat{\beta}_0$ e $\hat{\beta}_1$, minimizando o somatório dos quadrados dos erros amostrais dado por $\sum {\hat{\epsilon}_i}^2 = \sum (\hat{y}_i - y_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, ou seja, $\sum {\hat{\epsilon}_i}^2 = f(\hat{\beta}_0, \hat{\beta}_1)$, onde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ e $\hat{y}_i$ é o valor predito de y_i com base na equação da reta sem considerar o erro amostral $\hat{\epsilon}_i$. Isso significa que a função objetivo a ser minimizada varia em função dos coeficientes linear e angular da reta a ser ajustada para os dados da amostra. Após a aplicação do método dos mínimos quadrados, que permite encontrar o mínimo global de uma função quadrática em duas variáveis ($\hat{\beta}_0$ e $\hat{\beta}_1$ ótimos) para a função $\sum (\hat{y}_i - y_i)^2$, obtém-se a Expressão 5.3, e aplicando esse resultado na Expressão 5.1 chega-se à Expressão 5.4. Onde $\overline{x}$ e $\overline{y}$ representam as médias amostrais das variáveis x_i e y_i para todo i variando entre 1 e n.

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{(x_i - \overline{x})^2}. \tag{5.3}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}. \tag{5.4}$$

Para GWAS com fenótipo contínuo, um exemplo da relação linear estatística produzida pela regressão linear simples entre um SNP e o fenótipo com base no método dos mínimos quadrados ordinários é dado pela Figura 5.1. O genótipo do SNP é codificado como 0 para o homozigoto de referência AA, 1 para o heterozigoto Aa e 2 para o homozigoto aa, sendo A e a dois alelos possíveis. Desta forma, o $\hat{\beta}_0 = 271,8$ e $\hat{\beta}_0 = 438,5$. Em geral o alelo A é o mais frequente na população de estudo, logo, o alelo a é o menos frequente. Cabe destacar que o alelo A não é necessariamente dominante e o alelo a é a forma variante menos frequente, mas não é necessariamente o recessivo (NETO, 2013). Esse modelo pressupõe um efeito proporcional ao número de alelos a, ou seja, para $x = 0$, tem-se $y = 271,8$; para x=1, tem-se $y = 710,3$, e para $x = 2$, $y = 982,1$. Logo, a presença de um alelo a no genótipo do SNP, aumenta, em média, 271,8 no fenótipo em questão.

No contexto de estudos de associação em escala genômica, para um marcador SNP e o fenótipo de interesse, um modelo de regressão linear simples é construído pelo método

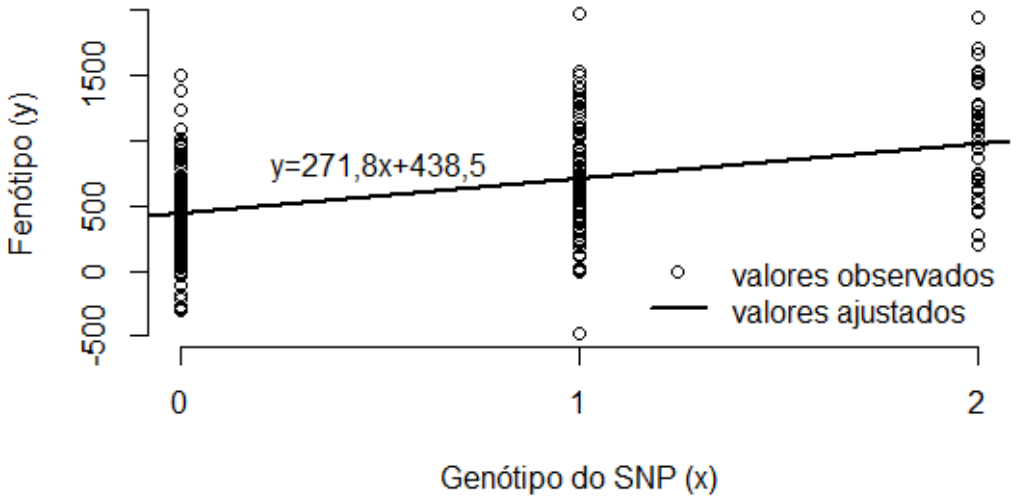


Figura 5.1 Gráfico dos fenótipos para uma amostra de 240 indivíduos em função do genótipo de um SNP e da reta de regressão ajustada pelo método dos mínimos quadrados.

dos mínimos quadrados ordinários e, então, o teste de hipóteses sobre o coeficiente β_1 é realizado para verificar se há evidências de que esse coeficiente é estatisticamente diferente de zero dado o nível de significância simbolizado por α. Assim, as hipóteses nula e alternativa são dadas pela Expressão 5.5.

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned} \tag{5.5}$$

O valor-p (do inglês, *p-value*), que significa valor da probabilidade, é a probabilidade de obter uma estatística de teste ou mais extrema do que aquela observada em uma amostra, presumindo a hipótese nula ser verdadeira (BIAU; JOLLES; PORCHER, 2010). Com isso, o valor-p dá aos pesquisadores uma medida para a evidência contra a hipótese nula (BIAU; JOLLES; PORCHER, 2010). Se o valor-p for igual ou menor que o nível de significância adotado *a priori*, onde geralmente $\alpha = 0,05$, então, os dados da amostra são inconsistentes com o pressuposto da hipótese nula, mostrando que há evidências de que a hipótese alternativa é verdadeira. Caso contrário, os dados da amostra corroboram com a hipótese nula.

O teste de hipótese geralmente usado após as estimativas dos betas serem calculadas é o teste-t cujas hipóteses nula e alternativa são dadas pela Expressão 5.5. A ideia fundamental por trás dos testes de hipóteses é a de um teste estatístico (estimador) e a distribuição amostral dessa estatística segundo a hipótese nula (GUJARATI, 2006). A decisão de aceitar ou rejeitar H_0 é tomada com base no valor-p do teste estatístico obtido a partir dos dados disponíveis (GUJARATI, 2006).

O valor-p associado ao teste de hipóteses dado por 5.5 pode ser usado como um método baseado em filtro para escolha dos SNPs mais informativos, pois não utiliza nenhum algoritmo indutor para aprendizagem. Para isso, basta ordenar os valores-p de dos SNPs de forma crescente e especificar um limite superior para escolher somente os marcadores que tenham valor-p menor ou igual ao limite. Logo, os coeficientes β_1 de todos os SNPs selecionados são significativamente diferentes de zero, o que indica uma associação estatística significativa com o fenótipo em questão. Além do mais, como são milhares de SNPs avaliados concomitantemente, são realizados múltiplos testes, o que implica na necessidade de aplicação de alguma correção para testes múltiplos. Nesse caso, pode-se utilizar a correlação de Bonferroni, multiplicando o valor-p bruto pelo número de SNPs (número de testes feitos), obtendo-se o valor-p corrigido. Após essa etapa, basta selecionar somente os SNPs cujos valores-p são menores que o α adotado. Esse método é o mais comum em GWAS para fenótipos contínuos devido à sua facilidade de aplicação e de interpretação.

5.2.1.2 Teste de Associação Qui-Quadrado

Uma importante aplicação do teste qui-quadrado ocorre quando o objetivo é verificar a associação entre duas variáveis (MARTINS, 2002). A representação das frequências observadas é construída a partir de uma tabela de dupla entrada denominada tabela de contingência.

Um evento B é dito independente de um evento A, se a probabilidade de B ocorrer não é influenciada pelo fato de A ter ou não ocorrido (LIPSCHUTZ, 1972). Equivalentemente, se a probabilidade de B é igual à probabilidade condicional de B dado que A ocorreu, o que é expresso matematicamente por $P(B) = P(B|A)$. Por definição, $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$, logo, se A e B são eventos independentes, $P(A \cap B) = P(A)P(B)$. Logo, para verificar se duas variáveis X e Y são associadas, é necessário avaliar se a distribuição conjunta de

probabilidades de X e Y é igual ao produto das distribuições marginais de X e Y, ou seja, $P(x_i, y_j) = P(x_i)P(y_j)$ para todo i e j.

O teste de hipótese do Qui-Quadrado construído com base nas hipóteses nula e alternativa são, respectivamente, H_0: as variáveis são independentes ou as variáveis não estão associadas e H_a: as variáveis são dependentes ou as variáveis estão associadas. O nível de significância α é fixado e o valor-p é calculado, com isso, se o mesmo é menor que α, existem indícios de que as variáveis são associadas. Caso o valor-p seja maior que α, então considera-se que as variáveis não estão associadas.

Da mesma forma que o valor-p baseado em regressão, o valor-p dos SNPs do teste qui-quadrado é usado como método de seleção baseado em filtro em problemas de seleção de SNPs em GWAS com fenótipo binário da mesma forma como o valor-p do coeficiente β_1 da regressão linear simples é utilizado em GWAS para fenótipo contínuo. Como existem milhares de SNPs avaliados simultaneamente, múltiplos testes Qui-Quadrado são realizados, é necessário aplicar algum tipo de ajuste, como, por exemplo, a correção de Bonferroni.

5.2.2 Lasso Bayesiano (Blasso)

Meuwissen et al. (2001) apresentaram dois métodos bayesianos denominados BayesA e BayesB como abordagens alternativas para a predição de valores genéticos genômicos em seleção genômica ampla (do inglês, *Genome-Wide Selection*-GWS). A possível vantagem desses métodos é assumir variâncias distintas para o efeito de cada marcador. Como Meuwissen et al. (2001) argumentam, os custos computacionais desses métodos podem ser elevados para conjuntos de dados com muitos marcadores SNP e, para contornar essa limitação, Meuwissen et al. (2001) propuseram o método Lasso Bayesiano (do inglês, *Bayesian Lasso* - Blasso) que é, de maneira geral, mais rápido computacionalmente do que BayesA e BayesB. O Blasso foi baseado em um tipo de regressão penalizada, proposta inicialmente por Tibshirani (1996), e adaptada para abordagem bayesiana por Park e Casella (2008). Esse método foi modificado e adaptado para seleção genômica ampla por Campos et al. (2009).

O modelo matemático do Blasso é mostrado na Expressão 5.6, onde y são os valores fenotípicos observados, μ é o efeito fixo estimado, a matriz M_{ij} é a matriz incidência dos SNPs, β_i são os efeitos a serem estimados para cada marcador, e são os resíduos,

n é o número de indivíduos na população, m é o número de marcadores SNP. A matriz $M_{ij} = [m_{ij}]_{m \times n}$, sendo $m_{ij} = 0$ ou -1 para homozigoto de referência (AA), $m_{ij} = 1$ ou 0 para heterozigoto (Aa) e $m_{ij} = 2$ ou 1 para homozigoto de variante (aa).

$$y = 1\mu + \sum_{i=1}^{m} M_{ij}\beta_i + e, \forall i \in \{1, ..., m\}, \forall j \in \{1, ..., n\}. \tag{5.6}$$

A função de verossimilhança é definida pela Expressão 5.7.

$$y|\mu, \beta_1, ..., \beta_m, \sigma_e^2 \text{ segue } N(1\mu + \sum_{i=1}^{m} X_{ij}\beta_i, I\sigma_e^2) \tag{5.7}$$

As Expressões 5.8, 5.9, 5.10 e 5.11 definem as distribuições *prioris* que são usadas juntamente com a função de verossimilhança dada em 5.7 no Teorema de Bayes para o cálculo das distribuições *posteriores*. O fator redutor do coeficiente β_i é o τ_i^2 e σ_e^2 é a variância do efeito a_i condicional ao redutor τ_i do marcador i.

$$\beta_i|\tau_i, \beta_i \text{ seguem } N(0, \underbrace{\tau_i^2\sigma_e^2}_{\sigma_i^2}) \tag{5.8}$$

$$\tau_i^2|\lambda^2 \text{ segue } Exp(\lambda^2) \tag{5.9}$$

$$\lambda^2 \text{ segue } Gamma(\alpha_1, \alpha_2) \tag{5.10}$$

$$\sigma_e^2 \text{ segue } \chi^{-2}(S, v) \tag{5.11}$$

$Exp(\lambda^2)$ simboliza a distribuição exponencial com parâmetro λ^2. A variância de cada marcador $V(\beta_i)$ é definida pela Expressão 5.12, logo, a variância conjunta dos m marcadores é dada pela Expressão 5.13. Finalmente, a herdabilidade (h^2) estimada pelos marcadores SNP é calculada como o percentual da variância explicada pelos SNPs em relação à variância do fenótipo (Expressão 5.14), que é a soma da variância dos SNPs com a variância dos resíduos.

$$V(\beta_i) = \tau_i^2\sigma_e^2 2p_i(1 - p_i) \tag{5.12}$$

$$V(\beta) = \sum_{i=1}^{m} \tau_i^2 \sigma_e^2 2p_i(1-p_i) \tag{5.13}$$

$$h^2 = \frac{V(\beta)}{V(\beta) + \sigma_e^2} \tag{5.14}$$

Como o Blasso foi construído inicialmente para predição de valor genético baseado em informação genômica obtida por marcadores SNPs, ele não é um método de seleção, mas sim de predição. Entretanto, como no cômputo dos efeitos dos marcadores são utilizados fatores redutores, alguns efeitos de marcadores convergem para valores próximos a zero, e outros não. Assim, pode-se considerar que essa convergência para zero de certa forma elimina o SNP considerado irrelevante no fenótipo predito, ocorrendo uma seleção de atributos. Uma adaptação possível para tornar o Blasso um método de seleção de atributos é ordenar de forma decrescente os marcadores SNPs pelo percentual de variância de cada marcador em relação à variância fenotípica total ($V(\beta) + \sigma_e^2$), e adotar um limite inferior para seleção dos SNPs considerados informativos. Como exemplo, pode-se considerar somente os SNPs que tenham variância maior ou igual a 1%, isto é, SNPs que expliquem pelo menos 1% da variância do fenótipo. Outra forma, é usar algum limite inferior para o efeito absoluto (considera-se o módulo do efeito) de cada marcador, como por exemplo, todos os coeficientes acima de 0,1 serão selecionados como relevantes, pois estes teriam impactos significativos. Logo, usando a variância ou o coeficiente de cada marcador como critério para selecionar os SNPs mais importantes, o Blasso pode ser classificado na categoria de métodos embutidos, pois a pesquisa pelos subconjuntos de atributos é realizada pelo mesmo algoritmo que avalia os mesmos.

5.3 Métodos Não-Paramétricos em GWAS

5.3.1 Random Forests em GWAS

A Tabela 5.1 permite observar o tipo de efeito genético de um marcador SNP que pode ser capturado pela RF. Ou seja, a RF pode tratar situações com ações no próprio *locus* somente aditivas, onde cada alelo variante aumenta ou diminui a variação fenotípica, ou ações não-aditivas de dominância onde a presença de ao menos um alelo variante aumenta a variação do fenótipo, ou ações não-aditivas de tipo recessivo, onde somente o

hetorozigoto promove a variação na característica.

Tabela 5.1 Diferentes efeitos genéticos. Adaptada de Goldstein, Polley e Briggs (2011).

Tipo	Mecanismo	Partição
Aditivo	Cada alelo variante aumenta a variação	0,1,2
Dominante	Presença de ao menos 1 alelo variante aumenta a variação	0, 1/2
Recessivo	Heterozigoto promove a variação	0/2,1

Com existem muitas vantagens na utilização de RF em GWAS, há um aumento rápido no número de publicações que associam RF e genética como pode ser notado na Figura 5.2. Por exemplo, Bureau et al. (2005) aplicaram a RF como método para identificar interação entre pares de SNPs em um conjunto de dados relativo à asma com casos e controles não afetados. Foram avaliados 42 SNPs no gene ADAM33, o qual já havia sido previamente identificado como tendo associação com a asma. Bureau et al. (2005) concluíram que SNPs e pares de SNPs altamente associados com a asma tendem a ter alto valor para o índice de importância, mas associação e importância preditiva nem sempre coincidem. Eles modificaram a medida de importância de variável da RF para tratar a importância de pares de SNPs e mostraram que essa medição pode ser mais vantajosa do que medir a importância individual de cada SNP.

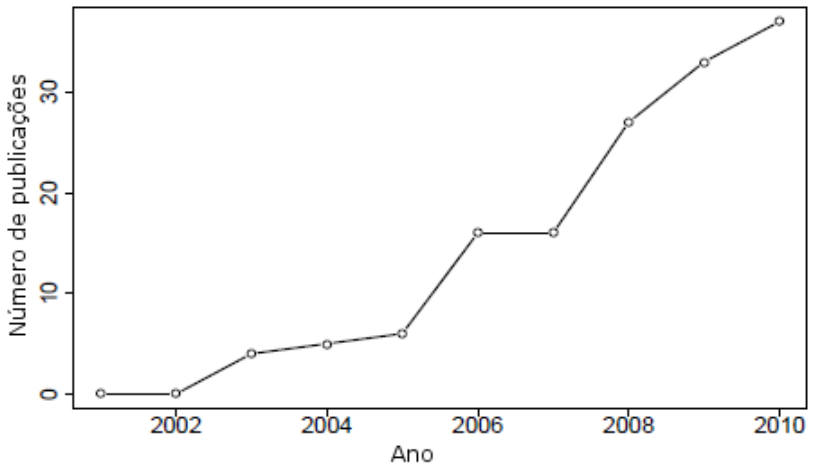


Figura 5.2 Artigos listados em PubMed usando os termos de busca (Random Forest OR Random Forests) AND (Gene OR SNP).

Foram encontrados mais de 125 artigos desde 2001 (os artigos do *Genetics Analysis Workshop* foram omitidos). Adaptado de Goldstein, Polley e Briggs (2011).

Em cenários com interações entre SNPs, Winham et al. (2012) analisaram o

comportamento da RF na seleção de SNPs quando a quantidade de marcadores aumenta e compararam com o método do valor-p da regressão logística univariada. Os resultados indicaram que a RF identifica interações em dados com baixa dimensionalidade. Mas, quando o número total de variáveis preditivas aumenta, a probabilidade de detecção diminui mais rapidamente para SNPs que interagem do que para SNPs que não interagem, o que indica que em dados com alta dimensionalidade as medidas de importância de variável da RF captura mais efeitos marginais do que os efeitos de interação entre SNPs.

Outro estudo de GWA que considerou um estudo de caso-controle para esclerose múltipla com RF foi realizado por Goldstein, Polley e Briggs (2011) para mostrar que os valores *default* para os parâmetros não são apropriados para conjuntos de dados usados em GWAS. Como sugeriram os autores, ganhos podem ser obtidos por sub-amostragem dos dados, poda baseada em desequilíbrio de ligação e remoção de SNPs com grande efeito da análise da RF (GOLDSTEIN; POLLEY; BRIGGS, 2011). Neste trabalho, novos genes foram identificados como potencialmente associados com a doença.

Meng et al. (2009) avaliaram que, em situações de correlação entre um verdadeiro SNP causal e SNPs em desequilíbrio de ligação, a diminuição da importância de variável para o verdadeiro SNP causal pode ocorrer. Uma abordagem para resolver este problema consiste em selecionar os SNPs em equilíbrio de ligação para análise (MENG et al., 2009). Métodos alternativos foram explorados para lidar com SNPs em desequilíbrio de ligação alterarando o algoritmo de construção de árvore por meio da construção de cada árvore apenas com SNPs em equilíbrio de ligação, modificando a medida importância, e usando haplótipos em vez de SNPs para construir a RF.

5.3.2 SVM ou SVR em GWAS

O SVM ou SVR originais são usados com bastante eficiência para a predição em problemas de classificação e de regressão com muitas variáveis, não sendo capazes de realizar a seleção de variáveis. Entretanto, algumas adaptações foram feitas por Villela, Leite e Neto (2014), como o método de seleção baseado numa busca ordenada, conhecida como *best-first*. Porém, essa técnica fica limitada a problemas de classificação multiclasses, não sendo possível o uso para fenótipos contínuos, desde que o mesmo seja discretizado.

Outros estudos de SVM aplicados em GWAS para fenótipos binários utilizaram uma abordagem de filtro baseado no valor-p de testes estatísticos como os realizados por Wei

et al. (2009), Ban et al. (2010) e Mittag et al. (2012). Um estudo mais recente em GWAS, realizado por Kim et al. (2013), usou um procedimento de duas fases: a primeira etapa selecionou SNPs promissores e identificou seus modelos genéticos usando MAX *test* descritos em Zheng, Freidlin e Gastwirth (2006), a segunda, ajustou um modelo preditivo usando SVM penalizado com pesos apropriados para os SNPs selecionados na primeira fase, os quais foram baseados em seus tipos genéticos. O algoritmo do SVM penalizado realiza seleção de atributos e predição simultaneamente, porém, trabalha somente em problemas de classificação. Além disso, quando o número de SNPs do conjunto inicial é relativamente grande, é necessário aplicar algum tipo de filtro como primeiro passo para que o SVM penalizado realize uma segunda seleção.

Yoon et al. (2003) utilizaram o SVM para estimar a suscetibilidade em indivíduos de alto risco para doença arterial coronariana. Neste estudo, a aplicação inicial de filtros não foi necessária porque foram investigados 14 SNPs de dez genes candidatos para doença arterial coronariana. Foram avaliados os *kernels* polinomial de ordens 1 (linear), 2 e 3, além do radial que apresentou a maior acurácia.

Cosgun e Duarte (2011) compararam *Boosted Regression Tree* (BRT), *Random Forest Regression* (RFR) e *Support Vector Regression* (SVR) em subconjuntos de SNPs selecionados, inicialmente pelo valor-p, para predizer doses de varfarina em afro-americanos. Além de SNPs, os modelos construídos incluíram covariáveis de idade, sexo, peso, altura, insuficiência cardíaca congestiva e doença renal crônica moderada ou grave para melhorar a predição das técnicas avaliadas. A RFR apresentou o melhor desempenho, entretanto, BRT e SVR tiveram resultados similares. Como Cosgun e Duarte (2011) sugerem, estudos posteriores sobre o SVR precisam ser feitos. Cabe destacar que tanto o SVR quanto a RFR (ou RF) demonstraram seu potencial na predição de fenótipo contínuo baseado em variáveis genotípicas e ambientais.

O SVR também é muito usado para predizer o valor genético genômico para animais em GWS. Nesse contexto, Long et al. (2011) usou o SVR com *kernels* linear e radial para predizer a produção de leite em gado e produção de grãos de trigo. Os resultados obtidos foram equivalentes ou superiores ao Blasso. Cabe ressaltar que neste caso, não foi aplicado nenhum tipo de filtro para reduzir a quantidade inicial de SNPs, ou seja, usou-se o SVR como um preditor para o fenótipo.

5.3.3 *Algoritmos Genéticos em GWAS*

Para a seleção de SNPs informativos com Algoritmos Genéticos, é necessário acoplar algum indutor para o aprendizado como função de aptidão do GA como foi discutido anteriormente nos métodos do tipo *wrapper*. Como exemplo, Shah e Kusiak (2004) usaram uma abordagem baseada em *data mining* e Algoritmos Genéticos para selecionar os SNPs que impactam a cura ou o desenvolvimento de drogas para várias doenças. O método consiste de um mecanismo de busca global, árvore de decisão ponderada, árvore de decisão baseada em *wrapper* e uma heurística baseada em correlação. A identificação do conjunto interseção dos atributos selecionados por cada uma das técnicas é realizada para selecionar os genes mais significantes. Após a aplicação do método, ocorreu uma redução de 85% do número de atributos e um aumento de 10% e 3,2% para acurácia e especificidade respectivamente.

Outro trabalho com Algoritmos Genéticos em GWAS, İlhan e Tezel (2013) usa GA com SVM para encontrar um subconjunto de *tag* SNPs que representem adequadamente os SNPs restantes, mas, nesse caso, a informação do fenótipo não é utilizada, pois a ideia é encontrar SNPs que representem adequadamente seu bloco LD. O objetivo é eliminar os SNPs redundantes, mas garantir que os SNPs representantes de cada bloco LD consigam predizer os SNPs restantes, que foram eliminados. Os parâmetros C e γ do SVM são otimizados pela técnica exame de partículas (do inglês, *Particle Swarm Optimization* - PSO).

5.4 Seleção de SNPs com interação

Existem duas classes de algoritmos que buscam por interações entre SNPs: métodos de busca exaustiva e não-exaustiva. A Tabela 5.2 demonstra uma comparação entre métodos para detecção de interações entre SNPs, onde o número de *loci* é especificado *a priori* em todos os algoritmos.

Os Algoritmos Genéticos se encaixam nos métodos não-exaustivos, já que eles não avaliam todas as possibilidades de interações. Para utilizar Algoritmos Genéticos na seleção de atributos é necessário embutir um algoritmo indutor de aprendizado em sua função de aptidão. Desse modo, a primeira vantagem de Algoritmos Genéticos aplicados em seleção de atributos, que possuem algum tipo de interação, com relação aos métodos

Tabela 5.2 Comparação de alguns métodos utilizados para detecção de interação de SNPs. Adaptado de Olazar (2013).

Algoritmo	**# inter.**[a]	**T. amostra**[b]	**Tipo de teste aplicado**	**Busca**[c]	**Referência**
FastANOVA	2 *loci*	100.000	Teste ANOVA	E	Zhang, Zou e Wang (2008)
COE	2 *loci*	100.000	Testes estatísticos convexos como Qui-Quadrado, razão de máxima verossimilhança, informação mútua e teste Cochran-Armitage.	E	Zhang et al. (2009)
TEAM	2 *loci*	500.000	Árvore de expansão mínima baseada em testes estatísticos convexos (como em COE)	E	Zhang et al. (2010)
MDR	k *loci*	10.000	*Data mining*	E	Moore et al. (2006)
BOOST	2 *loci*	500.000	Máxima verossimilhança	E	Wan et al. (2010a)
TROOST	3 *loci*	500.000	Máxima verossimilhança	E	Neto (2013)
EpiSNP	2 *loci*	500.000	Modelo Kempthorne	E	Ma et al. (2008)
FastEpistasis	2 *loci*	500.000	Regressão logística	E	Schüpbach et al. (2010)
PLINK	2 *loci*	500.000	Regressão logística	E/NE	Purcell et al. (2007)
AntEpiSeeker	k *loci*	100.000	Qui-Quadrado/ACO	NE	Wang et al. (2010)
SNPRuler	k *loci*	100.000	Aprendizado baseado em regras	NE	Wan et al. (2010b)
InterSNP	2 *loci*	300.000	Regressão logística	NE	Herold et al. (2009)
MECCPM	k *loci*	300.000	Critério de informação bayesiana (BIC)	NE	Miller et al. (2009)
SNPHarvester	k *loci*	500.000	Regressão logística penalizada	NE	Yang et al. (2009)
BEAM	2 *loci*	500.000	Modelo bayesiano	NE	Zhang e Liu (2007)
Epiforest	2 *loci*	100.000	Random Forest	NE	Jiang et al. (2009)
GENN	2 *loci*	100	Rede Neural	NE	Motsinger-Reif et al. (2008)
MIGA-2L	2 *loci*	500.000	Algoritmo Genético	NE	Olazar (2013)

[a] Número de SNPs interagindo.
[b] Número aproximado de SNPs na amostra.
[c] Estratégia de busca, onde "E" significa exaustiva e "NE" denota não-exaustiva.

greedy[1] é o esquema populacional adotado que explora soluções candidatas em diferentes partes do espaço de busca sem avaliar todas as possíveis interações, pois essa quantidade cresce com o aumento do número de atributos inicial (FREITAS, 2001). A segunda vantagem é que o operador *crossover* modifica vários genes (atributos) de uma só vez, permitindo um melhor desempenho do que estratégias que trabalham com um único atributo por vez (FREITAS, 2001). A terceira vantagem é que a função de aptidão dos Algoritmos Genéticos avalia um indivíduo (subconjunto de atributos) como um

[1]Um algoritmo indutor é dito ser *greedy* se: (1) ele contrói uma regra em um esquema incremental considerando um atributo por vez; (2) em cada passo a melhor escolha local é feita (FREITAS, 2001). Os métodos de seleção baseados em valor-p e *Stepwise* são exemplos de métodos *greedy*.

todo, conseguindo distinguir "bons grupos de atributos" de "grupos de atributos ruins" (FREITAS, 2001; PACKARD, 1990). Por último, os Algoritmos Genéticos, por meio dos operadores de seleção de pais, mutação e *crossover*, procuram somente um subconjunto das possíveis interações entre os atributos do conjunto inicial (CONGDON, 1995). Todavia, tendo encontrado uma importante interação, ele é frequentemente hábil em preservar esse padrão em gerações futuras (CONGDON, 1995). Assim, é muito provável que ao final da execução do GA, as interações relevantes detectadas em gerações distintas estejam presentes no subconjunto final de atributos.

Em relação à busca de interações entre SNPs, existem muitas abordagens propostas na literatura. Porém, as técnicas desenvolvidas pré-definem o número de interações entre os *loci*. Por exemplo, para pesquisar pares de SNPs interagindo, adapta-se o modelo de avaliação dos subconjuntos de SNPs e o algoritmo de busca para tratar somente dois SNPs. Na busca de interações entre trio de SNPs, a metodologia é adaptada para buscar e avaliar somente três SNPs.

Uma metodologia desenvolvida por Olazar (2013), denominada MIGA-2L, busca por pares significativos de SNPs para estudos de caso-controle (classificação) usando um algoritmo genético executado sobre máscaras de grupos de SNPs. Esse algoritmo é baseado na teoria da informação mútua para avaliação dos pares de SNPs. Esse método não faz uma busca exaustiva no espaço de soluções, pois é baseado em um algoritmo genético, o qual avalia uma amostra promissora do espaço de busca.

Para pesquisa de interações entre trios de SNPs para fenótipos binários, Neto (2013) desenvolveu um método, denominado TROOST, baseado no BOOST (WAN et al., 2010a) para pares de SNPS, que faz uma busca exaustiva de ternas de SNPs usando placas gráficas como processadores (GPGPU) em estudos de caso-controle. A pesquisa por trios de SNPS utiliza a busca por pares de SNPs como passo inicial (NETO, 2013).

6 O Método Proposto

Neste capítulo será descrito detalhadamente o método proposto para a seleção de marcadores SNP, bem como as justificativas para sua construção. Além disso, serão descritas duas versões do mesmo, no intuito de mostrar o processo evolutivo das principais ideias que nortearam a construção dessa abordagem para seleção de atributos em GWAS.

6.1 Introdução

Após o processo de genotipagem dos SNPs, faz-se o controle de qualidade (CQ) com a aplicação dos filtros, como por exemplo, *call rate*, MAF e HWE, com o objetivo de eliminar marcadores com informações ruidosas que podem gerar falsos-positivos caso não seja realizado esse processo. Como o número de SNPs no conjunto de dados inicial é na ordem de milhares, mesmo após o controle de qualidade, a quantidade de marcadores ainda é suficientemente grande para se extrair SNPs informativos com relação ao fenótipo analisado. Portanto, é necessário uma primeira etapa do *workflow* (6.1) denominada relevância, que objetiva colocar os SNPs individualmente em ordem decrescente de importância para o fenótipo em questão. Em seguida à fase de relevância, a etapa denominada corte que define o limite para a primeira seleção dos marcadores é de suma importância, pois dependendo da forma como esse ponto de corte é escolhido, pode-se flexibilizar a entrada de SNPs causais juntamente com muitos não-causais, ou restringir demasiadamente, perdendo-se SNPs informativos. Uma possível etapa subsequente à fase de corte é o refinamento, cujo objetivo é buscar por conjuntos de SNPs significativos principalmente quando existem interações entre SNPs. A partir das etapas descritas acima, a construção do *workflow* da Figura 6.1 facilita o entendimento da maioria dos métodos de seleção de SNPs em GWAS, além de permitir comparações sob o mesmo prisma.

O método proposto no presente trabalho é denominado *SNP Markers Selector* (SMS), cuja tradução livre é Seletor de Marcadores SNP. A concepção do SMS foi baseada nas três etapas do *workflow* da Figura 6.1. Na versão atual, ele é composto por três técnicas da Inteligência Computacional a saber: *Random Forests* (RF), Máquinas de Vetores Suporte

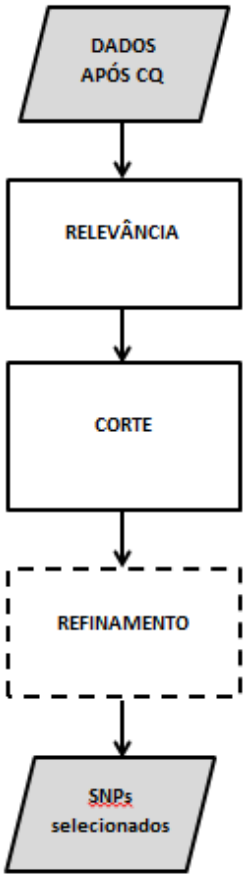


Figura 6.1 *Workflow* adotado para análise de métodos de seleção de atributos em GWAS.

O retângulo tracejado é uma das inovações do SMS em relação à maioria dos métodos de seleção em GWAS. O item **Dados após CQ** no fluxograma significa dados após controle de qualidade.

(do inglês, *Support Vector Machine* - SVM) e Algoritmos Genéticos (do inglês, *Genetic Algorithms* - *GA*). A RF é a técnica utilizada para a etapa de relevância, o SVM é usado para a etapa de corte, e finalmente, o GA com função de aptidão baseada no SVM realiza a etapa de refinamento.

A inovação do SMS está em dividir o processo de seleção nas seguintes etapas: ordenamento dos marcadores realizado pelo valor-p bruto (primeira versão) ou pelo *rank* da RF, avaliação dos subconjuntos gerados a partir dessa ordem, primeira seleção baseada no erro de predição do SVM (classificação) ou do SVR (regressão) na validação cruzada com 10-*folds* e, finalmente, a busca pela melhor combinação de marcadores pelo GA com base nos marcadores selecionados na primeira seleção. Todo o processo é repetido i vezes e uma das possibilidades para o conjunto final de marcadores é a união ou a interseção de

todos os subconjuntos gerados. Como o SMS usa a RF na fase de relevância e múltiplos *kernels* nas fases de corte e de refinamento, há a possibilidade de detectar diversos tipos de interações entre SNPs que impactam na determinação do fenótipo. Desta forma, o SMS atribui técnicas robustas para cada etapa de seleção, minimizando a perda de marcadores verdadeiro-positivos , os pontos fracos de cada abordagem quando aplicadas separadamente e, simultaneamente, maximizando o poder explicativo dos marcadores selecionados ao final de todo processo.

Outro ponto relevante considerado pelo SMS é busca dos próprios marcadores e não dos blocos haplótipos (blocos LD) construídos *a priori*, pois para se construir tais blocos é necessário somente o cálculo do LD entre os marcadores e o estabelecimento de um ponto limite para a definição do bloco, mas não a associação dos mesmos com a característica em questão. Com a definição do bloco, escolhe-se um representante para o mesmo (*tag* SNP), o que possibilita a eliminação de algum SNP do bloco que tenha informação complementar ao *tag* SNP para explicar parte do fenótipo.

6.2 Primeira Versão do SMS (SMS1)

A primeira versão desse método foi desenvolvida por Oliveira et al. (2014b) com base nos trabalhos de Wei et al. (2009) (diabetes do tipo 1) e Ban et al. (2010) (diabetes do tipo 2), pois os mesmos usaram o mesmo procedimento, com base no valor-p bruto de testes estatísticos para estudos de caso-controle, para construir uma sequência crescente de subconjuntos de SNPs, sendo esses subconjuntos definidos por cortes feitos tomando valores-p crescentes. Por exemplo, $S_1 \subset S_2 \subset ... \subset S_n$, onde S_1 são os SNPs com valores-p menores que 10^{-n}, S_2 são os SNPs com valores-p menores que 10^{-n+1}, S_3 são os SNPs com valores-p menores que 10^{-n+2}, e assim sucessivamente, até S_n que possui SNPs com valor-p menores que 10^n, onde n é escolhido diferentemente por cada estudo. Em seguida, cada subconjunto S_i é avaliado pelo SVM para todo i, escolhendo-se o subconjunto com melhor resultado para a métrica adotada. É importante destacar que Wei et al. (2009) executam somente as etapas de relevância e corte, baseadas, respectivamente, no valor-p bruto e no SVM com *radial*. Ban et al. (2010) executam as fases de relevância e corte do mesmo modo que Wei et al. (2009), porém, a etapa de refinamento é realizada por um processo de seleção de atributos denominado *Stepwise*.

O primeiro passo do SMS é baseado em um método de filtro para a primeira seleção de SNPs por meio do valor-p bruto da correlação de Spearman do marcador com o fenótipo contínuo, pois o valor-p corrigido por Bonferroni é muito restritivo para uma primeira etapa de filtro. O *rank* do valor-p bruto não é realizado por cromossomo, mas é realizado globalmente para todos os SNPs analisados. O segundo passo é agrupar os marcadores em subconjuntos por meio do limite superior para o valor-p dos marcadores estipulado *a priori*, avaliar o MSE de cada subconjunto e escolher o que possuir menor MSE. O terceiro passo é a avaliação desses subconjuntos de SNPs com base no SVR, e finalmente, o *wrapper* baseado no GA é usado para produzir subconjuntos de marcadores relevantes e o SVR é utilizado para avaliar os mesmos. O *kernel* adotado em todos os passos em que o SVR é utilizado foi o Pearson Universal *kernel* (PUK) o qual é explanado detalhadamente em Ünstü, Melssen e Buydens (2006). Nesta versão, o SMS é executado apenas uma vez, pois o objetivo inicial era analisar a possível melhoria da etapa de refinamento realizada pelo GA. A Figura 6.2 elucida o funcionamento do SMS na sua primeira versão.

A fase de relevância foi realizada pelo valor-p bruto e executadas no *software* R, mas as fases de corte e refinamento foram implementadas no *software* WEKA desenvolvido por Hall et al. (2009), pois o PUK já estava implementado no WEKA, mas não no R. A função de aptidão do GA no WEKA para o SVR foi o RMSE (*Root Mean Square Error*).

A descrição detalhada da primeira versão do SMS pode ser encontrada em Oliveira et al. (2014b) e Oliveira et al. (2014a). É importante alertar que nessa versão do SMS (SMS1), a fase de relevância foi realizada pelo valor-p bruto do coeficiente de correlação de Spearman e não pelo *rank* da RF baseado no pVI (regressão) ou gVI (classificação). De forma sintética, os passos do SMS são descritos a seguir.

1. Calcula-se o valor-p bruto da correlação de Spearman entre cada marcador e o fenótipo contínuo. No caso de fenótipo binário, calcula-se o valor-p bruto do teste qui-quadrado entre o SNP e o característica.

2. Agrupa os marcadores por meio do valor-p para construir uma sequência crescente de subconjuntos de marcadores em relação à quantidade de SNPs, ou seja, $S_1 \subset S_2 \subset ... \subset S_n$, onde S_1 são os SNPs com valores-p menores que 10^{-n}, S_2 são os SNPs com valores-p menores que 10^{-n+1}, S_3 são os SNPs com valores-p menores que 10^{-n+2}, e assim sucessivamente, até S_n.

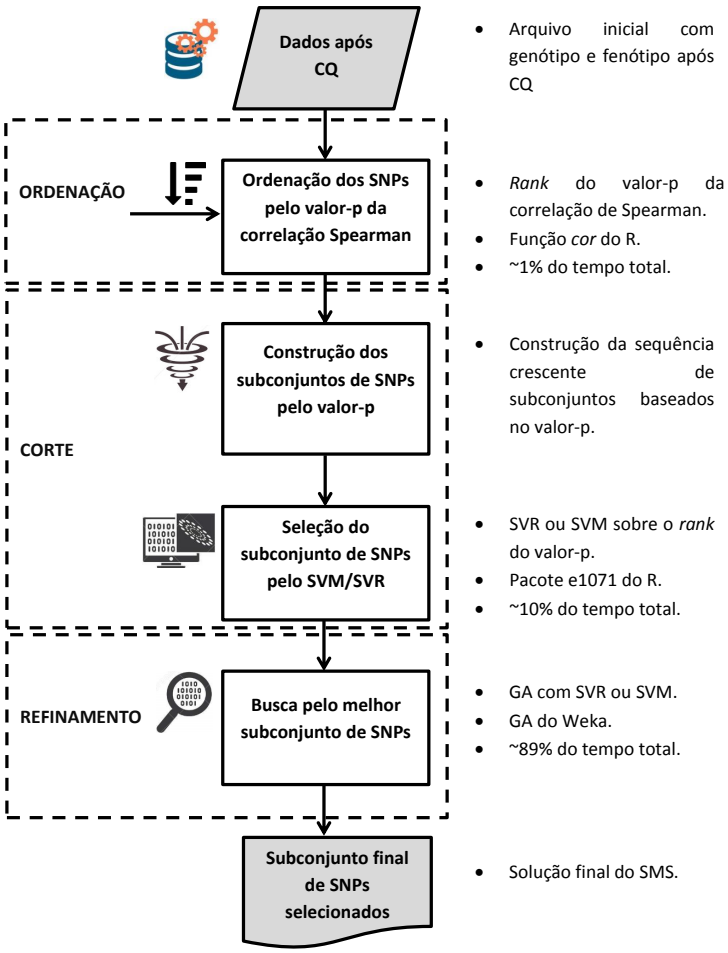


Figura 6.2 Fluxograma da primeira versão do SMS para o PUK adotado para o SVM/SVR. O item **Dados após CQ** no fluxograma significa dados após controle de qualidade.

3. Computa-se a correlação de Pearson (ou AUC) entre os valores preditos pelo SVR (ou SVM) baseados nos subconjuntos de SNPs construídos no passo 2 e escolhe-se o subconjunto que tem o melhor desempenho.

4. A partir do melhor subconjunto selecionado na etapa anterior, usa-se o GA para realizar uma busca não-exaustiva pelo "melhor" subconjunto de SNPs e realizar a

segunda e última seleção.

6.3 Versão Atual do SMS (SMS2)

Na atual versão, o valor-p bruto do coeficiente de correlação de Spearman foi substituído pelo *rank* da RF baseado na importância de permutação da variável (pVI) para fenótipos contínuos (regressão). Para fenótipos binários (classificação), o valor-p bruto do teste qui-quadrado foi substituído pelo *rank* da RF baseado na importância de gini da variável (gVI). Nesta versão, a ordenação baseada na RF é realizada por cromossomo, não considerando os marcadores de todos os cromossomos simultaneamente como feita na primeira versão do SMS. Os subconjuntos de SNPs são construídos de forma incremental e constante, sendo avaliados em cada cromossomo separadamente, o que não ocorreu na primeira versão do SMS, pois na fase de corte, os SNPs não eram avaliados por cromossomo, mas considerando como se os SNPs de todos os cromossomos fossem agrupados em um único cromossomo. A primeira seleção é definida a partir do menor MSE (AUC) gerado pelo SVR (SVM) aplicado sobre o *rank* da RF. Após o corte feito em cada cromossomo, a união dos subgrupos de SNPs por cromossomo é realizada, gerando o tamanho do cromossomo (subconjunto de SNPs selecionados pela primeira seleção) usado na otimização do GA. Em essência, o GA não foi modificado, mas a função de aptidão baseada no RMSE do SVR foi substituída pela correlação de Pearson. Além de todas as modificações anteriores, todas as etapas do SMS foram implementadas no *software* R, facilitando todas as análises subsequentes de desempenho do SMS.

A decisão em substituir o valor-p bruto do coeficiente de correlação de Spearman pelo *rank* da RF na etapa de ordenação ocorreu após a aplicação da primeira versão do SMS no conjunto de dados simulados do QTLMAS 2011. Esses dados possuem 9.990 SNPs e um fenótipo contínuo, porém, só oito SNPs são causais e estão dispersos em cinco cromossomos, sendo que dois SNPs interagem produzindo uma ação gênica não-aditiva com efeito significativo. Assim, caso a ordenação da RF fosse substituída pelo valor-p bruto nessa fase, somente três dos oito SNPs informativos iriam para as etapas de corte e de refinamento, isto é, o SMS não capturaria o par de SNPs com interação. Consequentemente, o SMS não teria chance de selecionar os outros cinco em nenhuma etapa posterior (corte e refinamento). O trabalho seminal que motivou a inserção do

rank da RF por cromossomo foi desenvolvido por Higa et al. (2014), que utilizou a RF para selecionar os SNPs mais informativos nos dados do QTLMAS 2011. Posteriormente, descobriu-se que a metodologia desenvolvida em Higa et al. (2014) foi baseada em um estudo de associação em escala genômica feito por Mokry et al. (2013), logo, todos os passos desse trabalho foram minuciosamente estudados para posterior aplicação da RF no SMS. Essa discussão será aprofundada no Capítulo 8, onde usou-se a nova versão do SMS (SMS2) no conjunto de dados do QTLMAS 2011.

O SMS é repetido i vezes, pois há duas fontes de aleatoriedade: uma na construção do *rank* da RF e outra durante o processo de busca do GA. Como exemplo, um SNP pode ser classificado na posição 10 numa execução da RF e ordenado na posição 13 em outra execução, entretanto, a variação do *rank* de um marcador é aceitável. Em relação ao GA, quando executa-se o mesmo duas vezes, em geral, as duas soluções são conjuntos de marcadores distintos, porém, a interseção entre esses conjuntos é, na maioria das vezes, não vazia, o que indica estabilidade na seleção do GA. Uma finalidade do GA, após a primeira seleção feita pelo SVR/SVM em conjunto com a RF, é analisar os subconjuntos de SNPs não considerando a ordenação dada pela RF, mas combinações promissoras entre SNPs. Essa abordagem possibilita a identificação de possíveis interações entre os marcadores e a eliminação de SNPs não-informativos que foram selecionados pela etapa de corte. Outra finalidade do GA é realizar uma busca por uma boa combinação de SNPs que esteja associada ao fenótipo sem avaliar exaustivamente todas as combinações possíveis.

Resumidamente, o SMS é a união de duas classes de métodos para seleção de atributos a saber: método de filtro, realizado pela RF, e método *wrapper*, realizado pelo GA mais o SVM/SVR. O método desenvolvido possui 5 etapas principais que podem ser observadas na Figura 6.3.

6.3.1 Etapas da Versão Atual do SMS

Na nova versão do SMS, além da substituição do valor-p bruto pelo *rank* da RF, houve a introdução de múltiplos *kernels*, sendo que cada um é executado i vezes, tanto para avaliar a estabilidade do método, quanto para possibilitar a seleção de vários tipos de subconjuntos ao final do processo. Todas as etapas do SMS para seleção de SNPs com fenótipo contínuo (regressão) estão descritas pelo pseudocódigo do Algoritmo 2.

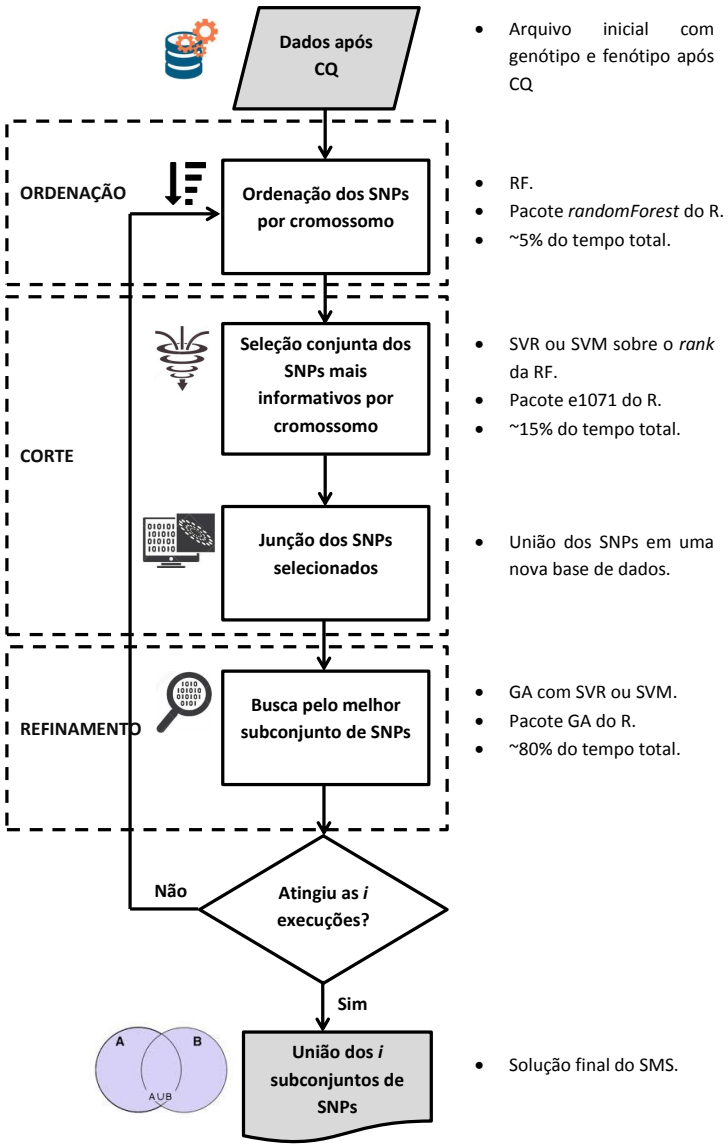


Figura 6.3 Fluxograma da versão atual do SMS para um determinado *kernel* adotado para o SVM/SVR.

O item **Dados após CQ** no fluxograma significa dados após controle de qualidade.

Para fenótipos binários (classificação), a medida de importância de permutação pVI é alterada para a medida de importância de gini gVI, o MSE médio é substituído pela AUC média e a função de aptidão baseada na média da correlação de Pearson é trocada pela AUC média. O Algoritmo 3 ilustra todos os passos necessários para a execução do SMS para classificação. Note que no Algoritmo 3 não aparece o parâmetro ϵ, pois o SVM

Algoritmo 2 SMS (*i*, *ntree*, *mtry*, *limitCHR*, *k*, *cost*, ϵ, *kernel*, γ, *numkernel*, *fold*, *iter*, *run*, *pop*, *pcross*, *pmut*,*elitism*, *fitness*)

para *contador*1 $\leftarrow$ 1 ... *i* **faça**
  para *contador*2 $\leftarrow$ 1 ... *numkernel* **faça**
    aplicar a função RandomForest(*ntree*, *mtry*) explicada pelo Algoritmo 1, ou seja, executa a *Random Forest*;
    construir o *rank* por cromossomo baseado na importância de permutação de cada marcador SNP (*pVI*);
    gerar a família crescente de subconjuntos de SNPs a partir do *rank* da RF por cromossomo com incremento de 10 SNPs;
    avaliar os suconjuntos anteriores por cromossomo pelo MSE médio do SVR(ϵ, *kernel*, γ) até o *limitCHR*;
    efetuar o corte em cada cromossomo para realizar a primeira seleção baseado em *k* passos de comprimento 10 após o valor mínimo do MSE médio;
    unir todos os subconjuntos construídos por cromossomo;
    executar o GA(*iter*, *run*, *pop*, *pcross*, *pmut*, *elitism*, *fitness*) baseado no conjunto união anterior.
  fim
fim
Selecionar a solução do GA (subconjunto de SNPs) com a maior função de aptidão (maior correlação de Pearson média) após *i* iterações;
construir a união das melhores soluções de cada *kernel*.

não faz uso do mesmo em sua formulação matemática, sendo o mesmo utilizado somente pelo SVR.

Algoritmo 3 SMS (*i*, *ntree*, *mtry*, *limitCHR*, *k*, *cost*, *kernel*, γ, *numkernel*, *fold*, *iter*, *run*, *pop*, *pcross*, *pmut*,*elitism*, *fitness*)

para *contador*1 $\leftarrow$ 1 ... *i* **faça**
  para *contador*2 $\leftarrow$ 1 ... *numkernel* **faça**
    aplicar a função RandomForest(*ntree*, *mtry*) explicada pelo Algoritmo 1, ou seja, executa a *Random Forest*;
    construir o *rank* por cromossomo baseado na importância de gini de cada marcador SNP (*gVI*);
    gerar a família crescente de subconjuntos de SNPs a partir do *rank* da RF por cromossomo com incremento de 10 SNPs;
    avaliar os suconjuntos anteriores por cromossomo pelo AUC média do SVM(*kernel*, γ) até o *limitCHR*;
    efetuar o corte em cada cromossomo para realizar a primeira seleção baseado em *k* passos de comprimento 10 após o valor máximo da AUC média;
    unir todos os subconjuntos construídos por cromossomo;
    executar o GA(*iter*, *run*, *pop*, *pcross*, *pmut*, *elitism*, *fitness*) baseado no conjunto união anterior.
  fim
fim
Selecionar a solução do GA (subconjunto de SNPs) com a maior função de aptidão (maior AUC média) após *i* iterações;
construir a união das melhores soluções de cada *kernel*.

A variável *i* representa o número de vezes que o SMS será executado para um determinado *kernel*. As variáveis *ntree* e *mtry* são, respectivamente, o número de árvores da floresta e o número de variáveis selecionadas aleatoriamente para a construção de cada árvore. O parâmetro *k* é o número de passos de comprimento 10 (10 SNPs acrescentados por vez) dados após o subconjunto que atinge o mínimo do MSE médio ou da AUC média encontrado na etapa de corte. Os parâmetros *cost*, ϵ, *kernel* e γ são referentes ao SVR usado nas etapas de corte e de refinamento. O número de *kernels* avaliados é dado pela variável *numkernel*. O parâmetro *fold* é o número de partições usadas na validação cruzada usada para avaliar dos subconjuntos de SNPs gerados nas etapas de

corte e de refinamento. Os parâmetros usados pelo GA são *iter*, *run*, *pop*, *pcross*, *pmut* e *elitism* representam respectivamente o número de gerações, a quantidade de gerações sem melhoria, o tamanho da população, a probabilidade de *crossing over*, a probabilidade de mutação e o número de melhores indivíduos do GA (maiores aptidões) que pertencerão à próxima geração sem alteração alguma. O parâmetro *fitness* é a função de aptidão usada pela GA, sendo a correlação de Pearson média entre os valores preditos pelo SVR e os observados, e a AUC média entre as classes preditas feitas pelo SVM e as verdadeiras classes. Detalhadamente, o SMS é composto dos seguintes passos:

1. Executa-se o modelo da RF sobre todos os marcadores de cada cromossomo em conjunto com o fenótipo avaliado para obter o ordenamento decrescente de importância dos marcadores. Essa importância é mensurada pelo acréscimo percentual do MSE, para problemas de regressão onde o fenótipo é contínuo, ou pelo decréscimo percentual da acurácia para problemas de classificação com fenótipo binário, quando uma permutação aleatória é realizada no marcador em questão conforme discutido na seção anterior sobre RF.

2. Constrói-se uma sequência crescente de subconjuntos de marcadores usando o *rank* gerado pela RF com passo igual a 10 marcadores. Por exemplo, o primeiro subconjunto possuirá somente o primeiro marcador do *rank* da RF, o segundo terá os 10 primeiros (incluindo o primeiro), o terceiro, os 20 primeiros (incluindo os 10 primeiros), e assim sucessivamente, até o último subconjunto que conterá parte ou todo o conjunto de dados inicial. Esse procedimento é feito para cada cromossomo.

3. Implementa-se o SVR (regressão) ou o SVM (classificação) com validação cruzada com 10-*fold* sobre cada um dos subconjuntos gerados no passo 2 e avalia-se o erro de predição por meio do MSE médio para regressão ou a média da AUC para classificação de cada um dos modelos construídos. Esse passo é também realizado para cada cromossomo.

4. Calcula-se a maior média da correlação de Pearson ou a maior média da AUC, dependendo do tipo de problema (regressão ou classificação), obtendo-se o grupo de marcadores que maximiza a correlação média ou a AUC média. Esse passo representa o primeiro filtro do método SMS.

5. No conjunto de dados intermediário, aplica-se um GA como segundo e último filtro para a selecionar o subconjunto final de marcadores. A última seleção com uso do GA objetiva maximizar a média da correlação de Pearson para o SVR ou a média da AUC do SVM, sendo ambas as médias geradas pela validação cruzada com 10-*fold*.

6. Todo o processo (passos do 1 ao 5) é repetido i vezes.

7. Todas as etapas anteriores são repetidas para cada *kernel* considerado no SVM/SVR. O número de *kernels* usados é dado por *numkernel*.

8. A união ou a interseção dos subconjuntos com maior aptidão construídos por cada kernel em i execuções pode ser o subconjunto final de SNPs selecionados. Outra opção é escolher o *kernel* que apresentou melhor desempenho e adotar a união ou a interseção dos subconjuntos gerados nas i execuções.

A medida de importância usada para a ordenação da RF é o pVI para problemas de regressão e o gVI para problemas de classificação. Como analisado anteriormente, o pVI indica uma qualidade preditiva do marcador, enquanto, o gVI não mede a predição, mas sim a relevância na construção das árvores da floresta.

Para realizar a construção da sequência crescente de subconjuntos de SNPs na etapa 2, foram realizadas várias simulações incrementando-se um SNP por vez ou dez SNPs por vez em relação ao subconjunto anterior. O cenário com incremento de SNP por vez mostrou uma variabilidade excessiva, não permitindo uma nítida identificação do ponto de mínimo para o MSE médio (regressão) ou o ponto de máximo para a AUC média (classificação), pois ocorreu uma frequente oscilação nessas medidas. Por outro lado, o passo incremental de dez SNPs por vez, permitiu um comportamento estável para o MSE e para a AUC, facilitando a identificação direta dos pontos de mínimo para o MSE médio e de máximo para a AUC média. Devido ao exposto anteriormente, o incremento de 10 SNPs foi adotado para a atual versão do SMS.

O ponto de corte da primeira seleção é flexibilizado de tal forma a aumentar a probabilidade de entrada dos marcadores verdadeiros-positivos com pequenos impactos sobre o fenótipo, entretanto, pode-se incorrer no efeito contrário, que é permitir a entrada de falsos-positivos no conjunto. Assim, se torna necessário utilizar um novo filtro para eliminar os marcadores não informativos e isso é feito pelo GA. Consequentemente, um novo filtro, o GA com SVR, é aplicado para retirar o maior número de marcadores não

causais introduzidos na etapa 4, garantindo que o grupo final de marcadores contenham o maior número de verdadeiros-positivos e o menor de falsos-positivos.

No passo 4, como existe a possibilidade de imprecisão nessas métricas do grupo avaliado, acrescenta-se uma margem de erro no índice deste subconjunto, permitindo a entrada de mais marcadores. Com isso, aumenta-se a probabilidade da entrada de verdadeiros-positivos que não estão no grupo com correlação ou AUC máximas, mas em algum outro subconjunto que contém estes marcadores causais. É importante ressaltar que esse artifício permite também a entrada de marcadores falso-positivos, entretanto, o objetivo principal do método é perder o mínimo de SNPs verdadeiros-positivos em relação ao fenótipo. Ao final, constrói-se uma base de dados intermediária formada pela união de todos os SNPs selecionados por cromossomo juntamente com o fenótipo.

O único parâmetro otimizado pelo SMS é o γ do *kernel* radial que é usado pelo SVM/SVR. Estipula-se um valor para γ entre 0,001 a 1 (incluindo os limites) com passo multiplicativo 10. Desse modo, todo processo do SMS será executado para cada *kernel* radial com γ entre 0,001 a 1, além do *kernel* linear. O melhor γ será aquele que demonstrará o subconjunto de marcadores com maior correlação média (fenótipo contínuo) ou maior AUC média (fenótipo binário), e após esse passo, uma das opções é fazer a união das i soluções geradas para o *kernel* com melhor desempenho. Isso permite de algum modo caracterizar a arquitetura genótipo-fenótipo, pois como explica (BEN-HUR; WESTON, 2007), valores pequenos de γ produzem padrões de classificação próximos ao *kernel* linear e valores maiores, geram padrões não-lineares próximos à algum *kernel* polinomial com grau superior a 1. Como exemplo, caso o melhor valor de γ seja 0,1, há um indício de alguma não linearidade entre os marcadores selecionados e o fenótipo e outros métodos para encontrar a(s) interação(ões) podem ser explorados.

O *kernel* linear é também avaliado como modelo controle para o *kernel* radial e também devido à sua simplicidade, todavia, não é realizada otimização no parâmetro C, sendo o mesmo valorado como 1. O *kernel* polinomial não foi avaliado porque como argumenta (BEN-HUR; WESTON, 2007), o *kernel* radial supera o *kernel* polinomial em acurácia e em tempo de convergência. O PUK também não foi usado na versão atual do SMS, pois o interesse era avaliar o SMS com *kernels* mais comumente usados em GWAS.

O número de marcadores do grupo selecionado no passo 3 é o número de genes (SNPs) em cada cromossomo de um indivíduo utilizado pelo GA. Além disso, a função de aptidão

adotada no GA para avaliar cada indivíduo da população é a média do erro de predição MSE para regressão (SVR) ou da AUC para classificação (SVM), da mesma forma em que são avaliados os subconjuntos construídos na etapa 3.

O GA não poderia ser usado na primeira seleção devido ao grande tamanho de cada cromossomo, filtrando lentamente os SNPs falso-positivos e podendo perder os informativos. Isso se deve pela avaliação que seria feita sobre grupos distintos de marcadores da base de dados original (população do GA) e não individualmente, diferentemente, do que é feito pelo *rank* gerado pela RF. O ordenamento da RF é realizado por um indicador computado para cada marcador, através da permutação aleatória de seus valores; porém, considerando todas as árvores construídas na RF, as quais podem também ser interpretadas como subconjuntos de marcadores. Ou seja, o *rank* da RF é uma medida robusta do ponto de vista estatístico e mais rápida do ponto de vista computacional do que o GA com função de aptidão baseada no erro predito MSE SVR ou na acurácia predita do SVM.

O SVR/SVM também não poderia ser usado para criar um *rank* com referência na avaliação individual por algum processo de seleção de atributos como o *Stepwise*, pois o custo computacional seria alto, inviabilizando essa abordagem. Além do mais, o SVR/SVM possui um tempo computacional elevado para ordenar de forma decrescente os marcadores mais relevantes, com base em alguma medida de erro de predição (MSE ou correlação de Pearson). Utiliza-se a métrica importância da variável gerada pela RF para essa tarefa, pois tal procedimento é computacionalmente viável para bases de dados na ordem de milhares ou até mesmo milhões de marcadores como mostrado em Goldstein et al. (2010).

Outro ponto importante do SVR é a capacidade de trabalhar com funções *kernel* universais (por exemplo, o *kernel* radial e o *kernel* universal de Pearson) que possuem capacidade de extrair vários tipos de relações subjacentes entre as variáveis explicativas e a variável explicada. Desta forma, é possível mapear relações lineares, não lineares polinomiais de baixas e altas ordens e não lineares não polinomiais, gerando um amplo espectro de relações possíveis entre o genótipo e o fenótipo. Deste modo, não é necessário especificar a forma matemática do modelo *a priori*, mas somente o *kernel* a ser usado e seus respectivos parâmetros.

Como a predição da RF para regressão está limitada aos valores médios da variável

explicada (y) na fase de treinamento, o que não permite mudanças significativas no erro de teste quando determinado marcador é avaliado se deve ou não ser selecionado para etapa seguinte do SMS, o SVR é usado para medir a sensibilidade do erro gerado pela entrada de tal marcador no conjunto atual. Esse fato pode ser explicado devido do SVR assumir, na maioria das vezes, valores preditos distintos dos valores observados no treinamento, quando o modelo é aplicado nas variáveis de entrada do conjunto de teste. Consequentemente, uma vantagem do SVR é a sensibilidade da medida de erro MSE à entrada de marcadores verdadeiros positivos no grupo avaliado. Além do mais, outra justificativa para a escolha do SVR como avaliador de determinado grupo de marcadores é dada pela simulação com interação de ordens 2 e 3, que será abordada na seção Base de Dados Simuladas (Experimentos Computacionais).

A RF não foi usada como função de aptidão do GA, pois o custo computacional do GA se torna impraticável. Esse fato se deve ao número de árvores (parâmetro *ntree*) ser maior ou igual a 1.000, tornando muito lenta a avaliação de cada grupo de marcadores gerado durante o processo evolutivo do GA.

A solução do SMS pode ser a união dos *numkernel* melhores subconjuntos de SNPs das i execuções de cada *kernel* avaliado no SMS. O conjunto interseção também pode ser adotado como solução, caso a busca seja por um subconjunto de SNPs mais restritivo. Outra possibilidade é escolher somente um dos *kernels* avaliados (por exemplo, o que tenha melhor desempenho) e fazer a união ou a interseção das i execuções do SMS, não utilizando a abordagem de múltiplos *kernels*. A última possibilidade é executar o SMS para uma quantidade de *kernels*, especificada pela variável *numkernel*; e para cada *kernel*, realizar i execuções, perfazendo o total de $i \times numkernel$ subconjuntos de SNPs, com isso, possibilitando a construção da solução final do SMS dada pelos conjuntos união ou interseção desses subconjuntos de marcadores. Um procedimento similar de interseção e união de subconjuntos de SNPs selecionados por abordagens distintas é discutido em Shah e Kusiak (2004).

Em relação às três categorias para classificação de métodos de seleção de atributos (filtro, *wrapper* e embutido) dos métodos de seleção apresentados no capítulo anterior, o SMS é uma junção de um método de filtro, baseado no *rank* da RF, com um método *wrapper*, onde o algoritmo de indução do aprendizado é o SVM ou o SVR e o processo que direciona a busca pelos melhores subconjuntos de atributos é realizado pelo GA. Desta

forma, o SMS é um método que pertence à interseção dos métodos de filtro com métodos *wrapper*.

6.3.2 Codificação dos Dados de Entrada

A Figura 6.4 mostra como foi realizada a codificação numérica dos SNPs. Por exemplo, o SNP1 em relação aos quatro touros na base de dados inicial original, indica que o alelo A (representado pela base nitrogenada A) é o possui maior frequência alélica na população, enquanto que o alelo a (representado pela base nitrogenada C), a menor frequência alélica. Desta forma, o número 0 representa a ausência do alelo a ou o homozigoto AA (não necessariamente dominante), o número 1 indica o heterozigoto Aa e o número 2 caracteriza o homozigoto aa (não necessariamente recessivo). A matriz de dados de entrada para o conjunto de dados reais (PTA do leite) processado pelo SMS possui o formato da terceiro conjunto de dados da Figura 6.4.

Para os conjuntos de dados simulados do SCRIME, usou-se uma codificação distinta do conjunto de dados reais mostrada pela última matriz da Figura 6.4. A codificação padrão do SCRIME é representar o homozigoto de referência AA como 1, o heterozigoto Aa como 2 e o homozigoto variante aa como 3. Em problemas de classificação no SCRIME, o indivíduo doente é representado por 0 e o sadio por 1, onde o processo de codificação numérica é exemplificado pela Figura 6.5. No Capítulo 7, todas as codificações serão explicadas detalhadamente.

6.3.3 A Random Forest usada no SMS

Para a ordenação de SNPs por meio da importância da variável da RF em problemas de GWAS, usualmente, otimiza-se os parâmetros: número de variáveis para a escolha dos nós que compõem as árvores e o número de árvores da floresta. Entretanto, essa otimização consome tempo e recursos computacionais, mas não implica sempre em uma ordenação mais correta para os marcadores SNPs. Posto isso, uma maneira de superar esse obstáculo é definir valores fixos para esses dois parâmetros e testar por meio de simulações quais valores são mais adequados para as bases de dados usadas de tal forma a apresentarem resultados satisfatórios para o desempenho global do SMS. Para o número de variáveis usado na quebra das ávores utilizou-se o número de marcadores de cada cromossomo e 4.000 árvores foi definido como o total de árvores na floresta.

Base de dados inicial original

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Touro 1	AA	GG	GG	CT	TT	60,94
Touro 2	AG	GG	GG	CT	TT	70,45
Touro 3	AG	GT	AG	TT	TT	100,34
Touro 4	GG	TT	AA	TT	CC	500,65

Base de dados inicial com codificação alélica

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Touro 1	*AA*	*aa*	*AA*	*Aa*	*AA*	60,94
Touro 2	*Aa*	*aa*	*AA*	*Aa*	*AA*	70,45
Touro 3	*Aa*	*Aa*	*Aa*	*aa*	*AA*	100,34
Touro 4	*aa*	*AA*	*aa*	*aa*	*aa*	500,65

Base de dados inicial com codificação numérica

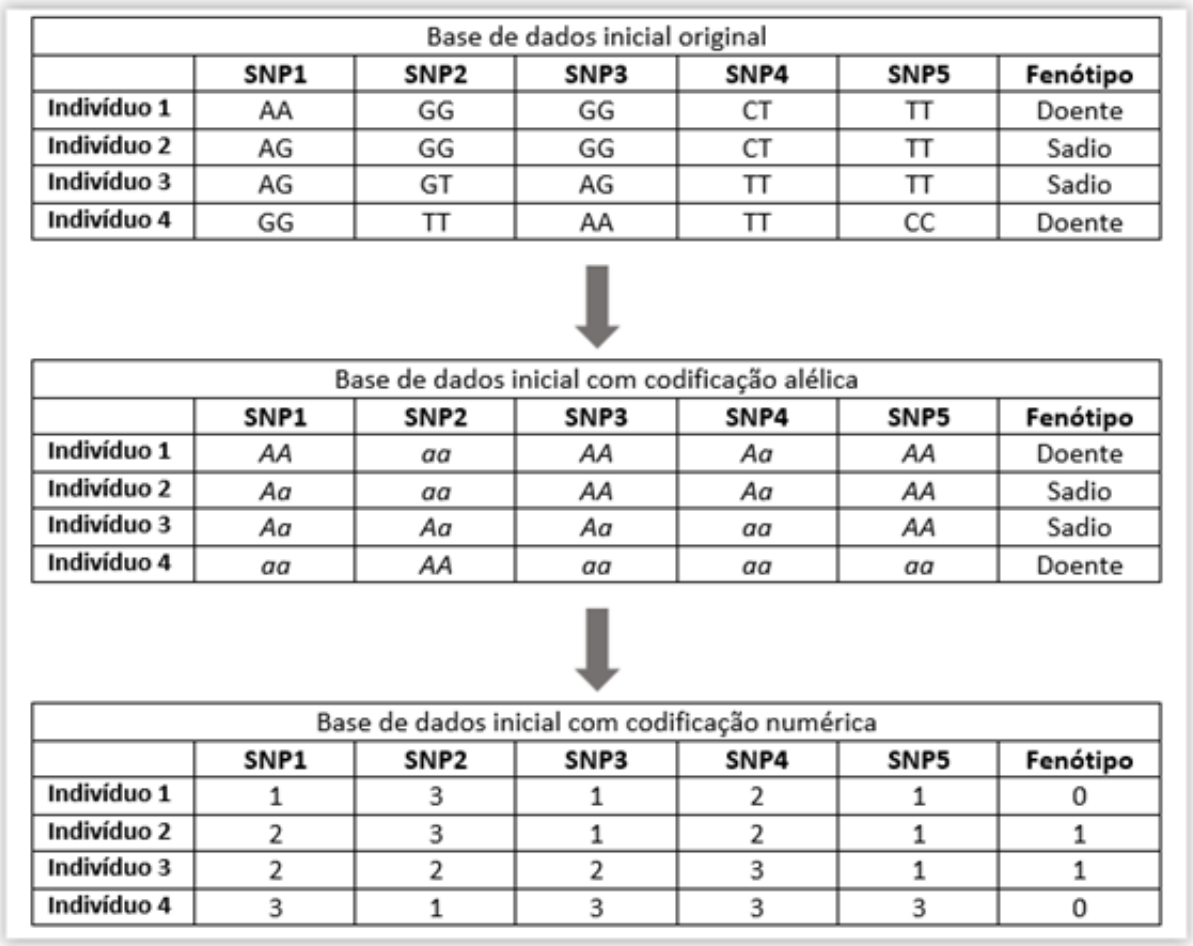
	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Touro 1	0	2	0	1	0	60,94
Touro 2	1	2	0	1	0	70,45
Touro 3	1	1	1	2	0	100,34
Touro 4	2	0	2	2	2	500,65

Figura 6.4 Exemplo hipotético do processo de codificação usado no conjunto de dados reais de genótipo-fenótipo de touros Gir.

O pacote do R utilizado para implementar a RF foi o randomForest com a função *randomForest* (LIAW et al., 2009). Para construir o *rank* da RF para cada cromossomo, insere-se o parâmetro *importance* com o valor lógico TRUE. Outro parâmetro importante para o controle da variabilidade da importância de cada marcador é a semente aleatória (parâmetro *seed*) que foi valorada como sendo o número atual das i execuções do SMS.

6.3.4 O Support Vector Machine usado no SMS

O pacote do R usado para a implementação do SVM/SVR foi o e1071, que oferece uma interface da libsvm em C++ implementada por Chih-Chung Chang e Chih-Jen Li (MEYER; WIEN, 2014; DIMITRIADOU et al., 2009). A libsvm é bastante flexível, permitindo trabalhar tanto em problemas de classificação (C-classificação, ν-classificação, uma-classe-classificação (detecção de novidade), quanto em problemas de regressão (ϵ-



Base de dados inicial original

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Indivíduo 1	AA	GG	GG	CT	TT	Doente
Indivíduo 2	AG	GG	GG	CT	TT	Sadio
Indivíduo 3	AG	GT	AG	TT	TT	Sadio
Indivíduo 4	GG	TT	AA	TT	CC	Doente

Base de dados inicial com codificação alélica

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Indivíduo 1	*AA*	*aa*	*AA*	*Aa*	*AA*	Doente
Indivíduo 2	*Aa*	*aa*	*AA*	*Aa*	*AA*	Sadio
Indivíduo 3	*Aa*	*Aa*	*Aa*	*aa*	*AA*	Sadio
Indivíduo 4	*aa*	*AA*	*aa*	*aa*	*aa*	Doente

Base de dados inicial com codificação numérica

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Indivíduo 1	1	3	1	2	1	0
Indivíduo 2	2	3	1	2	1	1
Indivíduo 3	2	2	2	3	1	1
Indivíduo 4	3	1	3	3	3	0

Figura 6.5 Exemplo hipotético do processo de codificação usado no conjunto de dados do SCRIME para classificação.

regressão e ν-regressão), além de aceitar o uso de diferentes funções *kernel* tais como linear, polinomial, radial, e sigmóide (CHANG; LIN, 2011). A função utilizada do pacote e1071 foi a *svm*, porém, a validação cruzada com 10-*fold* foi implementada separadamente com semente aleatória fixada em 123. Esse procedimento, realizado por um *script* separado da função *svm*, possibilita que todas as partições construídas na validação cruzada com 10-*fold* seja a mesma em qualquer etapa do SMS, o que não foi possível a partir do valor 10 inserido no parâmetro *cross* da função *svm*. Isto é, a função *svm* alterava a partição usada para o mesmo grupo de variáveis em execuções distintas do algoritmo.

Para os problemas de classificação, o parâmetro adotado para o *kernel* linear foi $C = 1$. Para o *kernel* radial, além do parâmetro $C = 1$, variou-se o parâmetro γ de 0,001 até 1 com passo multiplicativo 10. O parâmetro γ é estipulado como um parâmetro fixo para cada execução do SMS, sendo o mesmo usado em todas as fases do SMS (SVR/SVM+RF

e GA) em que se utiliza o SVR/SVM.

Para os problemas de regressão, os parâmetros do SVR para o *kernel* linear foram $C = 1$ e $\epsilon = 0,1$. Esses mesmos parâmetros são usados pelo *kernel* radial, sendo que o parâmetro γ varia da mesma maneira que nos problemas de classificação (SVM).

6.3.5 O Algoritmo Genético usado no SMS

O algoritmo genético utilizado no SMS é baseado na função *ga* do pacote GA do *software* R explanado por Scrucca (2012). No caso da seleção de atributos, a codificação de cada indivíduo do GA deve ser binária, onde o 1 indica a seleção do marcador e 0 a exclusão do mesmo para a formação de um indivíduo (cromossomo) do GA, que é uma base de dados formada a partir de outra composta m instâncias e $n + 1$ variáveis (marcadores SNP mais o fenótipo). A função de aptidão adotada para cada indivíduo do GA foi a média do coeficiente de correlação de Pearson em uma validação cruzada com 10-*fold*. A justificativa para essa escolha baseia-se na observação de que o uso do MSE predito com 10-*fold* como função de aptidão nas bases simuladas do SCRIME demonstrou soluções finais de marcadores SNP que possuíam correlação média próximas a 0, indicando baixa acurácia na predição. Os parâmetros do GA tais como geração da população inicial, tamanho da população, método de seleção para reprodução, probabilidades de *crossover* e de mutação, foram utilizados com os valores padrões da função *ga* explicadas a seguir.

O comprimento do cromossomo de cada indivíduo no GA é o número de marcadores que foi selecionado após o processo do SVR/SVM sobre a ordenação da RF. Como o objetivo do GA é realizar uma busca pelo subconjunto com a maior correlação média em 10-*folds*, a codificação natural é a binária. Assim, por exemplo, suponha que foram selecionados cinco marcadores após a terceira etapa (SVR + RF), um possível indivíduo da população inicial do GA seria $(0, 1, 0, 1, 1)$, o qual irá gerar uma matriz de treinamento-teste com m linhas e 3 colunas (consideram-se somente as colunas que possuem 1), onde as variáveis explicativas são os SNPs 2, 4 e 5 (os SNPs 1 e 3 são mascarados), e a variável explicada são os correspondentes valores fenotípicos. A Figura 6.6 elucida a estrutura dos indivíduos do GA e o cálculo de suas aptidões a partir de uma população do GA gerada aleatoriamente com três indivíduos (grupo de variáveis selecionadas e não selecionadas), além do cálculo da correlação dos valores previstos gerados pelo SVR com os valores fenotípicos observados para cada *fold*. Finalmente, a média dos quatro *folds* é

computada e adotada como aptidão de cada indivíduo (grupo de variáveis selecionadas e não selecionadas).

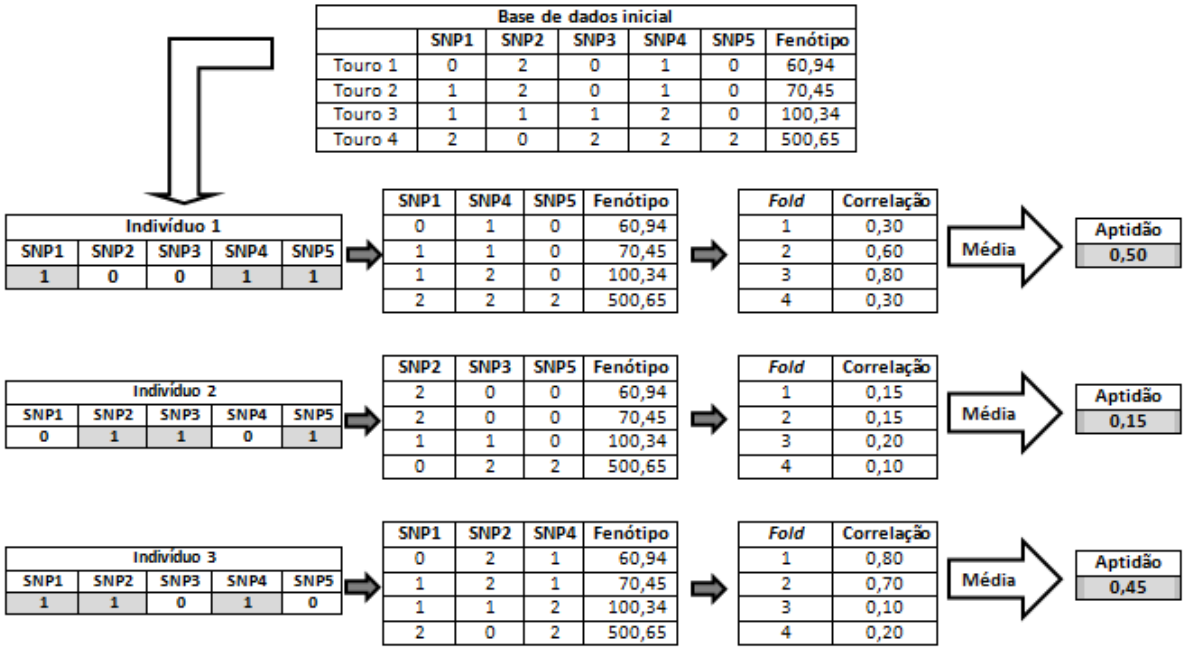


Figura 6.6 Exemplo de população inicial gerada aleatoriamente com 3 indivíduos pelo GA com codificação binária e suas respectivas aptidões computadas a partir da validação cruzada com 4-*fold* em um conjunto de dados inicial com fenótipo contínuo.

A população inicial do GA consta de 100 indivíduos, onde 99 foram gerados aleatoriamente e 1 indivíduo foi formado a partir dos marcadores selecionados pela etapa anterior ao GA (SVR/SVM sobre o *rank* da RF) que, obrigatoriamente, participará dessa população, com o intuito de garantir que o melhor subconjunto de marcadores criado pelos operadores genéticos do GA ao final da primeira geração apresente desempenho igual ou superior ao subconjunto selecionados após a primeira seleção (SVR/SVM sobre o *rank* da RF).

O método de seleção de indivíduos para o GA foi baseado no *ranking* linear proposto por (BAKER, 1985). Primeiramente, ordena-se de forma crescente todos indivíduos de 1 até N, a partir de seu valor dado pela função de aptidão original (correlação de Pearson média em 10-*fold*), e em seguida, avalia-se para cada indivíduo sua nova função de aptidão dada pela Expressão 6.1. Onde o Min é o valor da avaliação atribuído ao pior indivíduo do *ranking* (*rank* 1), Max é o valor da avaliação designado ao melhor indivíduo do *ranking* (*rank* N), N é o número de indivíduos da população e $rank(i, t)$ é o *ranking* do indivíduo i na geração t. Como relata (MITCHELL, 1998), o valor Max é escolhido pelo usuário,

além de ter-se as seguintes restrições $Max \geq 0$, $\sum_{i=1} E(i,t) = N$ (desde que a população mantenha seu tamanho inicial de geração para geração), $1 \leq Max \leq 2$ e $Min = 2 - Max$. O valor recomendado por (BAKER, 1985) é $Max = 1,1$, por conseguinte $Min = 0,9$, pois o autor mostrou que este esquema é favorável comparativamente com a seleção pela avaliação proporcional com roleta viciada em alguns problemas de teste (MITCHELL, 1998). Essa abordagem de seleção é para evitar a convergência prematura e a dominância de um "superindivíduo", pois caso o mesmo tenha uma avaliação muito superior a todo o restante da população e, supondo que o método de seleção pela roleta viciada seja usado, esse indivíduo terá uma probabilidade de ser selecionado muito superior aos outros (LINDEN, 2008), o que ocasiona perda de diversidade genética. Além disso, a pressão seletiva permanecerá constante ao longo de todas as gerações, não importando o grau de convergência genética que tenha ocorrido na população no decorrer do GA (LINDEN, 2008).

$$E(i,t) = Min + (Max - Min)\left[\frac{rank(i,t) - 1}{N - 1}\right] \tag{6.1}$$

O *crossover* utilizado foi operador de um ponto. Um ponto de corte é a posição entre dois genes de um cromossomo, como por exemplo, a Figura 6.7 indica um cromossomo com 5 genes, o que implica em 4 possibilidades para pontos de corte. Genericamente, um cromossomo com n genes possui $n - 1$ pontos de corte possíveis. Após a seleção dos pais pelo módulo de seleção, um ponto de corte é selecionado aleatoriamente a partir do parâmetro probabilidade de *crossover* especificado pelo usuário, que no caso do SMS é igual a 0,8. Suponha um exemplo com 5 pontos de corte, é atribuída uma probabilidade uniforme igual a 0,2 a cada ponto possível para o corte, e com isso, pode-se calcular uma distribuição de probabilidade acumulada 0,25 para o ponto 1, 0,50 para o 2, 0,75 para o 3 e 1,00 para o 4. Um número entre 0 e 1 é sorteado, por exemplo 0,55 e é escolhido o ponto de corte que tem a probabilidade acumulada mais próxima a esse ponto, que nesse exemplo o ponto 2. Um exemplo de aplicação do operador genético *crossover* de um ponto é dado na Figura 6.8, onde 2 pais trocam material genético para a formação de 2 filhos. Note que cada pai possui um ponto de corte selecionado de forma aleatória e independente do outro. Assim, depois do sorteio do ponto de cada pai, separa-se cada pai em duas partes: uma antes e outra após o ponto de corte. Em seguida, o filho 1 é gerado pela junção da primeira parte do pai 1 mais a segunda parte do pai 2. O filho 2 é

constituído pela primeira parte do pai 2 em conjunto com a segunda parte do pai 1.

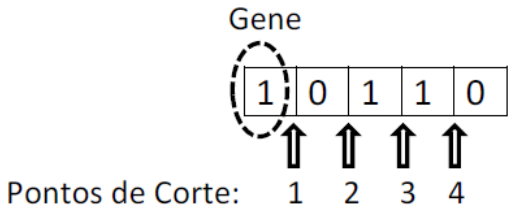


Figura 6.7 Possíveis pontos de corte em um cromossomo com 5 genes para o operador de um ponto.
Adaptado de Linden (2008).

O operador de mutação atua diretamente nos filhos gerados pela ação do operador de *crossover* sobre os pais. A mutação no SMS possui uma probabilidade igual a 0,1. Para cada gene de cada filho, é sorteado um número entre 0 e 1, caso esse número seja menor que 0,1, o operador mutação altera seu valor de 0 para 1 ou vice-versa. As probabilidades de *crossover* e de mutação são fixadas em 0,8 e 0,1 respectivamente, mantendo-se constantes até a convergência do GA ser atingida.

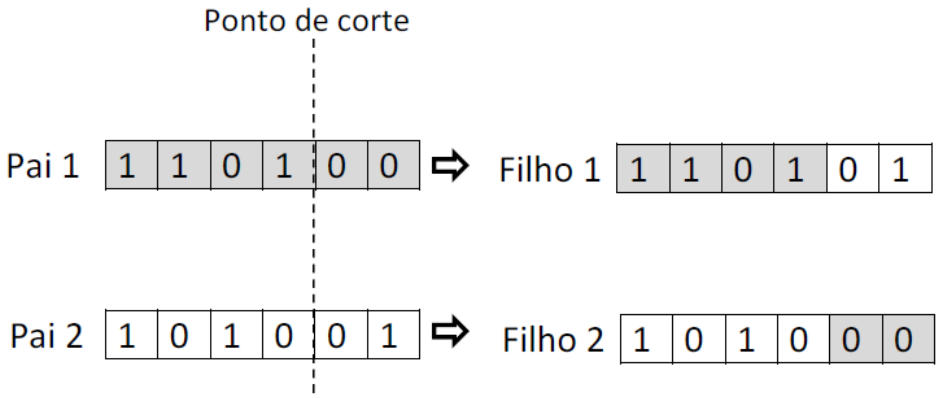


Figura 6.8 Exemplo *crossover* com um operador de um ponto.
Adaptado de Sivanandam e Deepa (2007).

O elitismo foi adotado como forma de garantir que os cinco indivíduos mais aptos fossem transferidos para a próxima geração, pois esse valor é padrão no pacote utilizado GA. Quanto maior o valor desse parâmetro, maior a chance de convergência prematura do GA. Essa estratégia garante que a função de aptidão será uma função não-decrescente ao longo das iterações do GA, isto é, o melhor indivíduo da geração atual possui aptidão

maior ou igual ao melhor indivíduo da geração anterior. O momento de parada do GA foi baseado em dois critérios: o número máximo de gerações ou o número de gerações em que o melhor indivíduo (subconjunto de marcadores) não apresenta melhoria, onde, 10.000 e 30, os respectivos valores destes parâmetros. Assim, o término do GA ocorre quando uma das condições for satisfeita. A técnica de seleção de variáveis baseada na abordagem *wrapper*, onde o SVR/SVM é o algoritmo de aprendizado embutido no GA, usada nesse trabalho é descrita detalhadamente por Kohavi e John (1997).

6.4 Vantagens e Desvantagens do SMS

A RF, usada na etapa de relevância, juntamente com o SVM/SVR e o GA adotados nas etapas de corte e refinamento respectivamente, permitem ao SMS capturar SNPs que demonstrem efeitos aditivos e/ou não-aditivos intralocus (dominância) ou interlocus (epistasia). Como o SMS é uma junção de técnicas não-paramétricas, ele é um método de seleção de atributos não-paramétrico, pois não necessita de premissa alguma sobre a distribuição dos dados de genótipo e de fenótipo. Além dos pontos abordados, o SMS pode capturar interações entre pares ou trios de SNPs simultaneamente sem a necessidade de definir o número de SNPs interagindo, ou seja, o GA do SMS não precisa de adaptação para tratar somente de interações entre pares ou trio de SNPs.

O ponto de corte escolhido pelo SMS é definido de forma automática, não sendo baseado em modelos teóricos ou definidos *a priori* pelo pesquisador como é feito com o valor-p de testes de hipoteses usados em GWAS. Além dessa questão, o ponto de corte é feito por cromossomo, o que permite a possibilidade de detecção de interações de SNPs entre cromossomos, além das interações intercromossômicas.

A constituição modular do SMS baseada nas etapas de relevância, corte e refinamento, permite que outras técnicas possam substituir as que foram usadas nas duas versões do SMS. Por exemplo, pode-se substituir o SVR pelo Blasso na etapa de corte e na função de aptidão do GA feita na etapa de refinamento, ou, até mesmo substituir a ordenação da RF pelo *rank* do Blasso com base na variância de cada marcador na etapa de relevância.

Uma possível desvantagem do SMS é seu tempo de processamento visto que o mesmo é composto de três técnicas distintas (RF, SVM e GA), porém, o benefício é a possibilidade de reduzir o número de SNPs informativos perdidos por outros métodos de seleção que se

direcionam por meio de medidas baseadas em pressupostos restritivos.

6.5 Resumo do Capítulo

Neste capítulo foi apresentado o método proposto SMS e o funcionamento de suas duas versões para mostrar como foi seu processo evolutivo de construção. Duas principais evoluções da primeira para segunda versão foram a substituição do *rank* feito valor-p bruto na fase de relevância pela ordenação da RF e todas as etapas do algoritmo foram unificadas para execução no *software* R. As três técnicas de Inteligência Computacional (RF, SVM/SVR e GA) foram descritas detalhadamente, além das vantagens e desvantagens do SMS.

7 Dados Experimentais

No presente capítulo, todas as simulações realizadas para a geração dos conjuntos de dados serão detalhadas, além de destacar as características distintas dos simuladores SCRIME e LDSO (simulador usado no QTLMAS 2011). O conjunto de dados reais será descrito detalhadamente em relação à codificação do genótipo e à variação do fenótipo.

7.1 Conjunto de Dados Simulados pelo SCRIME

Todos os modelos construídos com o simulador do pacote SCRIME (SCHWENDER; FRITSCH, 2013) do *software* R (TEAM, 2013) mimetizam somente marcadores SNP dentro de um único cromossomo, pois tal pacote não permite a criação de múltiplos cromossomos numa única simulação. Consequentemente, todas as interações geradas entre SNPs são intracromossômicas. Assim, os seis cenários simulados são baseados em 1.000 indivíduos com 100 marcadores em um único cromossomo. Outra característica da simulação do SCRIME a ser destacada é que a população criada não é baseada em gerações sucessivas, o que não permite o controle do parâmetro da herdabilidade h^2 por esse pacote. Apesar das limitações desse simulador, o mesmo possui a possibilidade de simular interações de qualquer ordem, o que não é possível em outros simuladores disponíveis para fenótipos contínuos.

Para os modelos do SCRIME, a MAF escolhida aleatoriamente a partir de uma distribuição uniforme com limites inferior e superior respectivamente iguais a 0,10 e 0,40 para todos os cinco modelos do SCRIME estão apresentadas na Tabela 7.1. Nota-se que o SNP6 ficou com a menor MAF (0,11) e isso pode dificultar a seleção desse marcador, porém, deve-se ressaltar que uma magnitude adequada escolhida para o coeficiente beta desse SNP pode compensar sua MAF pequena e permitir a sua detecção por algum método de seleção.

Tabela 7.1 MAF dos SNPs usados pelos 5 modelos do SCRIME.

SNPs causais	1	2	3	4	5	6	7	8	9
MAF	0,19	0,34	0,22	0,37	0,38	0,11	0,26	0,37	0,27

Os SNPs simulados pelo SCRIME não possuem parametrização para a distância física

(bp) e nem para a distância genética dada em cM, ou seja, eles são não-ligados. O LD entre entre eles é gerado indiretamente pelos parâmetros necessários à simulação e em todas as simulações ele não excedeu o valor de $r^2 = 0,30$. Portanto, nesse contexto, o SNP será considerado verdadeiro-positivo somente se o método de seleção encontrar exatamente o SNP usado na construção do modelo do SCRIME. Diferentemente, a análise do resultado nos dados do QTLMAS 2011 irá considerar como verdadeiro-positivo o marcador que estiver mais próximo fisicamente (que também é próximo geneticamente) do QTL em questão. Desta forma, a distância em centiMorgan (cM) será a medida de acurária do método de seleção.

7.1.1 Simulação 1 - Oito efeitos aditivos para regressão

O modelo 1 foi gerado para simular somente fenótipo contínuo. Foram definidos oito marcadores SNPs causais designados como SNP1, SNP2, SNP3, SNP4, SNP5, SNP6, SNP7 e SNP8. O simulador utilizado foi baseado no pacote *SCRIME* do R com o uso da função *simulateSNPglm* com os parâmetros tamanho da população igual a 1.000 indivíduos e número de marcadores igual a 100.

Para a simulação dessa base de dados, foi usado a função *simulateSNPglm* do pacote *scrime* do *software* R. Esse pacote considera a codificação 1 para homozigoto de referência AA, 2 para o heterozigoto Aa e 3 para o homozigoto variante aa.

O modelo gerado a partir da função *simulateSNPglm* é descrito pela Equação 7.1.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_5 + \beta_6 L_6 + \beta_7 L_7 + \beta_8 L_8 + erro \qquad (7.1)$$

Onde o *erro* é uma variável aleatória com média 0 (zero) e desvio padrão 1 (um), $L_1 = (SNP1 = 1)$, $L_2 = (SNP2 = 2)$, $L_3 = (SNP3 = 3)$, $L_4 = (SNP4 \neq 1)$, $L_5 = (SNP5 \neq 2)$ e $L_6 = (SNP6 \neq 3)$, $L_7 = (SNP7 = 1)$ e $L_8 = (SNP8 = 2)$. A Expressão $SNP1 = 1$ significa o efeito do SNP1 será destacado quando o mesmo for homozigoto de referência AA (codificado como 1), enquanto que $SNP4 \neq 1$ indica que o $SNP4$ terá seu efeito aumentado quando for diferente de 1 (diferente do homozigoto de referência AA), ou seja, quando o $SNP4$ apresentar indivíduos com Aa e aa, ele demonstrará relevância. Interpretações análogas podem ser usadas para as expressões

para os outros SNPs.

Para os coeficientes β_i com $i = 1, 2, 3, 4, 5$, foram atribuídos os seguintes valores: $\beta_0 = 640$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$, $\beta_4 = 1$, $\beta_5 = 1$, $\beta_6 = 1$, $\beta_7 = 1$ e $\beta_8 = 1$. Os valores para os betas foram escolhidos por tentativa e erro a fim de gerar uma distribuição próxima à da PTA do leite (fenótipo dos dados reais). Foram gerados 100 marcadores com 1.000 indivíduos, sendo a MAF simulada para cada SNP advinda de uma distribuição uniforme com os limites mínimo e máximo iguais a, respectivamente, 0,10 e 0,40. A variável contínua Y é uma combinação linear das variáveis L_i, a qual é simulada com base em um modelo de regressão múltipla linear.

O teste de normalidade de Shapiro-Wilk, desenvolvido por Shapiro e Wilk (1965), apresentou valor-p igual a $2{,}20 \times 10^{-1}$ (maior que $\alpha = 0,05$), o que sugere, a um nível de significância de 0,05, a normalidade para o fenótipo simulado. A Figura 7.1 elenca o histograma (A) e o *boxplot* (B) do fenótipo simulado, e a partir da mesma percebe-se que os valores fenotípicos são simétricos e não apresentam pontos aberrantes.

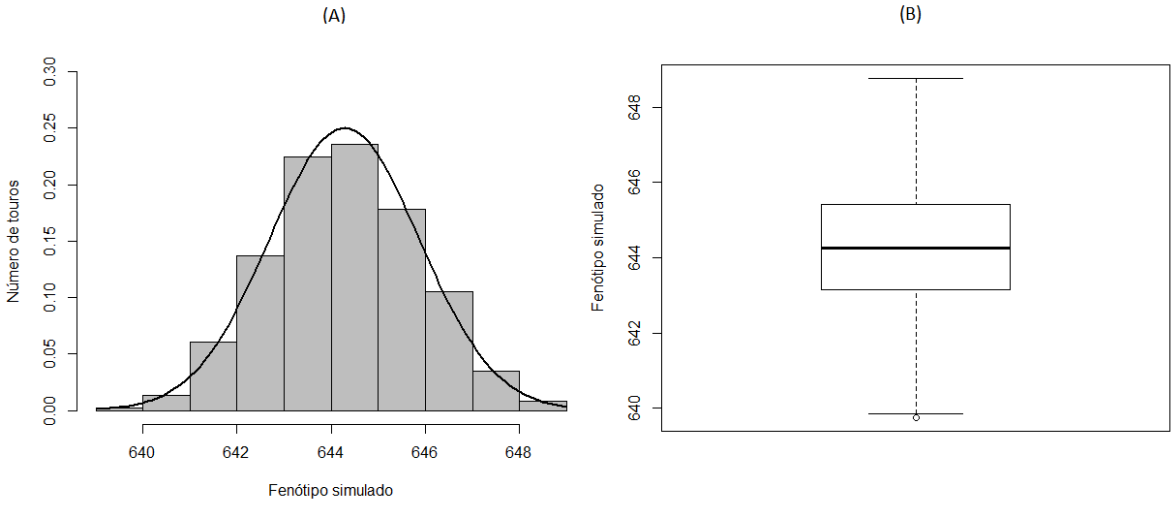


Figura 7.1 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação 1. (A) Histograma e (B) *boxplot* do fenótipo simulado pelo modelo 1.

7.1.2 Simulação 2 - Quatro interações de ordem 2 para regressão

Esse modelo foi construído para simular somente interações entre pares de SNPs de mesma intensidade. Assim, o modelo gerado a partir da função *simulateSNPglm* é mostrado pela

Equação 7.3.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + erro \tag{7.2}$$

Onde o *erro* é uma variável aleatória com média 0 (zero) e desvio padrão 1 (um), $L_1 = (SNP1 = 1)\&(SNP2 = 2)$, $L_2 = (SNP3 = 3)\&(SNP4 \neq 1)$, $L_3 = (SNP5 \neq 2)\&(SNP6 \neq 3)$, $L_4 = (SNP7 = 1)\&(SNP8 = 2)$.

Para os coeficientes β_i com $i = 1, 2, 3, 4, 5$, foram atribuídos os seguintes valores: $\beta_0 = 640$, $\beta_1 = 2$, $\beta_2 = 2$, $\beta_3 = 2$, $\beta_4 = 2$ e $\beta_5 = 2$. Os parâmetros para o número de indivíduos, o número de marcadores e a MAF foram os mesmos usados na simulação 1.

Neste caso, o teste de Shapiro-Wilk indicou um valor-p igual a $2{,}85\times10^{-4}$ (menor que $\alpha = 0,05$), ou seja, o fenótipo apresenta evidências de não seguir uma distribuição normal, apesar da Figura 7.2 demonstrar que o fenótipo gerado é também bastante simétrico. Esse fato pode ter ocorrido devido a relação genótipo-fenótipo ser baseada apenas em 4 interações de ordem 2.

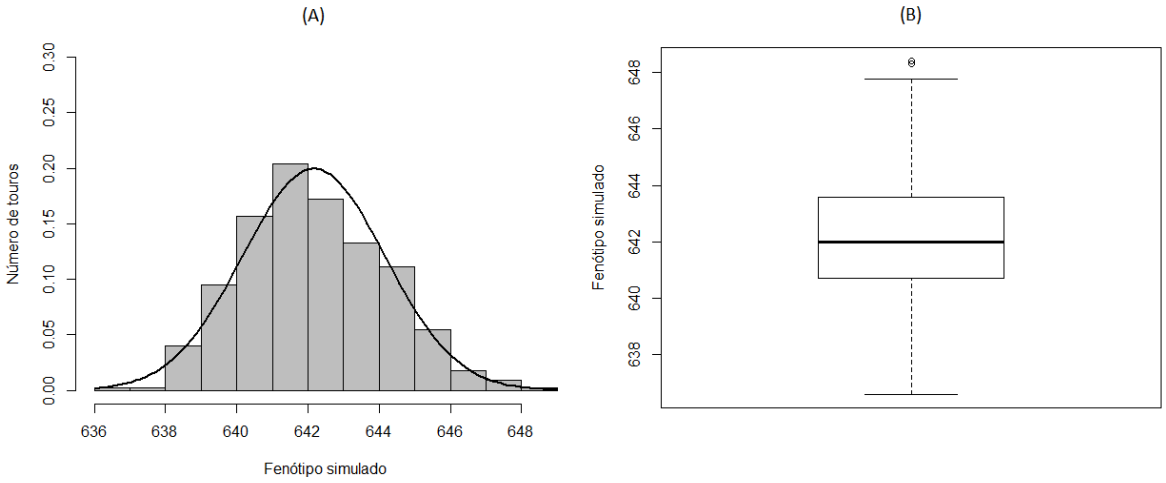


Figura 7.2 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação 2. (A) Histograma e (B) *boxplot* do fenótipo simulado pelo modelo 2.

7.1.3 Simulação 3 - Três interações de ordem 3 para regressão

Esse modelo foi construído para simular somente interações entre trios de SNPs de mesmo efeito. Assim, o modelo gerado a partir da função *simulateSNPglm* é mostrado pela

Equação 7.3.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + erro \tag{7.3}$$

Onde o *erro* é uma variável aleatória com média 0 (zero) e desvio padrão 1 (um), $L_1 = (SNP1 = 1)\&(SNP2 = 2)\&(SNP3 = 3)$, $L_2 = (SNP4 \neq 1)\&(SNP5 = 2)\&(SNP6 = 3$ e $L_3 = (SNP7 = 1)\&(SNP8 \neq 2)\&(SNP9 = 3)$.

Para os coeficientes β_i com $i = 1, 2, 3, 4, 5$, foram atribuídos os seguintes valores: $\beta_0 = 640$, $\beta_1 = 3$, $\beta_2 = 3$ e $\beta_3 = 3$. Os parâmetros para o número de indivíduos, o número de marcadores e a MAF foram os mesmos usados na simulação 1.

O histograma da Figura 7.3 demonstra que o fenótipo gerado é também simétrico, apesar de possuir alguns valores aberrantes como pode ser visto no *boxplot*. O teste de normalidade de Shapiro-Wilk foi realizado e o valor-p encontrado foi $2{,}072 \times 10^{-11}$ que é menor que o valor de significância $\alpha = 0,05$, logo, existem evidências que o fenótipo não segue a distribuição normal.

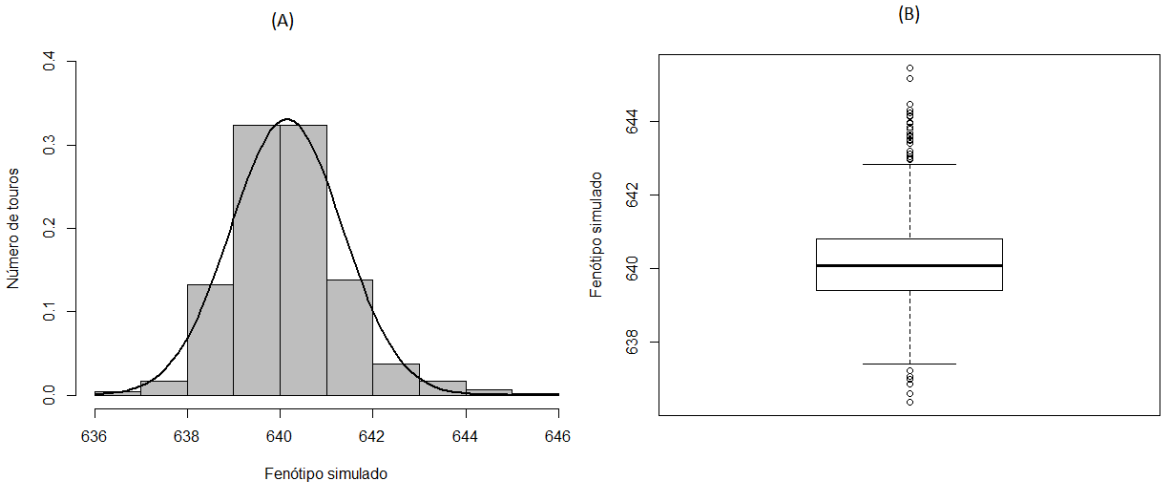


Figura 7.3 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação 3. (A) Histograma e (B) *boxplot* do fenótipo simulado gerado pelo modelo 3.

7.1.4 Simulação 4 - Três efeitos aditivos + uma interação de ordem 2 + uma interação de ordem 3 para regressão

Esse modelo foi construído para simular efeitos aditivos isolados com diferentes magnitudes e efeitos não aditivos significativos para demonstrar o potencial do SMS

em relação a outros métodos de seleção. Assim, o modelo gerado a partir da função *simulateSNPglm* é mostrado pela Equação 7.4.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_5 + erro \tag{7.4}$$

Onde o *erro* é uma variável aleatória com média 0 (zero) e desvio padrão 1 (um), $L_1 = (SNP1 = 1)$, $L_2 = (SNP2 = 2)$, $L_3 = (SNP3 = 3)$, $L_4 = (SNP4 \neq 1)\&(SNP5 = 3)$, $L_5 = (SNP6 = 1)\&(SNP7 = 2)\&(SNP8 = 3)$. A Expressão $L_4 = (SNP4 = 1)\&(SNP5 = 2)$ designa a interação entre os $SNPs$ 4 e 5, ou seja, quando o $SNP4$ for igual a 1 (genótipo Aa) e, simultaneamente, o $SNP5$ for igual a 2 (genótipo aa), o efeito da interação será destacado pelo coeficiente β_4. Interpretação análoga pode ser dada para a expressão $L_5 = (SNP6 = 1)\&(SNP7 = 2)\&(SNP8 = 3)$ que traduz a interação entre o trio de marcadores formado pelos SNPs 6, 7 e 8, sendo a mesma potencializada pelo coeficiente β_5.

Para os coeficientes β_i com $i = 1, 2, 3, 4, 5$, foram atribuídos os seguintes valores: $\beta_0 = 640$, $\beta_1 = 2$, $\beta_2 = 1,3$, $\beta_3 = 0,9$, $\beta_4 = 2$ e $\beta_5 = 3$. Os valores para os betas foram escolhidos para simular efeitos de ações gênicas aditivas (SNPs 1, 2 e 3) distintas e decrescentes, e efeitos não-aditivos proporcionais ao número de SNPs interagindo, isto é, coeficiente 2 para a interação entre o par de SNPs 4 e 5, e coeficente 3 para interação entre o trio de SNPs 6, 7 e 8. Os parâmetros para o número de indivíduos, o número de marcadores e a MAF foram os mesmos usados na simulação 1.

O teste de normalidade de Shapiro-Wilk indicou um valor-p igual a $2{,}39 \times 10^{-3}$ (menor que $\alpha = 0,05$). Com isso, a um nível de significância de 0,05; há evidências de que o fenótipo simulado não segue uma distribuição normal, apesar do histograma da Figura 7.4 ser praticamente simétrico.

7.1.5 Simulação 5 - Somente uma interação de ordem 4 para regressão

Esse modelo foi construído para simular somente uma interação entre quádruplas de SNPs. Assim, o modelo gerado a partir da função *simulateSNPglm* é mostrado pela Equação 7.5.

$$Y = \beta_0 + \beta_1 L_1 + erro \tag{7.5}$$

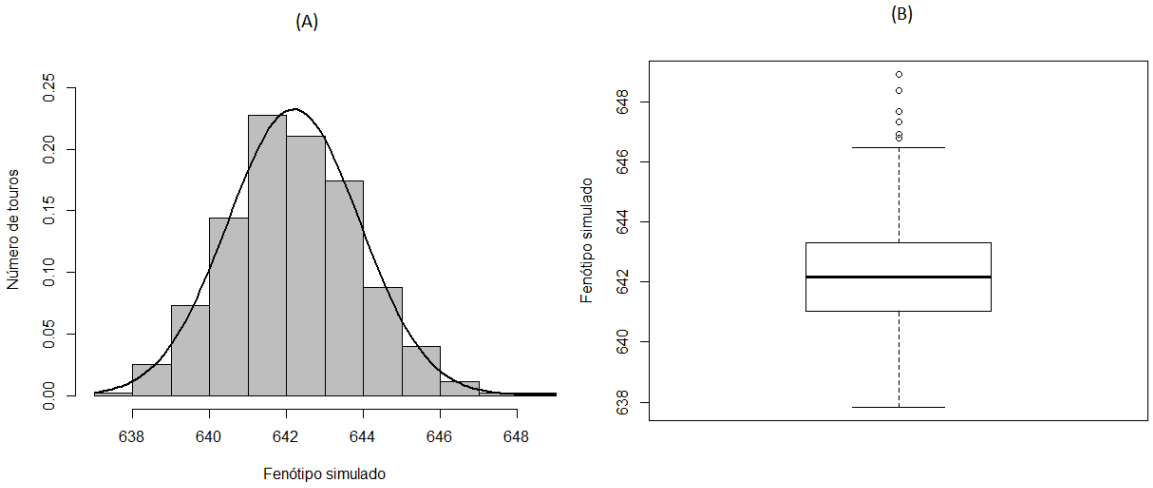


Figura 7.4 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação 4. (A) Histograma e (B) *boxplot* do fenótipo simulado pelo modelo 4.

Onde o *erro* é uma variável aleatória com média 0 (zero) e desvio padrão 1 (um), $L_1 = (SNP1 = 1)\&(SNP2 = 2)\&(SNP3 = 3)\&(SNP4 = 1)$. Para os coeficientes $\beta_0 = 640$ e $\beta_1 = 4$. Os parâmetros para o número de indivíduos, o número de marcadores e a MAF foram os mesmos usados na simulação 1.

Neste caso, o teste de Shapiro-Wilk indicou um valor-p igual a $3{,}817 \times 10^{-62}$ (menor que $\alpha = 0,05$), ou seja, o fenótipo demonstra evidências de não seguir uma distribuição normal, apesar do histograma da Figura 7.5 ser simétrico. Na simulação, alguns valores fenotípicos foram aberrantes em relação à mediana mostrada no *boxplot* (item (B) da Figura 7.5). Esse fato pode ter acontecido devido a relação genótipo-fenótipo ser baseada apenas em uma interação de ordem 4.

7.1.6 Simulação 6 - Três efeitos aditivos + uma interação de ordem 2 + uma interação de ordem 3 para classificação

A simulação 6 foi construída para mostrar que existe a possibilidade de utilizar o SMS em seleção de marcadores SNPs para problemas do tipo caso-controle, isto é, problemas de classificação. Esses problemas advêm de GWAS realizados em humanos, animais e plantas com relação a detecção de *loci* responsáveis pelo aumento do risco de desenvolvimento de doenças. Foi realizada a simulação de efeitos aditivos isolados com diferentes magnitudes e efeitos não-aditivos significativos de ordem 2 e 3 para demonstrar o potencial do SMS

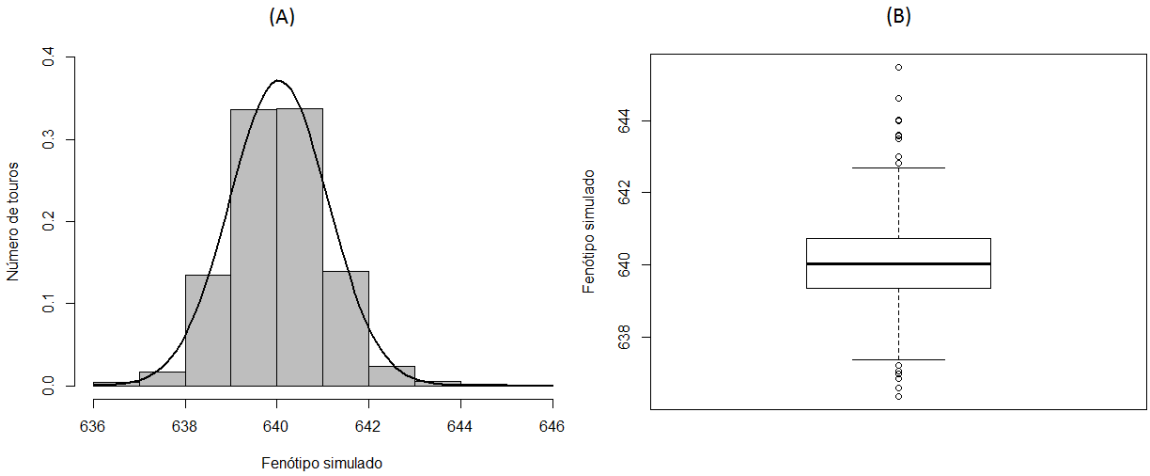


Figura 7.5 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação 5. (A) Histograma e (B) *boxplot* do fenótipo da simulação 5.

em relação a outros métodos de seleção. O número de controles (codificados como 0) e de casos (codificados como 1) foram iguais a, respectivamente, 138 e 862, o que mostra um cenário de classes desbalanceadas. Assim, a função que gera as classes é baseada em um modelo de regressão logística implementada a partir da função *simulateSNPglm* do pacote SCRIME do R e dada pela Expressão 7.6.

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_5 + erro \tag{7.6}$$

As variáveis explicativas são descritas como: $L_1 = (SNP1 = 1)$, $L_2 = (SNP2 = 2)$, $L_3 = (SNP3 = 3)$, $L_4 = (SNP4 \neq 1)\&(SNP5 = 3)$, $L_5 = (SNP6 = 1)\&(SNP7 = 2)\&(SNP8 = 3)$. A Expressão $L_4 = (SNP4 \neq 1)\&(SNP5 = 3)$ designa a interação entre os $SNPs$ 4 e 5, ou seja, quando o $SNP4$ for diferente de 1 (genótipo homozigoto de referência AA) e, simultaneamente, o $SNP5$ for igual a 3 (genótipo homozigoto variante aa), o efeito da interação será destacado pelo coeficiente β_4. Interpretação análoga pode ser dada para a expressão $L_5 = (SNP6 = 1)\&(SNP7 = 2)\&(SNP8 = 3)$ que traduz a interação entre o trio de marcadores formado pelos SNPs 6, 7 e 8, sendo a mesma potencializada pelo coeficiente β_5.

Para os coeficientes β_i com $i = 1, 2, 3, 4, 5$, foram atribuídos os seguintes valores: $\beta_0 = 0$, $\beta_1 = 2$, $\beta_2 = 1,3$, $\beta_3 = 0,9$, $\beta_4 = 2$ e $\beta_5 = 3$. Os valores para os betas foram escolhidos para simular efeitos de ações gênicas aditivas (SNPs 1, 2 e 3) distintas e decrescentes, e efeitos não-aditivos proporcionais ao número de SNPs interagindo, isto

é, coeficiente 2 para a interação entre o par de SNPs 4 e 5, e coeficente 3 para interação entre o trio de SNPs 6, 7 e 8. Os parâmetros para o número de indivíduos, o número de marcadores e a MAF foram os mesmos usados na simulação 1.

Na geração de cenários de classificação, é possível definir a probabilidade de um indivíduo (instância) ser um caso (codificado como 1), cujo parâmetro da função *simulateSNPglm* é denominado *p.cutoff* e foi adotado como 0,7. Deste modo, um número aleatório entre zero e um é sorteado a partir de uma distribuição uniforme, sendo atribuído o valor 1 ao indivíduo se o número sorteado é menor ou igual a 0,7, caso contrário, o valor 0, assim, aproximadamente, 70% dos indivíduos serão casos, e 30% controles.

7.2 Dados simulados do QTLMAS 2011

Os dados simulados foram propostos para avaliação de diversas técnicas de GWAS e GWS no 15º *workshop* QTLMAS 2011 (QTL **M**apping and **M**arker **A**ssisted **S**election), a fim de comparar o mapeamento de QTLs e técnicas de predição usadas em seleção genômica. A estrutura de marcadores SNPs é semelhante às situações encontradas em populações de animais, com um SNP em cada 0,05 cM (correspondente a um chip 60K para um genoma clássico com 3.000 cM), uma MAF média de 0,23, e um LD médio (0,05 cM) entre *loci* próximos igual a 0,27, semelhante aos resultados anteriormente descrito em bovinos (MCKAY et al., 2007). A relação de co-ascendência exibe uma grande variabilidade conforme o esperado em raças reais (ELSEN et al., 2012).

O conjunto de dados simulado no *workshop* QTLMAS 2011 mimetiza a estrutura familiar de suínos sob um modelo oligogênico, onde cada QTL é especificado inicialmente. O primeiro desafio é a seleção de marcadores SNPs que estão associados aos QTLs simulados, enquanto que o segundo desafio, é a determinação dos efeitos dos QTLs, o que é feito a partir dos marcadores selecionados na primeira etapa.

De acordo com Elsen et al. (2012), a população foi uma coleção de 20 famílias de porcos não independentes. Cada macho foi acasalado com 10 fêmeas, sendo que cada fêmea acasalou com apenas um reprodutor. Cada fêmea procriou dois conjuntos de 10 e 5 filhos, respectivamente. O primeiro grupo de progênie (n = 2.000 indivíduos) formaram a população experimental, com genótipos de referência e com informações fenotípicas. O segundo grupo com 1.000 indivíduos eram candidatos à seleção, possuindo

somente informação genômica referente aos marcadores genéticos. A intenção é predizer o fenótipo a partir do genótipo para selecionar animais superiores em relação à característica considerada.

A geração parental (20 machos e 200 fêmeas) foi gerada por uma amostra aleatória de dois gametas escolhidos a partir de um conjunto de 75 gametas. Esta grade de gametas 2 por 75 foi gerada após uma longa evolução de deriva genética aleatória e mutação simulada pelo *software* LDSO (YTOURNEL et al., 2012). A evolução da população ocorreu em duas etapas: 1.000 gerações de uma população compreendendo 1.000 gametas, seguido por uma restrição severa com 150 gametas evoluindo durante 30 gerações.

O genoma simulado consiste de 5 cromossomos autossômicos de 1 Morgan. SNPs bialélicos foram simulados, localizados a cada 0,05 cM (2.000 SNPs por cromossomo). O conjunto de 1.000 gametas foi gerado na primeira geração em equilíbrio de ligação. Durante as 1.150 gerações seguintes a este passo inicial, uma taxa de mutação de 0,0002 foi aplicada no processo (ELSEN et al., 2012).

A arquitetura genética da característica quantitativa foi provavelmente muito mais simples do que a maioria das situações prevalecentes para as características de produção: apenas 8 QTLs segregando, um ou dois por cromossomo (ELSEN et al., 2012). Diferentes tipos de relações alélicas foram escolhidos: aditividade para um único QTL com o maior efeito (cromossomo 1), genes ligados (cromossomas 2 e 3), um recurso de *imprinting* no cromossomo 4 e dois *loci* epistáticos no cromossomo 5 (ELSEN et al., 2012). Esta situação simplificada foi escolhida de propósito para evitar um possível efeito de confusão devido ao ruído poligênico e para enfatizar as habilidades das técnicas em relação ao lidar com tais casos extremos (ELSEN et al., 2012). Todas as propriedades dos 8 QTLs simulados estão descritas na Tabela 7.2.

No cromossomo 1, um QTL (QTL1) com 4 alelos, exibindo grandes efeitos aditivos (0.0, 2.0, 4.0 e 6.0 TU para alelos 1-4) foi posicionada perto da fronteira cromossomo (2,85 cM). O desvio entre os genótipos extremas (44 vs. 11) foi, assim, 12 TU, ou seja, cerca de 1,28 desvios-padrão fenotípicos. Os cromossomas 2 e 3 foram atribuídos a dois QTLs aditivos ligados mostrando um efeito alélico de 1-TU, agindo "em fase"no cromossoma 2, e "em repulsão"no cromossomo 3. A expressão "fase"e "repulsão"deveria ser clarificada na nossa contexto. Quatro classes nos cromossomos 2 e 3 respectivamente foram observados na última geração, definido pelos alelos presentes no QTL2 e QTL3 (respectivamente

Tabela 7.2 Características dos QTLs simulados. Adaptado de Elsen et al. (2012).

QTL	**Cr**	**Posição (cM)**	**Tipo**	**Efeitos**			
QTL1	1	2,85	4 alelos, aditivo e grande	Alelo 1 = 0,0; 2 = 2,0; 3 = 4,0; 4 = 6,0			
QTL2	2	81,90	em fase com QTL3		11	12	22
				11	-4	-2	0
QTL3	2	93,75	em fase com QTL2	12	-2	0	2
				22	0	2	4
QTL4	3	5,00	em oposição com QTL5		11	12	22
				11	0	2	4
QTL5	3	15,00	em oposição com QTL4	12	-2	0	2
				22	-4	-2	0
QTL6	4	32,20	*Imprinting*	11	12	21	22
				2	0	0	0
QTL7	5	36,30	Epistático com QTL8		11	12	22
				11	2	1	0
QTL8	5	99,20	Epistático com QTL7	12	0	0	0
				22	0	0	0

QTL4 e QTL5.): 1-1, 1-2, 2-1 e 2-2. As associações 1-1 e 2-2 sendo mais frequentes do que o 1-2 ou 2-1 em ambos os casos, que recebem a mesma direção dos efeitos de alelos 1 (respectivamente 2) e QTL2 em 1 (respectivamente 2) em QTL3, e alelos 1 (respectivamente 2) em QTL4 e 2 (respectivamente 1) em QTL5. O cromossomo 4 foi caracterizado por um QTL com *imprinting* genômico[1] de efeito moderado (2 TU). Todos os indivíduos que receberam um alelo do seu pai apresentaram um fenótipo quantitativo com 2 TU a mais, em comparação com os indivíduos recebendo alelo 2. No cromossomo 5, dois QTLs epistáticos foram posicionados com distância significativa um em relação ao outro. O efeito da QTL7 foi expresso (com valores médios de 0, 1 e 2 para genótipos 11, 12 e 22) apenas quando os animais apresentaram genótipo 11 no QTL8.

A codificação usada para o genótipo do conjunto de dados inicial foi $AA = 11$ para o homozigoto de referência, $Aa = 12$ para o heterozigoto ou 21 e $aa = 22$ para o homozigoto variante. Ademais, nenhum filtro de controle de qualidade (MAF, equilíbrio de Hardy-Weinberg, *call-rate*) foi aplicado no conjunto codificado de entrada no SMS. A justificativa para a não aplicação desses filtros está baseada na verificação da robustez do SMS em eliminar marcadores com pequenas MAFs. Além disso, o conjunto de dados do QTLMAS 2011 não possui genótipo ausente, por isso, não faz sentido o uso do filtro gerado pela *call-rate*.

A variabilidade do fenótipo foi devido à segregação de 8 QTLs e ao ruído ambiental. Os QTLs foram gerados, transformando SNPs que ainda foram polimórficos na última

[1]Significa a expressão diferencial do material genético quando o mesmo é herdado do macho ou da fêmea(PIERCE, 2013).

geração. Estes SNPs foram então removidos do conjunto de marcadores para representar o que geralmente ocorre em situações reais. O QTL localizado no cromossoma 1 foi gerado por divisão de alelos em dois SNPs adjacentes, a fim de criar um locus quadri-alélico. As características dos QTLs variaram entre os 5 cromossomos e foram escolhidos para representar situações extremas conforme Tabela 7.2. Os efeitos dos QTLs são dadas em unidades de "tratamento"(UT). A variância do ruído do ambiente foi ajustada à variação genética observada devido aos efeitos aditivos de QTL, a fim de dar uma herdabilidade de 0,30. O desvio-padrão fenotípico resultante foi de 9,37 UT.

Para a seleção de marcadores no QTLMAS 2011, diversos técnicas foram usadas, porém, as mesmas podem ser classificadas em duas categorias, a primeira denominada genômica (global), onde todos os SNPs são avaliados simultaneamente em uma única etapa, e uma local, onde os SNPs são testados um por vez (ELSEN et al., 2012). No grupo global, o método GBLUP assume que todos os marcadores contribuem para a característica [(NADAF et al., 2012), (ZENG et al., 2012)], enquanto que todos os outros métodos consideram que o conjunto total de SNPs é uma mistura composta de uma pequena parte dos SNPs que influenciam o fenótipo, e outra grande parte composta por SNPs neutros. Esse tipo de abordagem foi resolvida por diferentes métodos LASSO (o LASSO clássico usado por Nadaf et al. (2012) foi comparado com duas novas estratégias utilizadas por Usai, Carta e Casu (2012)) e por técnicas MCMC Bayes: Bayes A (NADAF et al., 2012), Bayes B [(NADAF et al., 2012), (ZENG et al., 2012)], Bayes C (DASHAB et al., 2012) e Bayes Cπ [(ZENG et al., 2012),(SCHURINK; JANSS; HEUVEN, 2012)]. Dashab et al. (2012) comparou diferentes maneiras de processar as informações dos marcadores e entre elas se destacou a clusterização de haplótipos baseado em genealogias locais usando o modelo GENMIX de Sahana et al. (2011).

Foi aplicado o teste de normalidade de Shapiro-Wilk ao fenótipo simulado e o mesmo indiciou valor-p igual a 0,72 (maior que $\alpha = 0,05$), logo existem evidências de que o mesmo segue uma distribuição normal com média 6,88 e desvio-padrão 9,20. Há alguns valores fenotípicos aberrantes como pode ser visto no *boxplot* da Figura 7.6, porém, eles não influeciaram no formato da distribuição a ponto de diferir de uma normal. O fenótipo simulado se situa entre os limites -24,48 e 36,96 (Tabela 7.3), o que mostra um ampla variação dada pela influência do genótipo e do meio-ambiente.

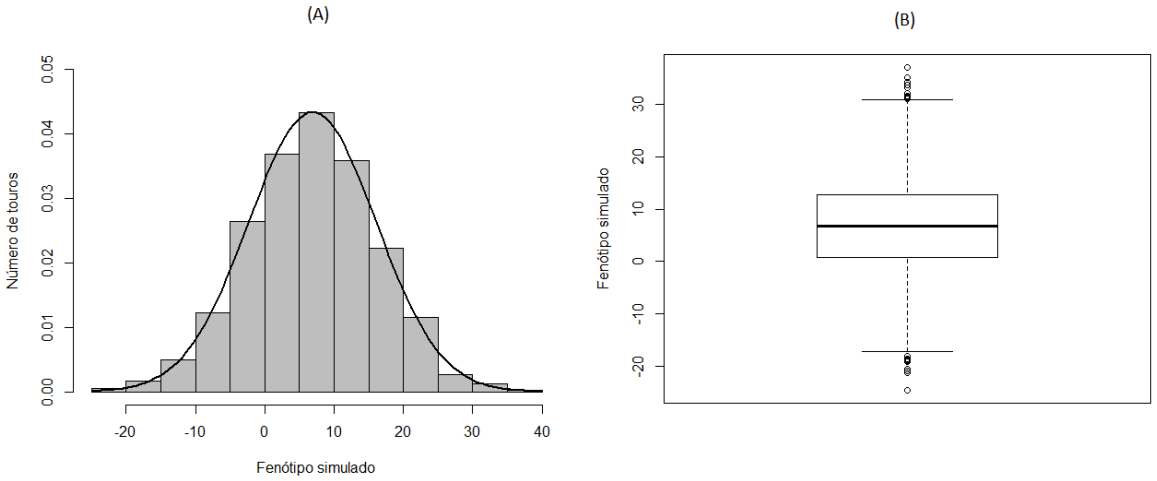


Figura 7.6 Histograma e *boxplot* do fenótipo contínuo gerado pela simulação feita pelo LDSO usado no QTLMAS 2011.

(A) Histograma e (B) *boxplot* do fenótipo simulado pelo LDSO no QTLMAS 2011.

Tabela 7.3 Medidas descritivas do fenótipo simulado no QTLMAS 2011.

Mínimo	1° Quartil	Mediana	Média	3º Quartil	Máximo
-24,48	0,76	6,79	6,88	12,81	36,96

7.3 Conjunto de Dados Reais

Inicialmente, uma discussão será feita sobre o fenótipo usado nesse estudo de associação. Em seguida, todas as características do genótipo serão descritas, além do controle de qualidade que foi realizado no mesmo.

7.3.1 A PTA do leite

A avaliação genética é um processo pelo qual faz-se a previsão do valor genético [2] dos animais com base em uma ou mais características (VALENTE et al., 2001). O principal objetivo desse procedimento é ordenar os indivíduos em uma amostra ou população a fim de selecionar os melhores indivíduos e descartar os piores. Diversas metodologias podem ser usadas, desde o mais simples e menos preciso, baseando-se somente em dados de produção do animal, até procedimentos mais sofisticados, em que informações genômicas e de parentesco são agregadas. Os métodos mais sofisticados baseiam-se em modelos estatísticos, requisitando processos computacionais demorados e complexos (VALENTE et al., 2001).

A eficiência de um programa de seleção para produção de leite é mais dependente da seleção de touros do que de vacas, pois os touros produzem mais descendentes do que as vacas, principalmente se a inseminação artificial for também usada (VALENTE et al., 2001). Consequentemente, pode-se praticar maior intensidade de seleção de machos do que de fêmeas (VALENTE et al., 2001). Outro ponto importante a considerar é que a produção de leite não pode ser medida diretamente nos touros, logo, tal característica deve ser medida em parentes próximos do sexo feminino, sendo denominada de pseudo-fenótipo.

O trabalho de Silva et al. (2014) explica detalhadamente esse tipo de característica. A capacidade prevista ou predita de transmissão (do inglês, *Predicted Transmitting Ability*-PTA) de um touro é computado a partir da produção de leite de sua prole feminina com base na metodologia desenvolvida em Verneque et al. (2012). A PTA é uma medida do desempenho esperado das filhas do touro em relação à média genética dos rebanhos (VERNEQUE et al., 2012). Assim, por exemplo, uma PTA de 500 kg para produção de

[2] Representa o que o animal transmite à progênie. Significa o quanto da diferença em produção em relação à média da população ou em relação às companheiras de rebanho que o animal transmite para os descendentes (VALENTE et al., 2001).

leite significa que, se o touro for usado numa população com nível genético igual ao usado para avaliá-lo, cada filha produzirá em média 500 kg por lactação a mais do que a média do rebanho (VERNEQUE et al., 2012). A PTA é a metade do valor genético e é o termo usado quando a avaliação genética é executada usando-se o modelo animal (VERNEQUE et al., 2012). A metodologia aplicada para a medição da PTA e os resultados obtidos estão descritos detalhadamente em Verneque et al. (2012).

Segundo (VALENTE et al., 2001), o modelo animal tornou-se a base para a avaliação genética de vacas e touros nos Estados Unidos em 1989. O uso desse método tem-se tornado frequente. Nesse modelo, a PTA do leite é obtida com base na avaliação simultaneamente de vacas e touros, baseando-se no histórico de produção das vacas, diferentemente dos modelos que avaliam separadamente machos e fêmeas. Assim, o modelo animal objetiva produzir preditores para o valor genético do animal. VanRaden e Wiggans (1991) demonstram os cálculos da PTA do leite de forma detalhada.

Para o cálculo da PTA do leite, somente o efeito genético do animal é considerado, reduzindo-se a maioria de fatores não-genéticos que influenciam na produção de leite das vacas tais como os efeitos externos (ambientais) e/ou internos (fisiológicos). Alguns efeitos externos são devido a região, o rebanho, as diferenças sazonais de ano para ano, etc (VALENTE et al., 2001). Como fatores internos ou de natureza fisiológica pode-se citar: idade, gestação, lactação, efeitos maternos entre outros (VALENTE et al., 2001). A maioria desses fatores podem ser medidos e os seus efeitos sobre a produção conhecidos, possibilitando o estabelecimento de padrões de variação (VALENTE et al., 2001). Dentre os mais importantes: duração da lactação, número de ordenhas, idade da vaca, época de parição, período de serviço e período seco (VALENTE et al., 2001). Assim, é coerente relacionar a PTA do leite de um touro, que contempla praticamente só informação genética relativa à produção de leite de sua progênie feminina, com informações genômicas de marcadores moleculares, com a finalidade de selecionar os SNPs informativos para esse fenótipo. Por conseguinte, busca-se a explicação do valor genético por meio de informação genômica dada pela variação alélica dos marcadores tipo SNP mais relevantes.

7.3.2 Descrição dos Dados

Para demonstrar a metodologia proposta, foi usada uma amostra de 343 touros genotipados da raça Gir (raça *Bos indicus* brasileira) fornecida pela Empresa Brasileira

de Pesquisa Agropecuária Gado de Leite (Embrapa Gado de Leite). Dos 343 animais, somente 240 possuem prole feminina, permitindo a mensuração da PTA do leite, que neste caso é o fenótipo considerado. A PTA do leite é uma característica que possui herdabilidade igual a 0,28 ($h^2 = 0,28$) para essa população de touros Gir (VERNEQUE et al., 2012). Os valores da PTA não foram derregredidos, pois as acurárias das PTAs dos touros foram todas superiores a 0,70. Esse conjunto de dados é parte do projeto de pesquisa descrito em Arbex et al. (2010) e no Apêndice A.1, que descreve sobre os termos de uso do conjunto de dados reais.

O genoma do bovino tem aproximadamente 3 bilhões de pares de bases e possui 30 pares de cromossomos, sendo 29 pares autossômicos e um par sexual. O genótipo dos touros foi gerado a partir do *Illumina BovineSNP50kv2 BeadChip* contendo um total de 56.947 marcadores.

As variáveis explicativas, descritas pela frequência de ocorrência do alelo B no *locus*, foram codificadas da seguinte maneira: $AA = 0$ (ausência do alelo a), $Aa = 1$ (presença de uma cópia do alelo a) e $aa = 2$ (presença de duas cópias do alelo a). Os valores faltantes, devido a erros de leitura, foram considerados como heterozigoto $Aa = 1$. Essa codificação é descrita detalhadamente em Illumina (2014).

7.3.3 Pré-processamento

Para o controle de qualidade (CQ) da base de dados real foram aplicados os filtros *call-rate* $\geq 0,95$, $MAF \geq 0,05$ e $HWE \geq 0,05/56.947$, sendo 0.05 a significância do HWE, 56.947 a quantidade de SNPs na base original e 0,05/56.947 o limite de corte do HWE com correção de Bonferroni. Após a aplicação dos filtros descritos acima, restaram 22.845 marcadores para a aplicação do método de seleção SMS, por conseguinte, a redução total promovida pelo controle de qualidade foi de 34.102 SNPs.

A Figura 7.7 demonstra o número de marcadores por cromossomo antes e após o controle de qualidade, sendo o cromossomo 99 um artifício para designar SNPs que estão presentes no *chip* de genotipagem, mas que não foram mapeados para nenhum dos 30 cromossomos. Note que os cromossomos 6 e X tiveram respectivamente, a menor e a maior redução na quantidade de SNPs após o controle de qualidade.

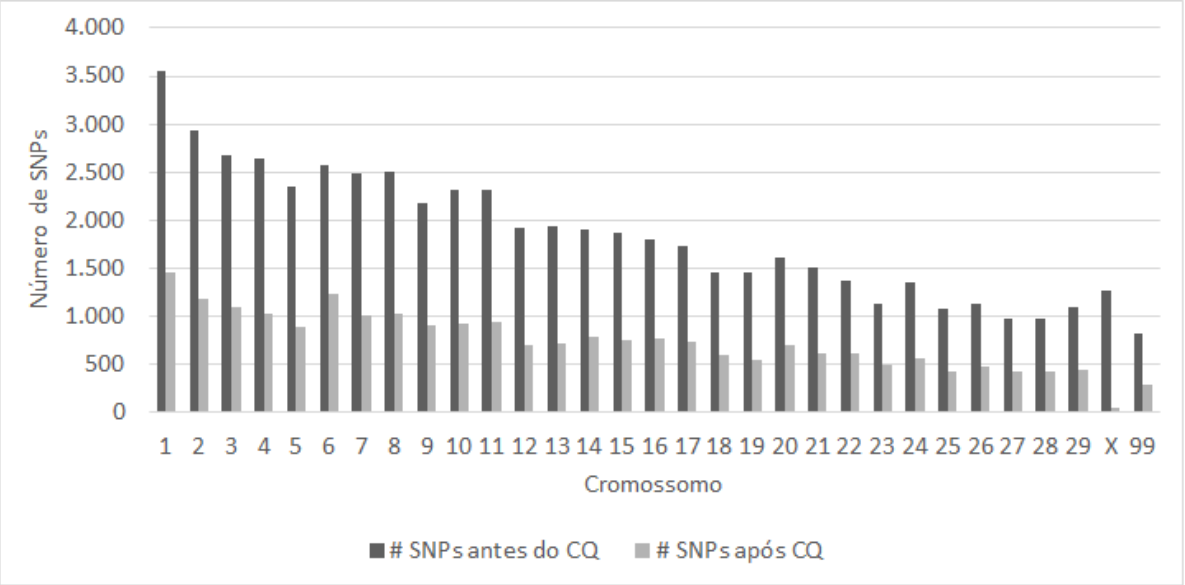


Figura 7.7 Número de marcadores SNPs antes e após o controle de qualidade (CQ).

7.4 Resumo dos Dados Experimentais

A Tabela 7.4 sintetiza os oito conjuntos de dados compostos de informações de genótipo (SNPs causais) e do tipo de variável que descreve os valores fenotípicos (contínuo ou binário) e mostra a diversidade de cenários que o SMS foi testado. A herdabilidade no sentido amplo simbolizada por H^2 na sétima coluna da Tabela 7.4 foi calculada pela fórmula $H^2 = 1 - \frac{var(erro)}{var(y)} = 1 - \frac{1}{var(y)}$ para as simulações realizadas pelo simulador SCRIME para fenótipo contínuo (simulaçõe de 1 a 5), onde $var(y)$ e $var(erro) = 1$ significam a variância do fenótipo e a variância do erro respectivamente. A herdabilidade da simulação 6 não foi calculada devido a falta de acessibilidade dessa medida nos resultados da simulação.

Tabela 7.4 Resumo das características dos conjuntos de dados usados para avaliar o método SMS.

Tipo dos dados	**Simulador**	**# SNPs causais**	**# Cr.**	**Fenótipo**	**LD**	H^2	**Interação**
Simulação 1	SCRIME	8	1	Contínuo	Baixo	0,61	Ordem 1 (aditivo)
Simulação 2	SCRIME	8	1	Contínuo	Baixo	0,68	Ordem 2
Simulação 3	SCRIME	9	1	Contínuo	Baixo	0,75	Ordem 3
Simulação 4	SCRIME	8	1	Contínuo	Baixo	0,31	Ordem 1, 2 e 3
Simulação 5	SCRIME	4	1	Contínuo	Baixo	0,13	Ordem 4
Simulação 6	SCRIME	8	1	Binário	Baixo	-	Ordem 1, 2 e 3
Simulação 7	QTLMAS 2011	8	5	Contínuo	Alto	0,30	Ordem 1 e 2
Real	-	Desconhecido	30	Contínuo	Alto	0,28	Desconhecido

7.5 Resumo do Capítulo

Foram gerados seis conjuntos de dados simulados com diferentes tipos de efeitos aditivos e/ou não aditivos no SCRIME para avaliar o SMS, porém, esses cenários possuem baixo desequilíbrio de ligação, o que não representa fielmente a realidade, visto que existe LD em dados reais. Entretanto, pode-se abordar esses SNPs gerados pelo SCRIME como *tag* SNPs sem perda de generalização e avaliar o SMS nesse contexto. Os dados do QTLMAS 2011 já possuem uma estrutura de LD próxima às bases de dados reais, o que permite uma análise do desempenho do SMS em cenários mais complexos. Finalmente, os dados reais trazem consigo vários obstáculos tais como incerteza, amostra pequena, dados faltantes, variabilidade no genótipo e no fenótipo e LD entre os SNPs, gerando um grande desafio para o SMS.

8 Experimentos Computacionais

O objetivo desse capítulo é comparar a seleção de marcadores em cada conjunto de dados construídos por meio do simulador SCRIME e por meio do simulador LDSO usado para construir o conjunto de dados (genótipo-fenótipo) do QTLMAS 2011. Os métodos de seleção comparados foram valor-p bruto, valor-p corrigido (correção de Bonferroni) e Blasso para os modelos do SCRIME enquanto que para o QTLMAS 2011, diversas técnicas de seleção foram usadas durante a competição e comparadas posteriormente com o SMS.

8.1 Parâmetros dos Métodos de Seleção

8.1.1 Valor-p Bruto e Valor-p Corrigido

Para os métodos valor-p bruto e valor-p corrigido usou-se o limite superior de 0,05 para selecionar os marcadores relevantes, ou seja, somente os SNPs com valor-p menor que 0,05 são selecionados. A função *lm* do R foi usada para estimar os coeficientes angulares das retas de regressão e calcular seus valores-p para fenótipos contínuos. Para fenótipos binários, a função *chisq.test* do R foi usada para realizar o teste de associação Qui-Quadrado.

8.1.2 Blasso

Para a seleção de SNPs informativos pelo Blasso, foi usado a variância explicada por cada marcador em relação ao fenótipo, com limite inferior igual a 0,01 (1%) da variância para marcadores informativos, isto é, somente os SNPs com variância maior que 0,01 são selecionados. Esse critério torna a seleção de marcadores rigorosa de forma a só selecionar SNPs que realmente expliquem uma fração importante da característica, apesar da maioria dos trabalhos que usam o Blasso para seleção de SNPs, utilizarem o percentual da variância explicada por janela (grupo) de SNPs como é realizado por Utsunomiya et al. (2014).

O pacote BLR foi usado para implementar o Blasso detalhado em Campos e Pérez (2010). Os valores adotados para σ_e^2 e σ_u^2 foram respectivamente iguais a 3,446538 e

1,089324, o grau de liberdade (do inglês, *degree of freedom* - *df*) igual a 2 e os parâmetros $\alpha_1 = 0,5$ e $\alpha_2 = 10^{-4}$ da distribuição Gamma como descrito em Pérez et al. (2010). O número de iterações da cadeia de Markov foi 20.000 (parâmetro $nIter = 20.000$), o número de iterações para o aquecimento da mesma cadeia foi 10.000 (parâmetro $burnIn = 10.000$) e o número de descartes da cadeia de Markov com objetivo de reduzir a autocorrelação foi igual a 1 (parâmetro $thin = 1$)[1]. Com esses parâmetros adotados para o Blasso, a cadeia de Markov convergiu em todos os seis cenários simulados e a convergência foi verificada pelo pacote do R denominado *boa* descrito por Smith et al. (2007).

8.1.3 SMS

A versão atual do SMS foi usada para seleção dos SNPs nos conjuntos de dados simulados do SCRIME e do QTLMAS 2011, e no conjunto de dados reais da Embrapa, porém, neste conjunto, alguns resultados da primeira versão(simbolizada por SMS1) foram inseridos a título de comparação com a versão atual (simbolizada por SMS2). Em alguma parte do texto que não seja especificada a versão do SMS, a mesma deverá ser adotada como sendo a atual.

Como o SMS permite diversas possibilidades de escolha para o subconjunto de SNPs selecionado, adotou-se como solução final, a união dos melhores subconjuntos construídos pelos cinco *kernels* utilizados no SMS para os problemas de regressão. Isso é justificado pelo desconhecimento *a priori* sobre a relação genótipo-fenótipo no conjunto de dados, além da flexibilidade criada pela junção de modelos com paradigmas distintos, o que permite aos modelos lineares encontrar os SNPs que possuem somente efeitos marginais aditivos, aos modelos não-lineares, os SNPs que somente apresentam efeitos não-aditivos (somente interações) e, finalmente, à união dos modelos lineares e não-lineares detectarem os SNPs que possuem ambos efeitos.

Para os problemas de regressão dados pelas simulações 1, 2, 3, 4 e 5; foram adotados para o SVR, o *kernel* linear e o *kernel* radial com γ variando de 0,001 a 1 com passo multiplicativo 10 foram usados como *mix* de modelos do SMS. O número de iterações do SMS para cada *kernel* foi $i = 10$, sendo que as sementes aleatórias da RF e do GA assumem os mesmos valores de i . A união dos SNPs selecionados por cada *kernel* avaliado será considerada a solução final do SMS. Para a RF utilizou-se $ntree = 4.000$ (número

[1]Para maiores detalhes sobre o parâmetro *thin* ver Link e Eaton (2012).

de árvores por floresta), $mtry = p = 100$ (número total de marcadores para construção de cada árvore na floresta). A ordenação realizada pela RF foi baseada na importância de permutação devido à propriedade preditiva desse *rank*. Para problemas de regressão, os parâmetros do SVR foram $C = 1$ e $\epsilon = 0,1$. O número de marcadores avaliados na etapa de corte foi $limitCHR = 0,95$ (95% de cada cromossomo) e adotou-se $k = 2$ passos de comprimento 10 após o menor MSE médio do SVR sobre o *rank* da RF. Para o GA, usou-se $run = 30$ (número máximo de gerações do GA sem melhoria), $iter = 10.000$ (número máximo de gerações do GA), $pcross = 0,8$ (a probabilidade de *crossing over*), $pmut = 0,1$ (probabilidade de mutação), $elitism = 5$ (número de melhores indivíduos do GA que irão para próxima geração sem alteração alguma) e a correlação média na validação cruzada com 10-*fold* para função de aptidão a ser maximizada.

Para o problema de classificação dado pela simulação 6, o parâmetro do SVM foi $C = 1$, pois na classificação não é necessário o ϵ, além disso, os mesmos *kernels* para regressão foram avaliados. A ordenação realizada pela RF foi baseada na importância de gini (gVI) devido ao melhor desempenho desse *rank* na detecção de interações entre variáveis (resultados não mostrados). Para a etapa de corte foi avaliada a média da área abaixo da curva ROC (média da AUC) em 10-*fold* no ponto com distância do mínimo igual a um passo de 10 marcadores. No GA, a única diferença em relação aos problemas de regressão, é a função de aptidão, que neste caso é a média da área abaixo da curva ROC em 10-*fold*. A união dos subconjuntos de SNPs escolhidos nas dez execuções do SMS com um único *kernel* que apresentou maior AUC média e menor número de marcadores. Essa abordagem foi adotada, pois o número de marcadores selecionados por alguns *kernels* foram praticamente todos os SNPs pertencentes ao conjunto inicial, ou seja, as etapas de corte e refinamento praticamente não eliminaram SNP algum.

Para os dados simulados do QTLMAS 2011 e para os dados reais, os *kernels* linear e radial com γ variando de 0,001 a 0,1 com passo multiplicativo 10 serão usados como *mix* de modelos do SMS. O $\gamma = 1$ não foi utilizado, pois ele gerou somente predições iguais para os 2.000 indivíduos do QTL e para os 240 touros do conjunto real, o que gerou desvio-padrão zero, não sendo possível o cálculo da correlação usada como função de aptidão no GA. O número de iterações do SMS para cada *kernel* foi $i = 3$ para o QTL e $i = 1$ para o conjunto real, sendo que as sementes aleatórias da RF e do GA assumem os mesmos valores de i, logo, no caso do QTLMAS 2011, são selecionados três

subconjuntos distintos de SNPs devido às alterações nas sementes aleatórias da RF e do GA. Com isso, não será realizada a união das três iterações do SMS para o QTLMAS 2011, mas o objetivo é verificar a estabilidade das soluções geradas quando a semente aleatória da RF e do GA são alteradas. Para a RF utilizou-se $ntree = 4.000$ (número de árvores por floresta), $mtry = p = 1.998$ (número total de marcadores para construção de cada árvore na floresta) para o QTLMAS 2011 e variou para cada cromossomo conforme a Figura 7.7 para os dados reais. Para o SVR, os parâmetros foram $C = 1$ e $\epsilon = 0,1$. O número de marcadores avaliados na etapa de corte foi $limitCHR = 0,20$ (20% de cada cromossomo) para o QTLMAS 2011 e $limitCHR = 0,95$ (95% de cada cromossomo) para os dados reais. Adotou-se 2 passos de comprimento 10 após o menor MSE médio do SVR sobre o *rank* da RF. Para o GA, usou-se $run = 30$ (número máximo de gerações do GA sem melhoria), $iter = 10.000$ (número máximo de gerações do GA), $pcross = 0,8$ (a probabilidade de *crossing over*), $pmut = 0,1$ (probabilidade de mutação), $elitism = 5$ (número de melhores indivíduos do GA que irão para próxima geração sem alteração alguma) e a correlação média na validação cruzada com 10-*fold* para função de aptidão a ser maximizada.

Os parâmetros *ntree* e *mtry* usados pela RF no problemas de regressão e classificação não são padrões de estudos anteriores na literatura, mas foram encontrados por meio de poucas simulações realizadas durante a fase de teste do SMS. Entretanto, nenhum processo de otimização na busca pelos melhores parâmetros foi feita, cabendo um estudo futuro nessa direção. Em relação aos parâmetros do SVM e do SVM, os parâmetros C e ϵ são padrões do pacote e1071 do R. Os parâmetros do pacote GA foram baseados nos valores padrões do mesmo.

Tanto a RF quanto o GA foram executados em paralelo para acelerar a execução do SMS, onde o algoritmo da RF foi executado em paralelo por 5 processadores e o GA por 64 processadores. O SMS foi executado em um computador AMD Opteron 6242 com 64 núcleos e 264 GB de memória RAM.

8.2 Critérios para Escolha do Melhor Método

O modelo ideal é aquele que seleciona somente os SNPs verdadeiros-positivos e nenhum SNP falso-negativo. Porém, como encontrar os SNPs que marcam os genes que influenciam

determinada característica é uma tarefa complexa devido à existência de ruídos em diversas etapas de GWAS, um modelo desse tipo é praticamente inviável para situações práticas. Consequentemente, é possível adotar critérios mais adequados para processos com diversos níveis de incerteza como são os estudos de associação em escala genômica. Por exemplo, pode-se escolher o melhor modelo como sendo o que seleciona o maior número de SNPs verdadeiros-positivos, mesmo selecionando mais SNPs falsos-positivos, ou, o que seleciona o menor número de SNPs verdadeiros-positivos, mas, o menor número de SNPs falsos-positivos, ou o maior número de SNPs causais com o menor número de SNPs não-causais.

No presente estudo, adotou-se o melhor modelo como sendo aquele que selecinou a maior quantidade de SNPs informativos, mesmo que o mesmo tenha selecionado maior número de SNPs não-informativos. A justificativa para tal escolha é que existe uma probabilidade maior de encontrar os genes que determinam a característica em questão quando selecionam-se o maior número de SNPs causais.

Outra consequência desse critério de escolha pelo melhor método é que, em relação ao tempo computacional, mesmo que o melhor método tenha um tempo de processamento superior aos demais, o que considerou-se mais importante foi a garantia da detectecção da maioria dos SNPs informativos. Isso se justifica pela importância dos desdobramentos dessa seleção de marcadores para estudos posteriores tais como as possíveis variações nas proteínas geradas a partir de mutações encontradas em genes que foram marcados pelos SNPs selecionados.

O tempo computacional não foi explicitado nas simulações do SCRIME, pois tais estudos não retratam a quantidade de SNPs em estudos de GWAS com dados reais atualmente. Por outro lado, em relação aos dados simulados do QTLMAS2011 e aos dados reais fornecido pela Embrapa Gado de Leite, o tempo do SMS foi explicado nos resultados. Ressalta-se que o tempo de processamento dos métodos, em ordem crescente, em todos os conjuntos de dados analisados é valor-bruto e valor-p corrigido, Blasso e SMS, ou seja, o SMS é o que consome mais tempo para selecionar os SNPs.

8.3 Simulação 1 - Oito efeitos aditivos para regressão

Os resultados da seleção por quatro abordagens distintas são demonstrados na Tabela 8.1 e conclui-se, que dentre os cinco modelos distintos de SVR usados pelo SMS, o que encontrou o maior número de SNPs informativos foi o *kernel* linear, o que intuitivamente era esperado, já que o modelo subjacente ao processo de simulação do SCRIME é linear (somente efeitos aditivos). Outra questão interessante observada é que à medida que aumenta-se o γ, o número de marcadores falsos-positivos diminui, todavia, perde-se o marcador 3 para os quatro γs estudados. Assim, surge uma questão: será que existe um $\gamma < 0,001$ capaz de selecionar simultaneamente os marcadores 3 e 6, além dos outros seis SNPs causais? Essa pergunta justifica-se porque o *kernel* linear capturou o SNP 3 mas não selecionou o SNP 6.

A união das soluções do SMS para modelos de SVR distintos gerou um subconjunto de marcadores com mais SNPs não-informativos, enquanto que o conjunto interseção das soluções do SMS não selecionou SNPs falsos-positivos e, consequentemente, demonstrou ser mais restritivo como era esperado. Portanto, a solução final do SMS adotada é o conjunto união com 12 marcadores, sendo sete verdadeiros positivos, pois o SNP 6 não foi detectado, talvez por possuir a menor MAF entre os oito SNPs relevantes.

O método do valor-p bruto foi o que indicou o melhor resultado do ponto de vista dos SNPs causais, pois encontrou todos os oito, por outro lado, inseriu cinco falsos-positivos. O método do valor-p corrigido encontrou somente quatro SNPs, sendo todos verdadeiros, o que mostra que a correção de Bonferroni é extremamente restritiva. O Blasso selecionou seis SNPs causais e somente um falso-positivo, o que mostra um equilíbrio entre SNPs verdadeiros e falsos. O Blasso não encontrou os SNPs 3 e 6.

Os oito SNPs informativos possuem desempenhos distintos em relação ao *kernel* usado pelo SVR como pode ser visto na Tabela 8.2, onde a maior correlação é referente ao *kernel* radial com $\gamma = 0,1$. Esse fato ocorreu também para o subconjunto de SNPs selecionado pelo SMS com *kernel* radial com o mesmo γ como observado na Tabela 8.1. Logo, uma observação importante é que a correlação média em cada *kernel* para os oito SNPs verdadeiros é inferior à correlação média encontrada por cada solução do SMS em cada *kernel*, inclusive para o $\gamma = 0,1$. Isso possibilita concluir que a medida usada para comparar subconjuntos de marcadores não é a ideal, entretanto, as outras medidas avaliadas, MSE e MAPE, também não conseguiram demonstrar melhores resultados que

Tabela 8.1 Resultado da seleção dos SNPs para a simulação 1.

Método	γ	SNPs selecionados	Iter[4]	$\overline{r}$	# SNPs (V)[5]
SMS Linear[1]	-	**7**, **1**, **4**, **8**, **5**, **2**, 51, **3**, 12, 97	4	0,512	10 (7)
SMS Radial[1]	0,001	**7**, **1**, **4**, **8**, **5**, **2**, 83, 51, 12	1	0,513	9 (6)
SMS Radial[1]	0,01	**7**, **1**, **4**, **8**, **5**, **2**, 51	2	0,592	7 (6)
SMS Radial[1]	0,1	**7**, **1**, **4**, **8**, **5**, **2**, 51, 14	1	0,724	6 (2)
SMS Radial[1]	1	**7**, **1**, **4**, **8**, **5**, **2**	1	0,643	6 (6)
União do SMS[6]	-	**1**, **2**, **3**, **4**, **5**, **7**, **8**, 10, 12, 14, 51, 83, 97	-	-	13 (7)
Inteseção do SMS[6]	-	**1**, **2**, **4**, **5**, **7**, **8**	-	-	6 (6)
Valor-p bruto[2]	-	**7**, **1**, **4**, **8**, **5**, **2**, 10, 60, 12, 71, **6**, 88, **3**	-	-	13 (8)
Valor-p corrigido[2]	-	**1**, **7**, **4**, **8**	-	-	4 (4)
Blasso[3]	-	**7**, **1**, **4**, **8**, **5**, **2**, 68	-	-	7 (6)

[1] Corte pelo MSE do SVR sobre o *rank* da RF.

[2] Valor-p $< 0{,}05$.

[3] Variância explicada por cada marcador $> 0{,}01$.

[4] Iteração do SMS com a solução de maior correlação média.

[5] Número de verdadeiros-positivos.

[6] O *rank* não é considerado no subconjunto de SNPs selecionados.

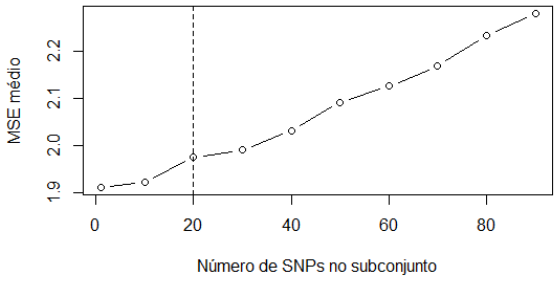
a correlação. Outra análise possível é que SNPs espúrios juntamente com SNPs causais geram sinal superior ao do conjunto formado somente pelos SNPs causais, o que sugere um filtro posterior para eliminar os SNPs falsos-positivos.

Mesmo que a correlação média do subconjunto de SNPs selecionado pelo SMS com um *kernel* seja maior do que em outro, não significa que a seleção de SNPs do primeiro seja melhor do que a do segundo. Um exemplo é dado pelas seleções do *kernels* linear e radial com $\gamma = 0,1$ na Tabela 8.1, onde a seleção do *kernel* linear encontrou mais SNPs causais do que o radial, entretanto, a correlação do primeiro *kernel* (0,512) é menor que a do segundo (0,724). A partir dessas observações, conclui-se que a abordagem multi-*kernel* é essencial para capturar estruturas distintas da relação genótipo-fenótipo.

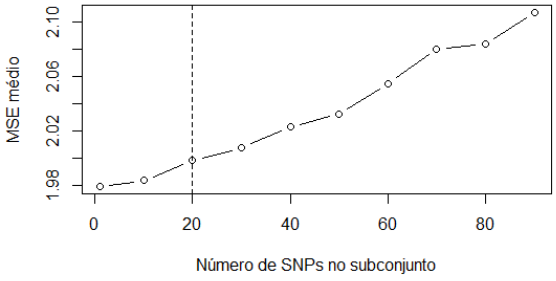
Tabela 8.2 Desempenho dos oito SNPs causais nos cinco *kernels* avaliados para a simulação 1.

kernel	γ	r (σ_r)	MSE (σ_{mse})	MAPE (σ_{mape})
Linear	-	0,506 (0,05)	1,903 (0,23)	0,172 (0,01)
Radial	0,001	0,507 (0,05)	1,971 (0,23)	0,176 (0,01)
Radial	0,01	0,589 (0,04)	1,675 (0,21)	0,162 (0,01)
Radial	**0,1**	**0,716 (0,04)**	**1,246 (0,21)**	**0,139 (0,01)**
Radial	1	0,537 (0,10)	1,814 (0,34)	0,166 (0,01)

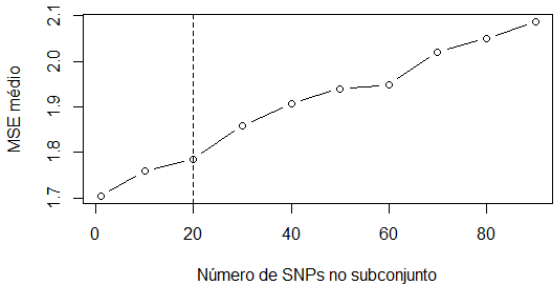
Como pode ser notado na Figura 8.1, o menor MSE médio ocorreu no SNP 7, que é o mais relevante ordenado pela RF (primeiro ponto do item (a) da Figura 8.1), assim, é de suma importância flexibilizar esse ponto de corte para possibilitar a entrada de outros SNPs causais que serão avaliados em conjunto pelo GA na etapa de refinamento, pois, caso contrário, o método SMS só selecionaria o SNP 7 ao final de todo o processo. Observe que no gráfico (a) da Figura 8.1 as variações do MSE médio absoluta e relativa estão entre os limites 1,9 e 2,3 para o *kernel* radial com $\gamma = 0,1$, respectivamente, próximas de 0,40 e 21%.



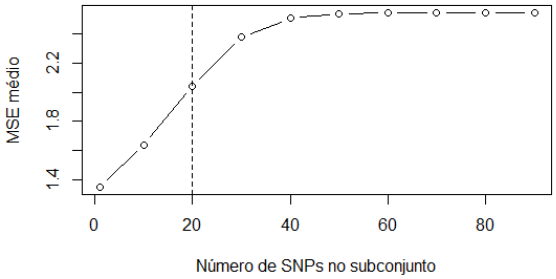
(a) *Kernel* linear na iteração 4.



(b) *Kernel* radial com $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial com $\gamma = 0,01$ na iteração 2.



(d) *Kernel* radial com $\gamma = 0,1$ na iteração 1.

Figura 8.1 Corte do SVR sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 1.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação ao modelo 1. A linha tracejada indica o ponto de corte.

Outro ponto a destacar é a coerência entre as ordenações geradas pelos cinco *kernels* usados no SMS indicados na Tabela 8.3. A RF demonstra plena estabilidade na ordenação dos SNPs 1, 2, 4, 5, 7 e 8, uma ligeira variação para o SNP 3 e uma grande variação para o SNP 6, que não foi selecionado em nenhuma etapa de corte, logo em nenhuma execução

do refinamento (GA). O valor-p bruto foi o único método que ordenou o SNP 6 na posição 11 que está entre os 13 selecionados pelo corte adotado (valor-p $< 0{,}05$), enquanto que o Blasso colocou-o na posição 90, naturalmente, não selecionando-o. O SNP 3 também não foi selecionado pelo Blasso e, no seu *rank*, este marcador ficou distante do corte realizado com variância maior que 0,01.

Tabela 8.3 Ordenação de cada método para os oito SNPs causais para a simulação 1.

Método	Iter[1]	Corte	SNPs causais							
			1	2	3	4	5	6	7	8
RF do SMS Linear	4	20	$2^{\underline{a}}$	$6^{\underline{a}}$	$15^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{54^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$
RF do SMS Radial $\gamma = 0,001$	1	20	$2^{\underline{a}}$	$6^{\underline{a}}$	$17^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{40^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$
RF do SMS Radial $\gamma = 0,01$	2	20	$2^{\underline{a}}$	$6^{\underline{a}}$	$13^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{46^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$
RF do SMS Radial $\gamma = 0,1$	1	20	$2^{\underline{a}}$	$6^{\underline{a}}$	$13^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{31^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$
RF do SMS Radial $\gamma = 1$	1	20	$2^{\underline{a}}$	$6^{\underline{a}}$	$13^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{48^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$
Valor-p bruto	-	13	$1^{\underline{a}}$	$6^{\underline{a}}$	$13^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$11^{\underline{a}}$	$2^{\underline{a}}$	$4^{\underline{a}}$
Valor-p ajustado	-	4	$1^{\underline{a}}$	$6^{\underline{a}}$	$13^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{11^{\underline{a}}}$	$2^{\underline{a}}$	$4^{\underline{a}}$
Blasso	-	7	$2^{\underline{a}}$	$6^{\underline{a}}$	$16^{\underline{a}}$	$3^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{90^{\underline{a}}}$	$1^{\underline{a}}$	$4^{\underline{a}}$

[1] Número da iteração nas 10 execuções do SMS.

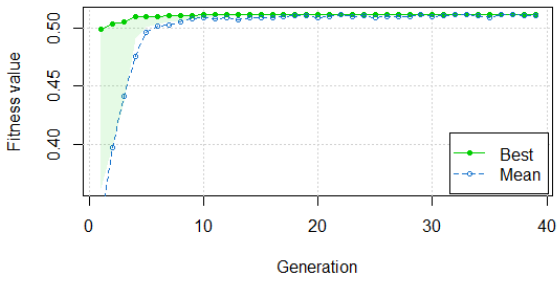
Outra questão relevante a ser ressaltada é que o SNP 6 não foi selecionado em nenhuma execução da etapa de corte do SMS nos cinco modelos de SVR analisados, por conseguinte, não foi selecionado pelo GA da etapa de refinamento (Tabela 8.4). Isso mostra que o *rank* da RF possui limitações para a detecção de SNPs homozigotos variantes com MAF pequena e efeito pequeno, o que não ocorreu com o método do valor-p bruto que ordenou o SNP 6 na posição 11 com valor-p igual a $2{,}92 \times 10^{-2}$, que é muito próximo ao ponto de corte 0,05, mas foi detectado. Por outro lado, o SNP 3 foi selecionado em todas as execuções da etapa de corte do SMS nos cinco modelos de SVR, mas foi selecionado apenas uma vez pela etapa de refinamento (GA) nas 10 execuções do SMS no *kernel* linear. Para os quatro *kernels* radiais analisados, o SNP 3 foi selecionado pela etapa de corte, mas praticamente não foi selecionado pelo GA nas 10 execuções. Portanto, o uso do *kernel* linear no SMS foi necessário para selecionar o SNP3.

A Figura 8.2 permite concluir que o GA do SMS convergiu rapidamente para o subconjunto de SNPs na $20^{\underline{a}}$ geração para todos os *kernels*, inclusive o com $\gamma = 1$ (gráfico não mostrado) e o melhor indivíduo não se alterou até a última geração. As correlações médias obtidas pelo melhor subconjunto de SNPs para os quatro *kernels* oscilaram entre 0,50 a 0,72.

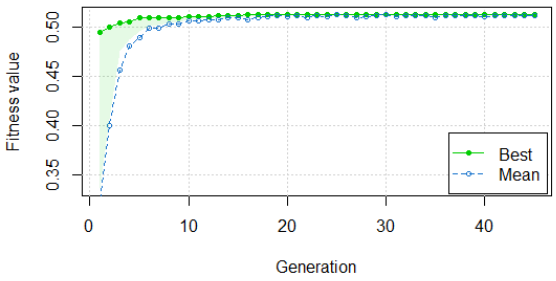
O método que apresentou o melhor resultado para esse conjunto de dados simulados

Tabela 8.4 Frequência da ausência dos oito SNPs causais da simulação 1 nas 10 execuções do SMS.

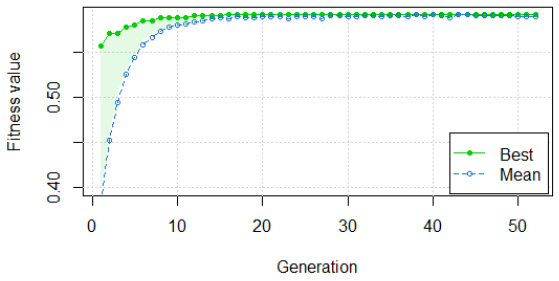
kernel	Etapas do SMS	SNPs causais							
		1	2	3	4	5	6	7	8
Linear	Relevância + Corte	0	0	0	0	0	**10**	0	0
	Refinamento	0	0	**1**	0	0	**10**	0	0
Radial $\gamma = 0,001$	Relevância + Corte	0	0	0	0	0	**10**	0	0
	Refinamento	0	0	**10**	0	0	**10**	0	0
Radial $\gamma = 0,01$	Relevância + Corte	0	0	0	0	0	**10**	0	0
	Refinamento	0	0	**9**	0	0	**10**	0	0
Radial $\gamma = 0,1$	Relevância + Corte	0	0	0	0	0	**10**	0	0
	Refinamento	0	0	**10**	0	0	**10**	0	0
Radial $\gamma = 1$	Relevância + Corte	0	0	0	0	0	**10**	0	0
	Radial	0	0	**10**	0	0	**10**	0	0



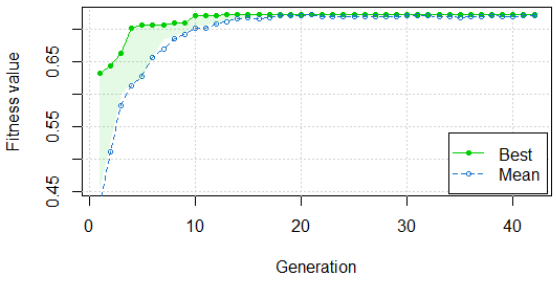
(a) *Kernel* linear.



(b) *Kernel* radial com $\gamma = 0,001$.



(c) *Kernel* radial com $\gamma = 0,01$.



(d) *Kernel* radial com $\gamma = 0,1$.

Figura 8.2 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 1.

foi o valor-p bruto, a partir do critério em relação ao maior número de SNPs causais selecionados, pois o mesmo selecionou oito SNPs verdadeiros-positivos e cinco falsos-positivos. Logo em seguida, o SMS ficou na segunda colocação com sete SNPs causais

e seis SNPs não-causais. O Blasso e o valor-p corrigido ficam nas terceira e quartas posições com seis e quatro SNPs verdadeiros-positivos respectivamente, apesar desses métodos selecionarem a menor quantidade de SNPs falsos-positivos em relação ao valor-p bruto e ao SMS.

8.4 Simulação 2 - Quatro interações de ordem 2 para regressão

A partir da Tabela 8.5 é possível inferir que nenhum método de seleção detectou os oito marcadores causais. Contudo, o SMS teve o melhor desempenho e identificou sete SNPs informativos e três não-informativos, sendo o SNP 4, o único marcador não encontrado. Todavia, entre os *kernels* usados, o que teve a maior eficiência foi o linear, pois identificou sete SNPs verdadeiros-positivos e somente um falso-positivo. Os métodos do valor-p bruto e valor-p corrigido identificaram seis e cinco SNPs causais respectivamente, mas detectaram, respectivamente, cinco e zero SNPs falsos-positivos. O Blasso encontrou oito marcadores, sendo seis causais e dois não-causais. Nenhum dos quatro métodos de seleção avaliados conseguiu detectar o SNP 4 e isso pode ter ocorrido pela forma como a interação entre os SNPs 3 e 4 foi gerada, onde na maioria das vezes que o SNP 3 for homozigoto variante (codificado como 3), o SNP 4 será diferente do homozigoto de referência (codificado como 1). Contudo, na etapa de filtro realizada pela RF, o SNP 4 foi ordenado em posições anteriores a 20ª (Tabelas 8.7 e 8.8), sendo o mesmo selecionado pela etapa de corte feita pelo SVR. Somente na etapa de refinamento executada pelo GA que esse SNP foi perdido e isso pode ter ocorrido pelos parâmetros avaliados do kernel ou pelo próprio algoritmo de busca do GA.

O subconjunto com a maior correlação média identificado pelo SMS com *kernel* radial com $\gamma = 0,1$, não foi o que identificou o subconjunto com maior número de SNPs informativos, como ocorreu nas simulações 1 e 2. Em relação aos*kernels* que selecionaram apenas SNPs informativos, apesar do *kernel* radial com $\gamma = 0,01$ selecionar mais SNPs causais do que o com $\gamma = 0,1$ e $\gamma = 1$, a correlação média do subconjunto selecionado pelo primeiro é 0,549, enquanto que as outras duas são respectivamente 0,811 e 0,806.

O desempenho dos oito SNPs causais estão evidenciados na Tabela 8.6. Note que o SVR com *kernel* radial com $\gamma = 0,1$ teve o melhor desempenho para os oito SNPs causais em todas as métricas avaliadas, e o linear o pior. Mas o subconjunto selecionado pelo SMS contém somente cinco dos oito SNPs informativos (quarta linha da Tabela 8.5) e esses dois subconjuntos apresentam correlações médias praticamente iguais. Isso significa que os SNPs 3, 4 e 6, que não pertencem ao subconjunto selecionado pelo SMS baseado no *kernel* radial com $\gamma = 0,1$, praticamente não aumentaram o poder preditivo desse *kernel*.

Tabela 8.5 Resultado da seleção dos SNPs para a simulação 2.

Método	γ	SNPs selecionados	Iter[4]	$\overline{r}$	# SNPs (V)[5]
SMS Linear[1]	-	**5**, **2**, **1**, **7**, **8**, **3**, **6**, 78	1	0,408	8 (7)
SMS Radial[1]	0,001	**5**, **2**, **1**, **7**, **8**, **3**, 46, 83	1	0,414	8 (6)
SMS Radial[1]	0,01	**5**, **2**, **1**, **7**, **8**, **3**	1	0,549	6 (6)
SMS Radial[1]	0,1	**5**, **2**, **1**, **7**, **8**	1	0,811	5 (5)
SMS Radial[1]	1	**5**, **2**, **1**, **7**, **8**	1	0,806	5 (5)
União do SMS[6]	-	**1**, **2**, **3**, **5**, **6**, **7**, **8**, 46, 78, 83	-	-	10 (7)
Inteseção do SMS[6]	-	**1**, **2**, **5**, **7**, **8**	-	-	5 (5)
Valor-p bruto[2]	-	**1**, **7**, **5**, **8**, **2**, **3**, 88, 10, 89, 60, 12	-	-	11 (6)
Valor-p corrigido[2]	-	**1**, **7**, **5**, **8**, **2**	-	-	5 (5)
Blasso[3]	-	**1**, **7**, **5**, **2**, **8**, **3**, 54, 88	-	-	8 (6)

[1] Corte pelo MSE do SVR sobre o *rank* da RF.

[2] Valor-p $< 0{,}05$.

[3] Variância explicada por cada marcador $> 0{,}01$.

[4] Iteração do SMS com a solução de maior correlação média.

[5] Número de verdadeiros-positivos.

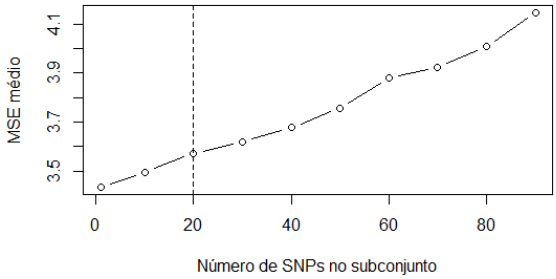
[6] O *rank* não é considerado no subconjunto de SNPs selecionados.

Tabela 8.6 Desempenho dos oito SNPs causais nos cinco *kernels* avaliados para a simulação 2.

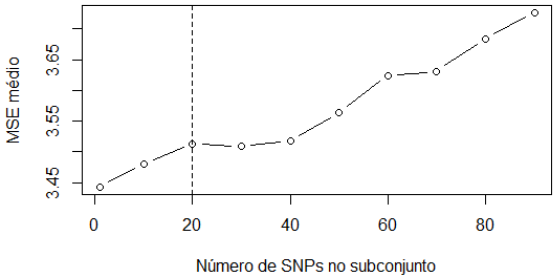
kernel	γ	$\overline{r}$	σ_r	$\overline{mse}$	σ_{mse}	$\overline{mape}$	σ_{mape}
Linear	-	0,40	0,07	3,40	0,30	0,23	0,011
Radial	1	0,65	0,08	2,27	0,29	0,18	0,009
Radial	**0,1**	**0,80**	**0,04**	**1,43**	**0,24**	**0,15**	**0,011**
Radial	0,01	0,54	0,06	2,84	0,27	0,21	0,011
Radial	0,001	0,41	0,07	3,43	0,40	0,23	0,014

A partir da Figura 8.3 nota-se que o ponto de corte foi o 20, isto é, a primeira seleção do SMS, baseada no MSE médio do SVR, escolhe os 20 SNPs mais importantes pelo *rank* construído pela RF. Novamente, é mostrada a importância de flexibilizar a entrada de mais SNPs informativos além do ponto de mínimo (SNP 5) no gráfico do MSE médio do SVR para permitir que o SMS selecione o maior números de SNPs verdadeiros-positivos. Note que no gráfico (a) da Figura 8.3 as variações do MSE médio absoluta e relativa estão entre os limites 3,5 e 4,1 para o *kernel* radial com $\gamma = 0,1$, respectivamente, próximas de 0,60 e 17%.

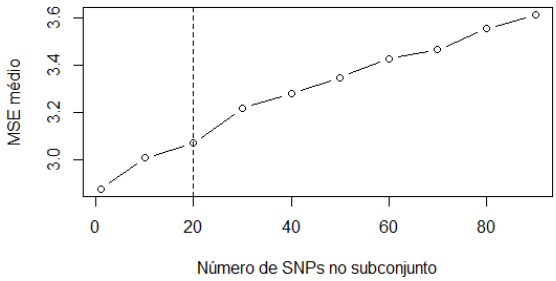
A ordenação produzida pela RF mostrou ser eficiente, pois todos os oito SNPs causais foram colocados em posições anteriores ao corte realizado pelo SVR como pode ser notado na Tabela 8.7. Para os métodos baseados no valor-p, os SNPs 4 e 6 não seriam identificados



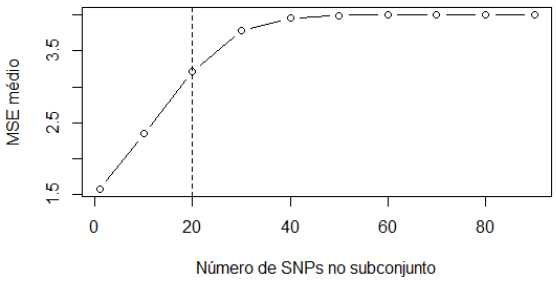
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.3 Corte do SVR sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 2.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação a simulação 2. A linha tracejada indica o ponto de corte.

mesmo se o limite superior fosse flexibilizado. Para o método Blasso, o SNP 4 ficou na 99ª posição (penúltimo lugar), e o SNP 6 na 22ª posição. Essas observações permitem concluir que o SMS foi o método com melhor desempenho nesse cenário com interações de ordem 2. É importante ressaltar que o SMS pode ser melhorado tanto no SVR quanto no GA para a possível identificação do SNP 4.

Pela Tabela 8.8, percebe-se que todos os *kernels* analisados não detectaram o SNP 4 na fase de refinamento do GA, entretanto, esse SNP foi selecionado pelo corte do SVR em todas as 50 execuções do SMS. Em relação ao SNP 6, o mesmo só foi selecionado pelo *kernel* linear em nove execuções do SMS, ocorrendo somente uma ausência na 4ª execução.

A Figura 8.4 mostra que o GA convergiu para a solução ótima em 20 gerações aproximadamente. Esse comportamento é análogo ao SMS executado para a simulação 1.

O método que demonstrou o melhor resultado para esse conjunto de dados simulados

Tabela 8.7 Ordenação de cada método para os oito SNPs causais para a simulação 2.

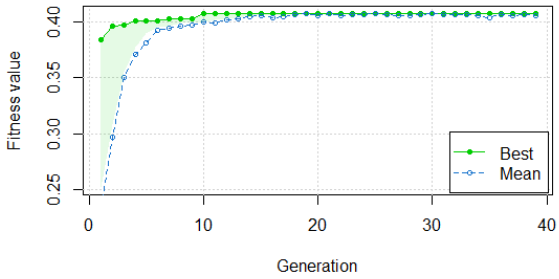
Método	Iter[a]	Corte	SNPs causais							
			1	2	3	4	5	6	7	8
RF do SMS Linear	1	20	$3^{\underline{a}}$	$2^{\underline{a}}$	$7^{\underline{a}}$	$12^{\underline{a}}$	$1^{\underline{a}}$	$9^{\underline{a}}$	$4^{\underline{a}}$	$5^{\underline{a}}$
RF do SMS Radial $\gamma = 0,001$	1	20	$3^{\underline{a}}$	$2^{\underline{a}}$	$7^{\underline{a}}$	$18^{\underline{a}}$	$1^{\underline{a}}$	$9^{\underline{a}}$	$4^{\underline{a}}$	$5^{\underline{a}}$
RF do SMS Radial $\gamma = 0,01$	1	20	$3^{\underline{a}}$	$2^{\underline{a}}$	$7^{\underline{a}}$	$12^{\underline{a}}$	$1^{\underline{a}}$	$9^{\underline{a}}$	$4^{\underline{a}}$	$5^{\underline{a}}$
RF do SMS Radial $\gamma = 0,1$	1	20	$3^{\underline{a}}$	$2^{\underline{a}}$	$7^{\underline{a}}$	$11^{\underline{a}}$	$1^{\underline{a}}$	$9^{\underline{a}}$	$4^{\underline{a}}$	$5^{\underline{a}}$
RF do SMS Radial $\gamma = 1$	1	20	$3^{\underline{a}}$	$2^{\underline{a}}$	$7^{\underline{a}}$	$17^{\underline{a}}$	$1^{\underline{a}}$	$9^{\underline{a}}$	$4^{\underline{a}}$	$5^{\underline{a}}$
Valor-p bruto	-	11	$1^{\underline{a}}$	$5^{\underline{a}}$	$6^{\underline{a}}$	$\mathbf{85^{\underline{a}}}$	$3^{\underline{a}}$	$\mathbf{24^{\underline{a}}}$	$2^{\underline{a}}$	$4^{\underline{a}}$
Valor-p ajustado	-	5	$1^{\underline{a}}$	$5^{\underline{a}}$	$\mathbf{6^{\underline{a}}}$	$\mathbf{85^{\underline{a}}}$	$3^{\underline{a}}$	$\mathbf{24^{\underline{a}}}$	$2^{\underline{a}}$	$4^{\underline{a}}$
Blasso	-	8	$1^{\underline{a}}$	$4^{\underline{a}}$	$6^{\underline{a}}$	$\mathbf{99^{\underline{a}}}$	$3^{\underline{a}}$	$\mathbf{22^{\underline{a}}}$	$2^{\underline{a}}$	$5^{\underline{a}}$

[a] Número da iteração nas 10 execuções do SMS.

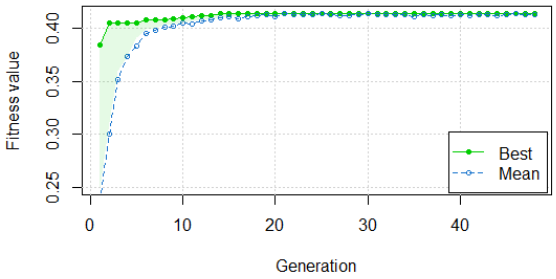
Tabela 8.8 Frequência da ausência dos oito SNPs causais da simulação 2 nas 10 execuções do SMS.

kernel	Etapas do SMS	SNPs causais							
		1	2	3	4	5	6	7	8
Linear	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	**10**	0	**1**	0	0
Radial $\gamma = 0,001$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	**10**	0	**10**	0	0
Radial $\gamma = 0,01$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	**10**	0	**10**	0	0
Radial $\gamma = 0,1$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	**10**	**10**	0	**10**	0	0
Radial $\gamma = 1$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	**10**	**10**	0	**10**	0	0

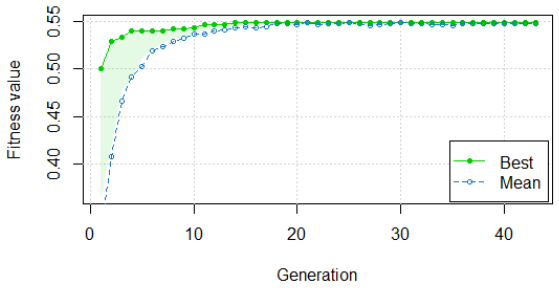
foi o SMS, a partir do critério em relação ao maior número de SNPs causais selecionados, pois o mesmo selecionou sete SNPs verdadeiros-positivos e três falsos-positivos. Logo em seguida, o valor-p bruto ficou na segunda colocação com seis SNPs causais e cinco SNPs não-causais. O Blasso e o valor-p corrigido ficam nas terceira e quartas posições com seis e cinco SNPs verdadeiros-positivos respectivamente, apesar desses métodos selecionarem a menor quantidade de SNPs falsos-positivos em relação ao valor-p bruto e ao SMS.



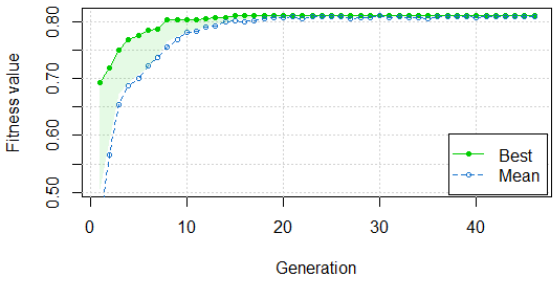
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.4 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 2.

8.5 Simulação 3 - Três interações de ordem 3 para regressão

A Tabela 8.9 mostra que nenhum dos métodos usados capturou os nove marcadores informativos, mas somente a solução dada pelo conjunto união do SMS teve o melhor desempenho selecionando 6 marcadores causais, onde foram identificadas as interações de ordem 3 do primeiro trio (SNPs 1, 2 e 3) e do terceiro trio (SNPs 7, 8 e 9). Em contrapartida, nenhum dos outros métodos selecionou mais de 4 marcadores informativos, além de nenhum subconjunto de marcadores possuir pelo menos um dos 3 trios de SNPs que interagem entre si.

Dentre os modelos SVR avaliados pelo SMS, o *kernel* radial com $gamma = 0,1$ foi o que apresentou maior correlação média e também selecionou o maior número de marcadores causais, além de selecionar apenas 2 falsos-positicos. Entretanto, a interação do trio de SNPs 9, 7 e 8 só foi capturada pelo *kernel* radial com $gamma = 1$, onde o mesmo não detectou SNP não-informativo. Uma possível justificativa para isso é que o γ mais adequado para selecionar a maior parte dos SNPs informativos, que interagem em trios, pode estar entre os valores 0,1 e 1.

A Tabela 8.10 mostra um resultado diferente em relação às simulações 1, 2 e 3, pois para os quatro *kernels* radiais, as três medidas preditivas apresentaram o mesmo valor tanto para média quanto para o desvio-padrão. O *kernel* linear demonstrou resultado ligeiramente superior para as três medidas avaliadas tanto em relação à média quanto ao desvio-padrão.

A partir da Figura 8.5 nota-se que o ponto de corte assumiu os valores 20 ou 30, o que não difere das simulações anteriores. Todavia, as curvas geradas pelo MSE médio do SVR demonstraram um comportamento distinto com a variação do γ, pois o ponto de mínimo nas *kernels* linear e radial com $\gamma = 0,001$ não ocorreu no SNP colocado na primeira posição pela RF. Outro ponto a ser destacado no gráfico (a) da Figura 8.7 são as variações do MSE médio absoluta e relativa entre os limites 1,45 e 1,65 para o *kernel* linear, respectivamente, próximas de 0,20 e 14%.

A Tabela 8.11 permite inferir que os SNPs 4, 5 e 6 não foram detectados pelo SVR na primeira seleção do SMS, pois a RF os ordenou em posições posteriores ao ponto de corte. Todavia, os outros seis SNPs foram ordenados nas primeiras posições, o que

Tabela 8.9 Resultado da seleção dos SNPs para a simulação 3.

Método	γ	SNPs selecionados	Iter[4]	$\overline{r}$	# SNPs (V)[5]
SMS Linear[1]	-	**9**, **7**, 68, 11, 60, 49, 10, 40, 30	1	0,198	9 (2)
SMS Radial[1]	0,001	**9**, **7**, **3**, 68, 49, 60, 11, 10, 40, 44, 63	2	0,198	12 (3)
SMS Radial[1]	0,01	**9**, **7**, **3**, 49, 11, 10, 79, 22	1	0,281	8 (3)
SMS Radial[1]	0,1	**9**, **7**, **3**, **2**, **1**, 45, 20	1	0,397	7 (5)
SMS Radial[1]	1	**9**, **7**, **8**	7	0,373	3 (3)
União do SMS[6]	-	**1**, **2**, **3**, **7**, **8**, **9**, 10, 11, 20, 22, 30, 40, 44, 45, 49, 60, 63, 68, 79	-	-	19 (6)
Inteseção do SMS[6]	-	**9**, **7**	-	-	2 (2)
Valor-p bruto[2]	-	**9**, 10, **7**, **3**, 60, **6**, 82, 40	-	-	8 (4)
Valor-p corrigido[2]	-	**9**	-	-	1 (1)
Blasso[3]	-	**9**, **7**, 10, 68, 11, 82, **3**, 69, 87, 60, 88, 40, 53, 13, 36, 44, 67, 83, 20, 34, **5**	-	-	21 (4)

[1] Corte pelo MSE do SVR sobre o *rank* da RF.

[2] Valor-p $< 0{,}05$.

[3] Variância explicada por cada marcador $> 0{,}01$.

[4] Iteração do SMS com a solução de maior correlação média.

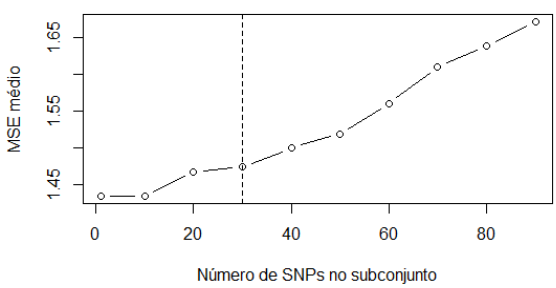
[5] Número de verdadeiros-positivos.

[6] O *rank* não é considerado no subconjunto de SNPs selecionados.

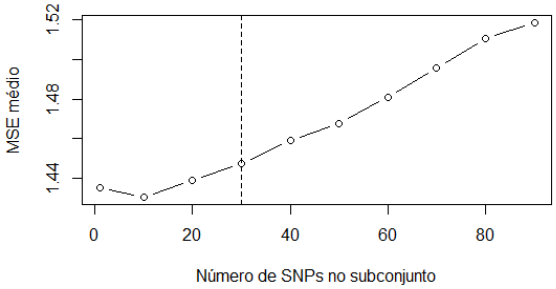
Tabela 8.10 Desempenho dos nove SNPs causais nos cinco *kernels* avaliados para a simulação 3.

kernel	γ	r (σ_r)	MSE (σ_{mse})	MAPE (σ_{mape})
Linear	-	0,118 (0,09)	1,441 (0,25)	0,144 (0,01)
Radial	0,001	0,091 (0,11)	1,545 (0,29)	0,149 (0,01)
Radial	0,01	0,091 (0,11)	1,545 (0,29)	0,149 (0,01)
Radial	0,1	0,091 (0,11)	1,545 (0,29)	0,149 (0,01)
Radial	1	0,091 (0,11)	1,545 (0,29)	0,149 (0,01)

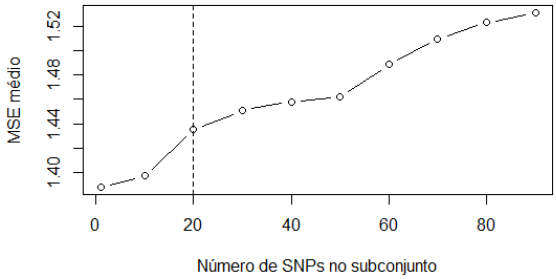
permitiu sua seleção pelo GA. Os SNPs 1, 2 e 8, não selecionados pelos dois métodos do valores-p bruto e corrigido, mostraram diferenças significativas em suas ordenações em relação à da RF, entretanto, os *ranks* gerados pelo valor-p e pela RF para os SNPs 4 e 5 foram equivalentes. A ordenação do Blasso apresentou comportamento próximo aos dos valores-p bruto e corrigido para os SNPs 1, 2 e 8; e à RF para os SNPs 4, 5 e 6. Essa constatação mostra como existem vieses distintos para as quatro abordagens de seleção de marcadores utilizadas em cenários de epistasia com trios de SNPs, além de indicar que a RF consegue capturar duas das três interações construídas.



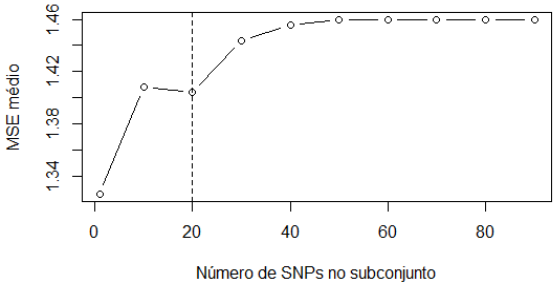
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.5 Corte do SVR sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 3.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação à simulação 3. A linha tracejada indica o ponto de corte.

Tabela 8.11 *Rank* gerado por cada método para os nove SNPs causais para a simulação 3.

Rank	Iter	Corte	SNPs causais								
			1	2	3	4	5	6	7	8	9
RF do SMS Linear	1	20	6ª	5ª	4ª	**31ª**	**44ª**	**51ª**	2ª	3ª	1ª
RF do SMS Radial $\gamma = 0,001$	1	30	5ª	6ª	4ª	**51ª**	**58ª**	**55ª**	2ª	3ª	1ª
RF do SMS Radial $\gamma = 0,01$	2	20	5ª	6ª	4ª	**40ª**	**39ª**	**31ª**	2ª	3ª	1ª
RF do SMS Radial $\gamma = 0,1$	1	20	5ª	6ª	4ª	**24ª**	**66ª**	**44ª**	2ª	3ª	1ª
RF do SMS Radial $\gamma = 1$	1	20	5ª	6ª	4ª	**35ª**	**61ª**	**39ª**	2ª	3ª	1ª
Valor-p bruto	-	13	**35ª**	**72ª**	4ª	**44ª**	**51ª**	6ª	3ª	**50ª**	1ª
Valor-p ajustado	-	4	**35ª**	**72ª**	4ª	**44ª**	**51ª**	6ª	3ª	**50ª**	1ª
Blasso	-	7	**77ª**	**25ª**	7ª	**32ª**	21ª	**79ª**	2ª	**44ª**	1ª

De acordo com a Tabela 8.12, a fase de refinamento feita pelo GA, baseado nos *kernels* linear e radial com $\gamma = 0,001$ e com $\gamma = 0,01$, não capturou os SNPs 1, 2, 3 e 8, os quais foram identificados pela fase de corte. Logo, ou o kernel usado não conseguiu perceber o sinal gerado pelos trios de SNPs causais que foram pelo GA, ou durante o processo evolutivo do GA, as 3 ternas de SNPs não foram construídas em um único

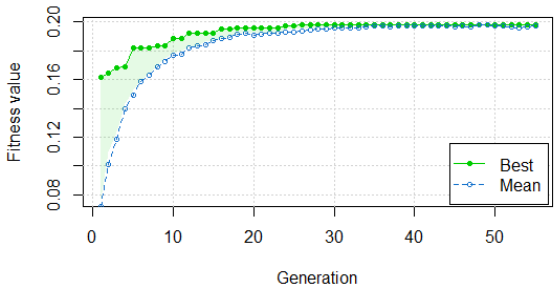
subconjunto para ser avaliado adequadamente pela função de aptidão. Para o kernel radial com $\gamma = 0,1$, houve uma variação significativa no corte, pois os SNPs 4 e 5 foram selecionados nessa fase nas 10 execuções do SMS, mas não foi pelos outros *kernels* na mesma etapa. Entretanto, o GA não gerou, ou não identificou subconjuntos com essa dupla de SNPs, ou, até mesmo, avaliou esse subconjunto, mas o mesmo demonstrou desempenho inferior a algum outro sem esses marcadores, não sendo selecionados ao final do GA. Em relação ao *kernel* radial com $\gamma = 1$, o terno de SNPs 1, 2 e 3 foi identificado nas 10 execuções do SMS para a etapa de corte, mas somente algumas vezes na fase de refinamento, e, como escolhe-se o subconjunto final com maior correlação média, que, neste caso, não possui esses SNPs, os mesmos não foram selecionados. A partir dessa análise, parece que algum γ distinto e próximo de 1, pode selecionar todos os nove SNPs mesmo com alguns SNPs falso-positivos, porém, isso tem que ser verificado em um trabalho posterior.

Tabela 8.12 Frequência da ausência dos nove SNPs causais da simulação 3 nas 10 execuções do SMS.

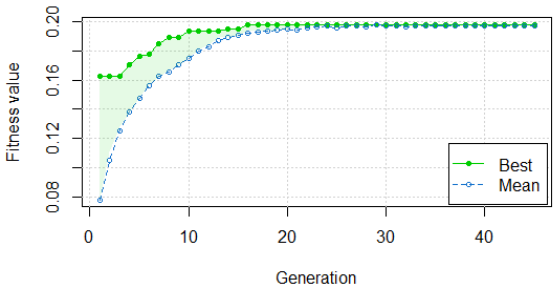
kernel	Etapas do SMS	SNPs causais								
		1	2	3	4	5	6	7	8	9
Linear	Relevância + Corte	0	0	0	**10**	**10**	**10**	0	0	0
	Refinamento	**10**	**10**	**10**	**10**	**10**	**10**	0	**10**	0
Radial $\gamma = 0,001$	Relevância + Corte	0	0	0	**10**	**10**	**10**	0	0	0
	Refinamento	**10**	**10**	0	**10**	**10**	**10**	0	**10**	0
Radial $\gamma = 0,01$	Relevância + Corte	0	0	0	**10**	**10**	**10**	0	0	0
	Refinamento	**10**	**10**	0	**10**	**10**	**10**	0	**10**	0
Radial $\gamma = 0,1$	Relevância + Corte	0	0	0	0	0	**10**	**10**	**10**	0
	Refinamento	**3**	**3**	0	**10**	**10**	**10**	**10**	**10**	0
Radial $\gamma = 1$	Relevância + Corte	0	0	0	**6**	**9**	**8**	0	0	0
	Refinamento	**6**	**5**	**7**	**8**	**10**	**10**	**6**	**4**	**2**

A Figura 8.6 mostra que o GA convergiu para a solução ótima em 30 gerações aproximadamente. Note também que as correlações médias geradas pelos subconjuntos finais de SNPs do GA com os *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$ são, respectivamente, iguais a 0,30 e 0,35, as quais foram superiores às correlações médias dos outros *kernels*, próximas ao valor de 0,20.

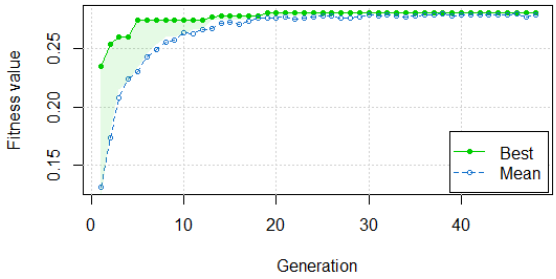
O método que obteve o melhor desempenho para esse conjunto de dados simulados foi o SMS, a partir do critério em relação ao maior número de SNPs causais selecionados, pois o mesmo selecionou seis SNPs verdadeiros-positivos e treze falsos-positivos. Logo em



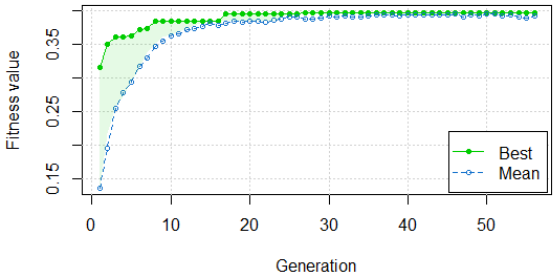
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.6 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 3.

seguida, o valor-p bruto ficou na segunda colocação com quatro SNPs causais e quatro SNPs não-causais. O Blasso e o valor-p corrigido ficam nas terceira e quartas posições com quatro e um SNPs verdadeiros-positivos respectivamente. Apesar do valor-p bruto e do Blasso empatarem no número de SNPs causais selecionados, a quantidade de SNPs não-causais selecionada pelo valor-p bruto (quatro SNPs não-causais) foi inferior à quantidade selecionada pelo Blasso (17 SNPs não-causais), portanto, o valor-p bruto demonstrou melhor resultado. Alem disso, esses dois métodos selecionaram SNPs distintos, pois o valor-p bruto selecionou o SNP6, mas não selecionou o SNP5 e o Blasso selecionou o SNP5, mas não o SNP6. O SMS não selecionou nem o SNP5 e nem o SNP6, o que mostra a complexidade dessa simulação e a possibilidade de utilizar em trabalhos futuros a união de métodos distintos para aumentar a chance de capturar o maior número possível de SNPs informativos.

8.6 Simulação 4 - Três efeitos aditivos + uma interação de ordem 2 + uma interação de ordem 3 para regressão

A escolha pelo melhor modelo SMS foi o referente ao $\gamma = 0,1$ que apresentou maior correlação média igual a 0,75. Entretanto, esse modelo selecionou 7 marcadores relevantes, mas perdeu o SNP 3. Com o $\gamma = 1$, o SMS eliminou os SNPs 3 e 6, porém não introduziu SNP falso-positivo. O modelo SMS com $\gamma = 0,01$ capturou todos os oito SNPs informativos, mas inseriu quatro marcadores não-informativos. Uma possível explicação para esse comportamento é que deve-se buscar outros γs intermediários entre os limites 0,01 e 0,1 para possibilitar ao SMS encontrar todos os marcadores causais e reduzir o número de marcadores espúrios. Outra análise possível é que o ruído gerado pelos SNPs não-informativos aumenta a correlação média do subconjunto final selecionado em relação aos oito SNPs

O único método que selecionou todos os marcadores foi o SMS, apesar do mesmo inserir nove marcadores falso-positivos. De um lado, o método do valor-p bruto não selecionou os SNPs 6 e 7 e inseriu sete marcadores falso-positivos. Por outro lado, o valor-p ajustado não selecionou os SNPs 3, 4, 6 e 7, mas não introduziu marcador falso-positivo, o que demonstra ser um método extremamente restritivo. O Blasso adicionou apenas o marcador 4 em relação à seleção do valor-p ajustado.

Dentre os cinco *kernels* avaliados, o que demonstrou melhor resultado na seleção de SNPs foi o radial com $\gamma = 0,01$, que capturou os oito SNPs causais e inseriu somente quatro falsos-positivos como pode ser notado na Tabela 8.13. Entretanto, o subconjunto de SNPs selecionado por esse *kernel* apresentou a terceira maior correlação média, sendo a maior referente à seleção do *kernel* radial com $\gamma = 0,1$, o qual não detectou o SNP 3. Assim, usar a maior correlação como critério de escolha do subconjunto de SNPs não é adequado para garantir a seleção de todos os SNPs causais. Ou seja, a melhor seleção é dada pelo *kernel* que melhor se adapta à relação matemática entre o genótipo e o fenótipo, e não pelo *kernel* que apresenta a maior correlação dos fenótipos observados com os preditos.

O desempenho dos oito SNPs causais são evidenciados na Tabela 8.14. O comportamento dos cinco *kernels* para os oito SNPs causais foi similar ao do modelo

Tabela 8.13 Resultado da seleção dos SNPs para a simulação 4.

Método	γ	SNPs selecionados	Iter[4]	$\overline{r}$	# SNPs (V)[5]
SMS Linear[1]	-	**1**, **2**, **5**, **4**, **8**, **7**, **3**, 27, 71, 91, 62, 10	10	0,589	12 (7)
SMS Radial[1]	0,001	**1**, **2**, **5**, **4**, **8**, **7**, **3**, 77, 71, 10	8	0,588	9 (7)
SMS Radial[1]	0,01	**1**, **2**, **5**, **4**, **8**, **7**, **3**, 54, **6** , 63, 83, 10	1	0,651	12 (8)
SMS Radial[1]	0,1	**1**, **2**, **5**, **4**, **8**, **7**, **6**, 9	1	0,753	8 (7)
SMS Radial[1]	1	**1**, **2**, **5**, **8**, **7**	1	0,708	4 (4)
União do SMS[6]	-	**1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, 9, 10, 27, 54, 62, 63, 71, 83, 91	-	-	17 (8)
Inteseção do SMS[6]	-	**1**, **2**, **4**, **5**, **8**	-	-	5 (5)
Valor-p bruto[2]	-	**1**, **2**, **8**, **5**, **3**, 89, **4**, 88, 10, 11, 60, 83, 77	-	-	13 (6)
Valor-p corrigido[2]	-	**1**, **2**, **8**, **5**	-	-	4 (4)
Blasso[3]	-	**1**, **2**, **8**, **5**, **4**	-	-	5 (5)

[1] Corte pelo MSE do SVR sobre o *rank* da RF.

[2] Valor-p $< 0{,}05$.

[3] Variância explicada por cada marcador $> 0{,}01$.

[4] Iteração do SMS com a solução de maior correlação média.

[5] Número de verdadeiros-positivos.

[6] O *rank* não é considerado no subconjunto de SNPs selecionados.

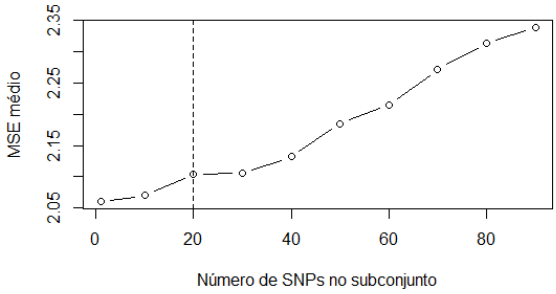
1, onde o radial com $\gamma = 0,1$ demonstrou a maior correlação, e o radial com $\gamma = 1$ a menor correlação. Como a correlação média dos oito SNPs informativos para o *kernel* radial com $\gamma = 0,1$ foi 0,74 (Tabela 8.14), a qual é menor que 0,75, que foi referente ao subconjunto selecionado pelo mesmo *kernel*, não seria possível que o SMS encontrasse exatamente os oitos SNPs causais, a menos que o GA ficasse preso em um máximo local.

Tabela 8.14 Desempenho dos oito SNPs causais nos cinco *kernels* avaliados para a simulação 4.

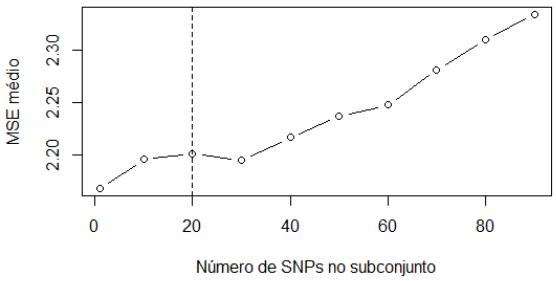
kernel	γ	$\overline{r}$	σ_r	$\overline{mse}$	σ_{mse}	$\overline{mape}$	σ_{mape}
Linear	-	0,59	0,05	2,05	0,33	0,17	0,012
Radial	0,001	0,58	0,05	2,16	0,35	0,18	0,012
Radial	0,01	0,64	0,05	1,37	0,20	0,15	0,012
Radial	**0,1**	**0,74**	**0,04**	**1,37**	**0,20**	**0,14**	**0,012**
Radial	1	0,55	0,09	2,15	0,46	0,18	0,015

Para a etapa de corte realizada pela RF, percebe-se uma estabilidade nos quatro *kernels* apresentados na Figura 8.7. As curvas do MSE médio do SVR sobre os SNPs ordenados pela RF demonstraram comportamento não-decrescente ao longo dos subconjuntos avaliados. Isso mostra que existe similaridade entre as ordenações da RF e

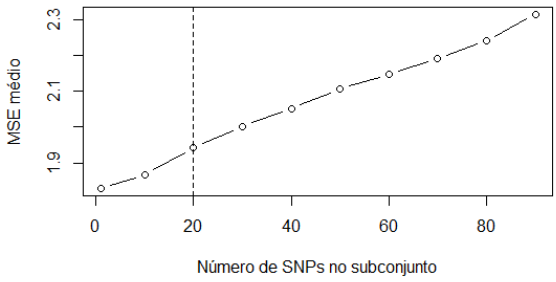
do SVR, pois o mínimo global do MSE médio foi o primeiro ponto. Caso contrário, haveria outro ponto de mínimo global para o MSE médio, o qual ocorreria em alguma posição além da primeira. É importante destacar no gráfico (a) da Figura 8.7 que as variações do MSE médio absoluta e relativa estão entre os limites 2,05 e 2,35 para o *kernel* linear, respectivamente, próximas de 0,20 e 14%.



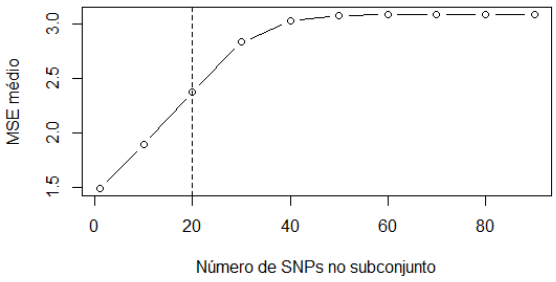
(a) *Kernel* linear na iteração 10.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 8.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.7 Corte do SVR sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 4.

textitKernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação ao modelo 4. A linha tracejada indica o ponto de corte.

O *rank* da RF para todos os *kernels* usados no SMS demonstrou estabilidade, pois dos oitos SNPs causais, somente o SNP 6 oscilou entre as posições 10 e 11, e os demais foram ordenados nas mesmas posições nas diferentes execuções do SMS como mostrado na Tabela 8.15. Entretanto, os SNPs 6 e 7 foram classificados, respectivamente, nas posições 83 e 42 (posições em negrito nas linhas 6 e 7 da Tabela 8.15) para os métodos baseados nos valores-p, não sendo selecionados para o limite de corte de 0,05. Além desses dois SNPs, os SNPs 3 e 4 não foram selecionados pelo valor-p corrigido. O Blasso classificou os SNPs 3, 6 e 7, respectivamente, nas posições 6, 100 e 12 (posições em negrito na última

linha da Tabela 8.15), ou seja, não selecionou esses marcadores.

Tabela 8.15 Ordenação de cada método para os oito SNPs causais para a simulação 4.

Método	Iter[a]	Corte	SNPs causais							
			1	2	3	4	5	6	7	8
RF do SMS Linear	10	20	1^{a}	2^{a}	7^{a}	4^{a}	3^{a}	10^{a}	6^{a}	5^{a}
RF do SMS Radial $\gamma = 0,001$	8	20	1^{a}	2^{a}	7^{a}	4^{a}	3^{a}	11^{a}	6^{a}	5^{a}
RF do SMS Radial $\gamma = 0,01$	1	20	1^{a}	2^{a}	7^{a}	4^{a}	3^{a}	10^{a}	6^{a}	5^{a}
RF do SMS Radial $\gamma = 0,1$	1	20	1^{a}	2^{a}	7^{a}	4^{a}	3^{a}	11^{a}	6^{a}	5^{a}
RF do SMS Radial $\gamma = 1$	1	20	1^{a}	2^{a}	7^{a}	4^{a}	3^{a}	10^{a}	6^{a}	5^{a}
Valor-p bruto	-	13	1^{a}	2^{a}	5^{a}	7^{a}	4^{a}	$\mathbf{83^{a}}$	$\mathbf{42^{a}}$	3^{a}
Valor-p ajustado	-	4	1^{a}	2^{a}	$\mathbf{5^{a}}$	$\mathbf{7^{a}}$	4^{a}	$\mathbf{83^{a}}$	$\mathbf{42^{a}}$	3^{a}
Blasso	-	5	1^{a}	2^{a}	$\mathbf{6^{a}}$	5^{a}	4^{a}	$\mathbf{100^{a}}$	$\mathbf{12^{a}}$	3^{a}

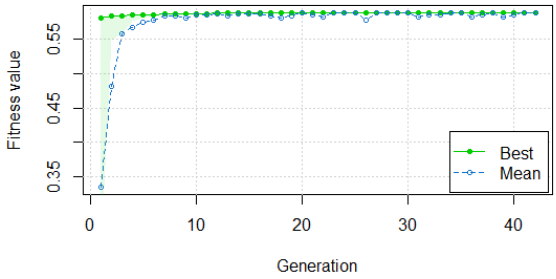
[a] Número da iteração nas 10 execuções do SMS.

A partir da Tabela 8.16 é possível inferir que todos os oito SNPs informativos foram selecionados na etapa de corte, entretanto somente o *kernel* radial com $\gamma = 0,01$ conseguiu selecionar todos SNPs causais na etapa de refinamento. Consequentemente, para que a seleção final capture o maior número possível de SNPs, deve ocorrer a compatibilidade entre a seleção da RF com a do GA cuja aptidão é baseada na correlação média gerada entre os valores observados e os preditos pelo SVR.

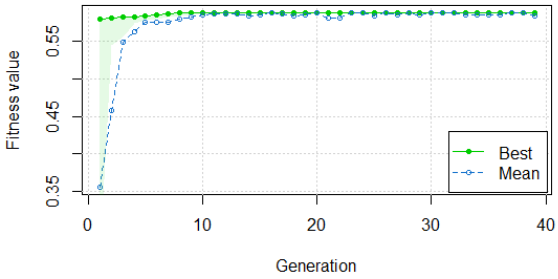
Tabela 8.16 Frequência da ausência dos oito SNPs causais da simulação 4 nas 10 execuções do SMS.

kernel	Etapas do SMS	SNPs causais							
		1	2	3	4	5	6	7	8
Linear	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	0	0	**10**	**3**	0
Radial $\gamma = 0,001$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	0	0	**10**	**3**	0
Radial $\gamma = 0,01$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	0	0	0	**2**	0	0
Radial $\gamma = 0,1$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	**10**	0	0	**3**	0	0
Radial $\gamma = 1$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	0	**10**	**8**	0	**10**	**2**	**2**

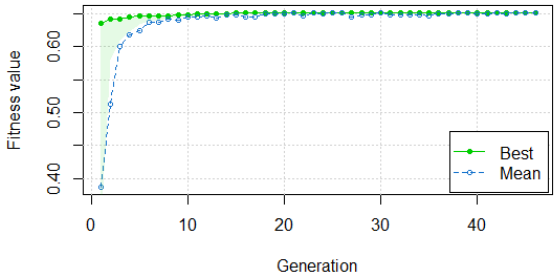
A Figura 8.8 permite concluir que o GA do SMS convergiu rapidamente para o subconjunto de SNPs na 20^{a} geração para todos os *kernels*, inclusive o referente ao $\gamma = 1$ (gráfico não mostrado) e o melhor indivíduo não se alterou até a última geração. As correlações médias obtidas pelo melhor subconjunto de SNPs para os quatro *kernels* oscilaram entre 0,55 a 0,75.



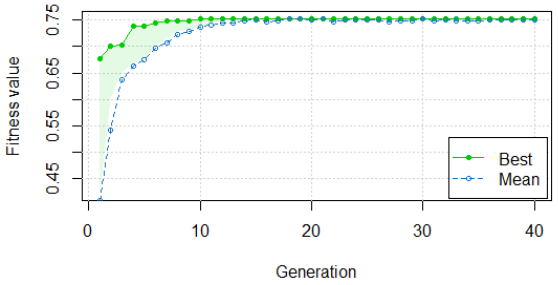
(a) *Kernel* linear.



(b) *Kernel* radial com $\gamma = 0,001$.



(c) *Kernel* radial com $\gamma = 0,01$.



(d) *Kernel* radial com $\gamma = 0,1$.

Figura 8.8 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 4.

Uma conclusão a partir dessa simulação e da simulação 3 é que os trios formados pelos SNPs 1, 2 e 3; configurados como respectivamente, homozigoto de referência, heterozigoto e homozigoto variante; foram selecionados pelo SMS. Esse fato pode ter acontecido pelo coeficiente 3, designado para essa terna, ter sido suficiente para a captação do sinal pelo SMS.

O método que mostrou o melhor desempenho para esse conjunto de dados simulados foi o SMS, a partir do critério em relação ao maior número de SNPs causais selecionados, pois o mesmo selecionou todos os oito SNPs verdadeiros-positivos e nove falsos-positivos. Logo em seguida, o valor-p bruto ficou na segunda colocação com seis SNPs causais e cinco SNPs não-causais. O Blasso e o valor-p corrigido ficam nas terceira e quartas posições com cinco e quatro SNPs verdadeiros-positivos respectivamente. O Blasso e o valor-p corrigido não selecionaram SNPs falsos-positivos, mas os mesmos, juntamente com o método do valor-p bruto, não capturaram a interação de ordem 3 formada pelo trio de SNPs 7, 8 e 9.

8.7 Simulação 5 - Somente uma interação de ordem 4 para regressão

A Tabela 8.17 mostra que nenhum dos métodos usados capturou os quatro marcadores informativos, mas somente os subconjuntos dados pela união do SMS e do Blasso tiveram melhor desempenho selecionando dois marcadores causais, onde os SNPs 1 e 3 foram detectados pelo SMS, e os SNPs 2 e 4 pelo Blasso.

Tabela 8.17 Resultado da seleção dos SNPs para a simulação 5.

Método	γ	SNPs selecionados	Iter[4]	$\overline{r}$	# SNPs (V)[5]
SMS Linear[1]	-	11, 68, 10, 53, 40, 60, 83, 7, 92, 95, 77, 19	4	0,152	12 (0)
SMS Radial[1]	0,001	11, 10, 68, 92, 60, 40, 95, 90, 83, 53, 71	6	0,153	11 (0)
SMS Radial[1]	0,01	**3**, 11, 10, 68, **1**, 92, 83, 40, 60, 22, 30, 76, 95, 29, 84, 53	4	0,182	16 (2)
SMS Radial[1]	0,1	68, 10, 27, 60, 76, 22, 20, 92, 30, 23, 29, 84, 95, 18, 77, 44, 9, 7, 24	4	0,285	19 (0)
SMS Radial[1]	1	27, **1**, 83, 30, 22, 75, 63, 91, 20, 70, 19, 98, 53, 86, 88	6	0,263	13 (1)
União do SMS[6]	-	**1**, **3**, 7, 9, 10, 11, 18, 19, 20, 22, 23, 24, 27, 29, 30, 40, 44, 53, 60, 63, 68, 70, 71, 75, 76, 77, 83, 84, 86, 88, 90, 91, 92, 95, 98	-	-	35 (2)
Inteseção do SMS[6]	-	-	-	-	-
Valor-p bruto[2]	-	10, 60, 53, 11, 40	-	-	5 (0)
Valor-p corrigido[2]	-	10	-	-	1 (0)
Blasso[3]	-	10, 11, 53, 68, 88, 87, 82, 13, 31, 69, 67, 60, 20, 36, 71, 83, 5, **2**, 84, 50, 59, 21, 44, **4**, 40, 37, 89, 97, 34, 92, 16, 48, 27, 32, 24, 65, 58, 7, 12, 33, 8, 78, 26, 61, 25, 22, 9, 28, 72, 42, 94, 73, 52	-	-	53 (2)

[1] Corte pelo MSE do SVR sobre o *rank* da RF.

[2] Valor-p $< 0{,}05$.

[3] Variância explicada por cada marcador $> 0{,}01$.

[4] Iteração do SMS com a solução de maior correlação média.

[5] Número de verdadeiros-positivos.

[6] O *rank* não é considerado no subconjunto de SNPs selecionados.

É importante destacar que o *kernel* radial com $\gamma = 1$ selecionou de maneira dispersa

cada um dos quatro SNPs causais durante as dez execuções do SMS para esse *kernel* conforme a Tabela 8.18. Por conseguinte, caso a união dos dez subconjuntos gerados pelo SMS para esse *kernel* fosse adotada, os quatro SNPs causais seriam selecionados, porém, SNPs falsos-positivos também seriam selecionados. Em contrapartida, nenhum dos outros métodos, valores-p bruto e corrigido, selecionaram SNP causal algum, o que mostra ser um cenário muito complexo para a seleção de SNPs. Neste cenário, também ficou evidente que o subconjunto de marcadores com maior correlação não apresentou a melhor seleção, pois a maior correlação foi do *kernel* radial com $\gamma = 0,1$ que selecionou somente SNPs falsos-positivos, totalizando 19 marcadores.

Todos os *kernels* tiveram desempenho praticamente iguais quando somente os quatro SNPs causais foram avaliados para as medidas de correlação, MSE e MAPE como pode ser visto na Tabela 8.19. Essa observação juntamente com as correlações médias dos subconjuntos de SNPs gerados pelo SMS para os cinco *kernels* na quarta coluna da Tabela 8.17, mostra que o GA não conseguiria encontrar exatamente os quatro SNPs causais, a menos que ele fosse adaptado para buscar somente quádruplas de SNPs, pois essa estratégia não permitiria a seleção de muitos SNPs falsos-positivos como ocorreu nesse cenário simulado.

A partir da Figura 8.9 percebe-se que o comportamento dos MSEs médios, para os *kernels* linear e radial com $\gamma = 0,001$ e $\gamma = 0,01$, exibiram padrão semelhante, além de terem o mesmo ponto de corte nos 30 SNPs mais importantes indicados pelo *rank* da RF. O *kernel* radial com $\gamma = 1$ demonstrou um comportamento diferente dos outros *kernels*, pois MSE médio apresentou comportamento decrescente ao longo dos 20 primeiros SNPs ordenados pela RF, assumindo o ponto de mínimo global no 20^{o} SNP. Após o mínimo, o MSE médio aumentou até o 50^{o} SNP, mantendo o MSE médio constante em torno de 1,16. O gráfico (a) da Figura 8.9 que as variações do MSE médio absoluta e relativa estão entre os limites 1,15 e 1,40 para o *kernel* linear,respectivamente, próximas de 0,25 e 22%.

Com base na Tabela 8.20, nota-se um comportamento instável na ordenação da RF somente para o SNP causal 2 e uma das possíveis causas para isso é o coeficiente igual a 4 da quádrupla de SNPs causais usado na simulação dos dados, que nesse caso pode ser considerado pequeno, pois na distribuição dos efeitos marginais para os quatro SNPs informativos, a parte referente ao SNP 2 é relativamente pequena. Uma informação adicional é que o SNP 2 destaca seu efeito em indivíduos heterozigotos, isto é, quando a

Tabela 8.18 SNPs selecionados pelo *kernel* radial com $\gamma = 1$ para a simulação 5.

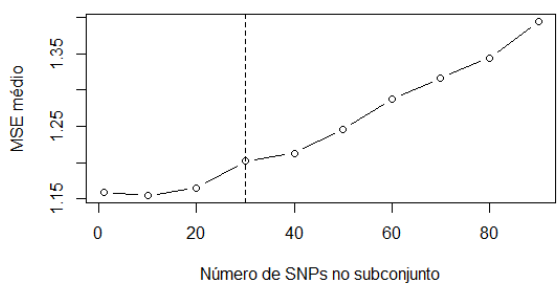
Iter.	# Ger.	SNPs	# SNPs	# SNPs causais	Correlação
1	48	68, **1**, 27, 83, 60, 12, 22, 76, 92, 35, 20, 18, 84, 26, 77, 95, 5, 30	18	1	0,17
2	44	**4**, 11, 68, 10, 27, 92, 83, 40, 20, 30, 22, 63, 97, 19, 51, 75, 7	17	1	0,20
3	39	11, 68, **1**, 22, 12, 93, 76, 60, 63, 8, 39, 89, 79, 20	14	1	0,16
4	85	**4**, **1**, 40, 22, 83, 92, 76, 12, 35, 98, 20, 53, 18, 93, 26, 48, 41, 44, 85, 15, **2**, 79	22	3	0,23
5	61	**4**, 68, 11, 22, 60, 83, 48, 29, 69, 85, 96, 18	12	1	0,20
6	93	27, **1**, 83, 30, 22, 75, 63, 91, 20, 70, 19, 98, 53, 86, 88	15	1	0,26
7	70	10, 68, 12, 77, 83, 97, 20, 76, 85, 35, 30, 46	12	0	0,18
8	58	68, 27, **1**, 10, 90, 53, 29	7	1	0,19
9	69	**3**, 11, 10, **1**, 68, 22, 20, 83, 60, 48, 99, 7, 23, 53, 100, 76, 64, 95, 8, 55, 32	21	2	0,20
10	50	11, 60, 30, 76, 12, 29, 85, **2**, 40, 75, 73, 53	12	1	0,17
Média	61,70	-	15,00	1,20	0,20
Desvio-padrão	17,68	-	4,59	0,79	0,03

variável desse marcador assume o valor 2.

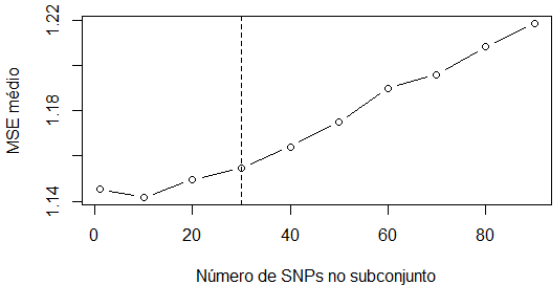
O *kernel* radial com $\gamma = 0,1$ demonstrou um ponto de corte com maior acurácia entre os *kernels* avaliados, pois selecionou dez vezes os SNPs 1, 3 e 4, e sete vezes o SNP 2, porém, na etapa de refinamento não selecionou o SNP 2 em execução alguma como observado na Tabela 8.21. É importante comentar que o SNP 2 foi selecionado duas

Tabela 8.19 Desempenho dos quatro SNPs causais nos cinco *kernels* avaliados para a simulação 5.

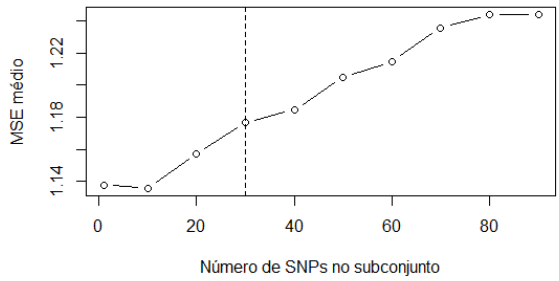
kernel	γ	r (σ_r)	MSE (σ_{mse})	MAPE (σ_{mape})
Linear	-	0,132 (0,01)	1,164 (0,23)	-0,072 (0,07)
Radial	0,001	0,131 (0,01)	1,160 (0,23)	-0,073 (0,08)
Radial	0,01	0,132 (0,01)	1,160 (0,22)	0,005 (0,09)
Radial	0,1	0,133 (0,01)	1,144 (0,20)	0,112 (0,13)
Radial	1	0,132 (0,01)	1,134 (0,20)	0,166 (0,16)



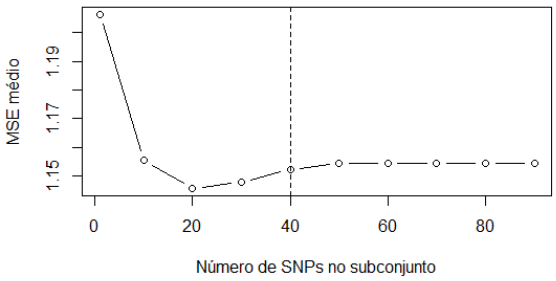
(a) *Kernel* linear na iteração 4.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 6.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 4.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 4.

Figura 8.9 Corte do SVR sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 5.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação à simulação 5. A linha tracejada indica o ponto de corte.

vezes na etapa de refinamento somente para o *kernel* radial com $\gamma = 1$ nas dez execuções realizadas para o SMS.

A Figura 8.10 mostra que o GA convergiu para a solução ótima em 30 gerações aproximadamente nos *kernels* analisados. Esses resultados mostram que o GA do SMS demonstrou comportamento semelhante para todas os conjuntos de dados simulados, pois o número de SNPs foi o mesmo em todas as simulações.

O método que mostrou o melhor desempenho para esse conjunto de dados simulados

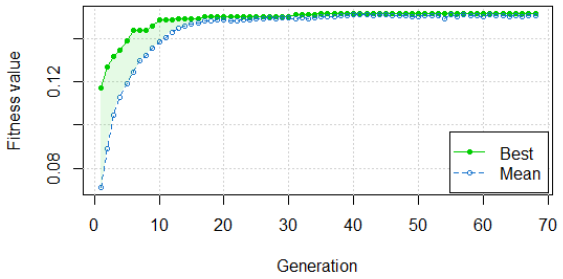
Tabela 8.20 *Rank* gerado por cada método para os quatro SNPs causais para a simulação 5.

Rank	Iter	Corte	SNPs causais			
			1	2	3	4
RF do SMS Linear	4	30	$6^{\underline{a}}$	$\mathbf{75^{\underline{a}}}$	$1^{\underline{a}}$	$2^{\underline{a}}$
RF do SMS Radial $\gamma = 0,001$	6	30	$5^{\underline{a}}$	$\mathbf{56^{\underline{a}}}$	$1^{\underline{a}}$	$2^{\underline{a}}$
RF do SMS Radial $\gamma = 0,01$	4	20	$7^{\underline{a}}$	$\mathbf{28^{\underline{a}}}$	$1^{\underline{a}}$	$2^{\underline{a}}$
RF do SMS Radial $\gamma = 0,1$	4	20	$6^{\underline{a}}$	$\mathbf{37^{\underline{a}}}$	$1^{\underline{a}}$	$2^{\underline{a}}$
RF do SMS Radial $\gamma = 1$	6	20	$7^{\underline{a}}$	$\mathbf{78^{\underline{a}}}$	$1^{\underline{a}}$	$2^{\underline{a}}$
Valor-p bruto	-	5	$\mathbf{51^{\underline{a}}}$	$\mathbf{80^{\underline{a}}}$	$\mathbf{22^{\underline{a}}}$	$\mathbf{49^{\underline{a}}}$
Valor-p ajustado	-	1	$\mathbf{51^{\underline{a}}}$	$\mathbf{80^{\underline{a}}}$	$\mathbf{22^{\underline{a}}}$	$\mathbf{49^{\underline{a}}}$
Blasso	-	53	$\mathbf{73^{\underline{a}}}$	$18^{\underline{a}}$	$\mathbf{61^{\underline{a}}}$	$32^{\underline{a}}$

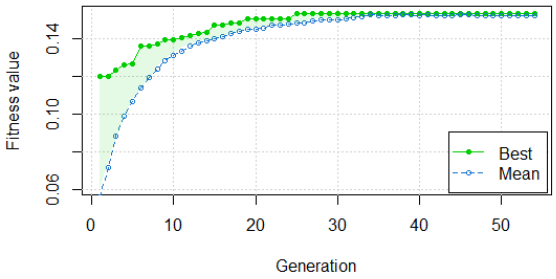
Tabela 8.21 Frequência da ausência dos quatro SNPs causais da simulação 5 nas 10 execuções do SMS.

kernel	Etapas do SMS	SNPs causais			
		1	2	3	4
Linear	Relevância + Corte	0	**9**	0	0
	Refinamento	**10**	**10**	**10**	**9**
Radial $\gamma = 0,001$	Relevância + Corte	0	**8**	0	0
	Refinamento	**10**	**10**	**10**	**10**
Radial $\gamma = 0,01$	Relevância + Corte	0	**9**	0	0
	Refinamento	**7**	**10**	**6**	**9**
Radial $\gamma = 0,1$	Relevância + Corte	0	**3**	0	0
	Refinamento	**6**	**10**	**7**	**9**
Radial $\gamma = 1$	Relevância + Corte	0	**8**	0	0
	Refinamento	**4**	**8**	**9**	**8**

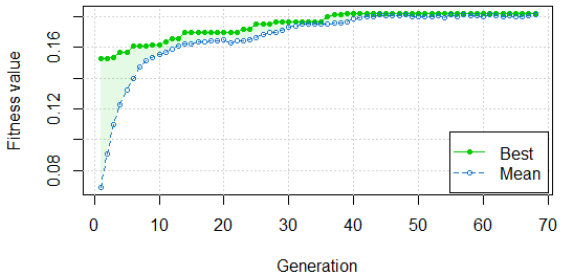
foi o SMS, a partir do critério em relação ao maior número de SNPs causais e o menor número de SNPs não-causais selecionados, pois o mesmo selecionou dois SNPs verdadeiros-positivos e 33 falsos-positivos. Logo em seguida, o Blasso ficou na segunda colocação com também dois SNPs causais e 51 SNPs não-causais. O valor-p corrigido e o valor-p bruto ficam nas terceira e quartas posições com zero SNPs verdadeiros-positivos e, um e cinco SNPs falsos-positivos respectivamente. O SMS selecionou os SNPs causais 1 e 3, enquanto o Blasso selecionou os SNPs causais 2 e 4, indicando um cenário simulado mais complexo do que os anteriores.



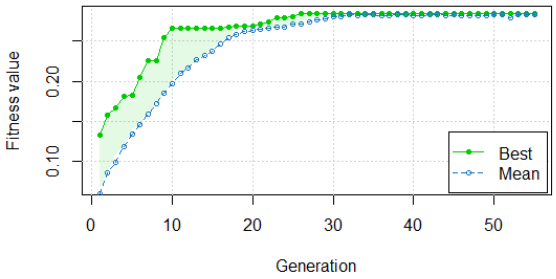
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,001$ na iteração 1.



(c) *Kernel* radial $\gamma = 0,01$ na iteração 1.



(d) *Kernel* radial $\gamma = 0,1$ na iteração 1.

Figura 8.10 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 5.

8.8 Simulação 6 - Três efeitos aditivos + uma interação de ordem 2 + uma interação de ordem 3 para classificação

A simulação 6 gerou um conjunto de dados com 138 controles (codificados como 0) e 862 casos (codificados como 1), isto é, um conjunto de dados desbalanceado entre as duas classes. Tal simulação permitiu avaliar o desempenho do SMS em uma nova situação de alta complexidade: a classificação em dados desbalanceados com efeitos aditivos e não-aditivos, onde estes foram compostos de uma interação entre pares de SNPs, e uma interação entre trios.

O desbalanceamento das classes simula um estudo de associação de escala genômica para uma doença rara, pois quaisquer amostras de indíviduos extraídas da população de interesse, terão mais indivíduos sadios do que doentes. Devido às instâncias ocorrerem com pequena frequência, modelos que descrevem a classe rara tendem a ser altamente especializados, por conseguinte, tais modelos são suscetíveis à presença de ruídos nos dados de treinamento (PANG-NING et al., 2006). Consequentemente, muitos algoritmos podem não detectar eficazmente instâncias da classe rara (PANG-NING et al., 2006). Existem diversas técnicas para abordar o problema do desequilíbrio de classes, porém, nenhuma delas foi usada no SMS com intuito de testar o SMS na sua versão mais simples em problemas de classificação com desbalanceamento entre as classes.

Tabela 8.22 Número de SNPs selecionados (SNPs), número de SNPs causais selecionados (V), AUC média em 10-*fold* por *kernel* para cada iteração do SMS, média e desvio-padrão (σ) das medidas anteriores para 10 execuções do SMS em relação à simulação 6.

Iter	Linear		Radial $\gamma = 0,001$		Radial $\gamma = 0,01$		Radial $\gamma = 0,1$		Radial $\gamma = 1$	
	SNPs (V)	AUC	SNPs (V)	AUC	SNPs (V)	AUC	SNPs (V)	AUC	SNPs (V)	AUC
1	**53 (2)**	**0,619**	**50 (2)**	**0,712**	9 (2)	0,584	70 (8)	0,881	19 (7)	0,778
2	54 (5)	0,589	47 (2)	0,700	11 (4)	0,600	71 (7)	0,883	19 (8)	0,820
3	54 (4)	0,593	48 (4)	0,712	17 (1)	0,613	71 (7)	0,868	19 (6)	0,825
4	11 (0)	0,586	48 (2)	0,689	12 (1)	0,629	70 (8)	0,834	19 (7)	0,828
5	41 (2)	0,584	54 (4)	0,703	11 (4)	0,593	70 (8)	0,831	19 (6)	0,844
6	49 (1)	0,609	47 (2)	0,695	8 (3)	0,598	71 (7)	0,870	19 (8)	0,819
7	49 (3)	0,601	55 (3)	0,694	23 (4)	0,671	70 (8)	0,837	20 (5)	0,877
8	9 (2)	0,584	46 (3)	0,692	22 (2)	0,660	71 (8)	0,853	18 (8)	0,838
9	47 (2)	0,585	50 (3)	0,670	28 (6)	0,651	70 (8)	0,827	19 (8)	0,841
10	50 (3)	0,595	49 (3)	0,689	**36 (1)**	**0,713**	70 (8)	0,817	19 (7)	0,865
Média	41,7 (2,4)	0,595	49,4 (2,8)	0,696	17,7 (2,8)	0,631	70,4 (7,7)	**0,850**	19 (7)	**0,833**
σ	17,2 (1,4)	0,012	3 (0,8)	0,012	9,3 (1,7)	0,042	0,5 (0,5)	0,024	0,5 (1,1)	0,027

Como as maiores médias da área abaixo da curva ROC (AUC), nas dez iterações do

SMS, são as relativas aos *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$, repectivamente, iguais a 0,850 e 0,833 (as médias em negrito na penúltima linha da Tabela 8.22). Tomou-se a decisão de verificar se a diferença estatística entre elas é significativa, pois em caso negativo, o subconjunto de SNPs escolhido será a união dos subconjuntos selecionados pelo *kernel* radial com $\gamma = 1$, devido a média dos SNPs selecionados ser menor e igual a 19 (8.22). Os *boxplots* da Figura 8.11 permitem concluir que a média da AUC do *kernel* radial com $\gamma = 0,1$ é ligeiramente superior à média do kernel radial com $\gamma = 1$, mas para a tomada de decisão foi realizado um teste de hipóteses para a comparação das médias.

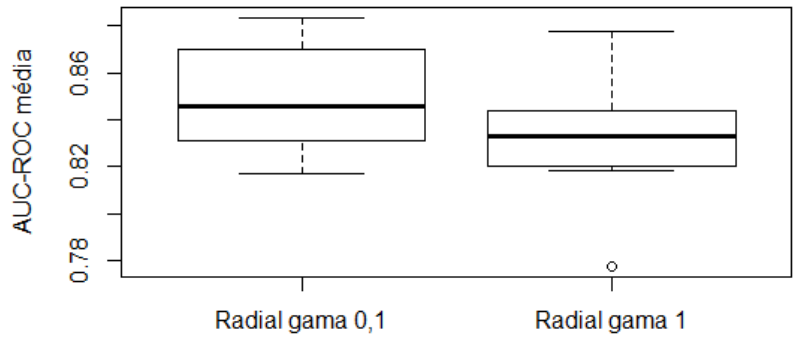


Figura 8.11 *Boxplots* das médias da AUC nas 10 execuções do SMS para os *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$ em relação à simulação 6.

Segundo Gravetter e Wallnau (2013), um experimento com medidas repetidas é aquele em que a variável dependente é medida duas ou mais vezes para cada indivíduo em uma única amostra, ou seja, o mesmo grupo de indivíduos é utilizada em todas as condições de tratamento. Consequentemente, como as 10 execuções para os dois *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$ do SMS são baseadas no mesmo conjunto de dados; então, fez-se um teste de hipóteses para comparação de médias em amostras pareadas denominado teste-t pareado (GRAVETTER; WALLNAU, 2013). Assim, caso existam evidências das médias serem iguais, o subconjunto de marcadores escolhido será o relativo ao *kernel* radial com $\gamma = 1$, pois seus dez subconjuntos de SNPs possuem um número substancialmente menor do que os referentes ao *kernel* radial com $\gamma = 0,1$ como notado na Tabela 8.22.

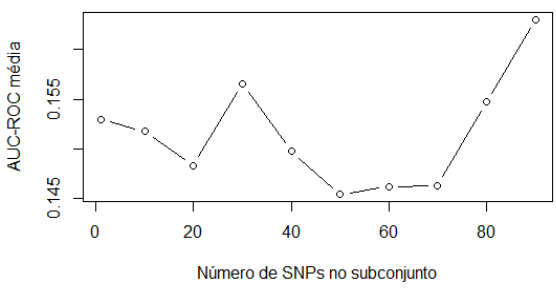
O teste t pareado possui duas hipóteses para sua correta aplicação a saber: é necessário que as amostras sejam extraídas de populações normais e as variâncias populacionais

das duas amostras sejam iguais (GRAVETTER; WALLNAU, 2013). Posto isso, para a verificação da hipótese de normalidade das AUC médias, realizaram-se dois testes de Shapiro-Wilk, os quais apresentaram valores-p iguais 0,3018 e 0,6667, respectivamente, para os *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$. Daí, como ambos valores-p são maiores que 0,05, existem evidências de que as distribuições amostrais das AUC médias para os dois *kernels* seguem distribuições normais a um nível de significância de 0,05. Para a verificação da igualdade entre variâncias, usou-se o teste F, onde o mesmo indicou valor-p igual a 0,7221, ou seja, há evidências de que as variâncias populacionais são iguais, pois esse valor-p é superior ao nível de significância 0,05. Finalmente, aplicou-se o teste t pareado bilateral e obteve-se valor-p igual a 0,3006

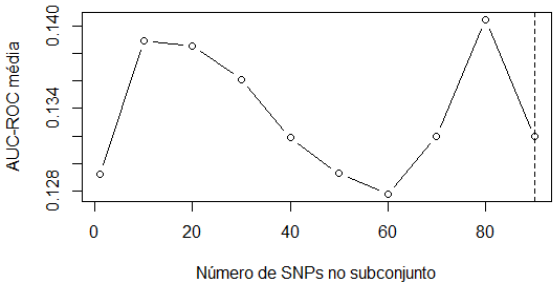
A partir da Figura 8.12, nota-se que os cortes foram o 100 para o *kernel* linear, 90 para os radiais com $\gamma = 0,001$ (não mostrado na Figura 8.12) e $\gamma = 0,01$, 30 para o radial com $\gamma = 0,1$ e 20 para o radial com $\gamma = 1$. Os valores máximos assumidos pela AUC média para os cinco *kernels* variaram consideravelmente, com 0,165 para o linear, 0,140 para o radial com $\gamma = 0,01$, 0,80 para o radial com $\gamma = 0,1$ e 0,70 para o radial com $\gamma = 1$. Contudo, o comportamento mais adequado foi o do *kernel* radial com $\gamma = 1$, pois ele aumentou até o ponto de máximo e depois diminuiu até estabilizar em 0,50.

O *rank* da RF foi baseado no gVI (importância de gini da variável), pois o mesmo apresentou resultados superiores ao pVI (importância de permutação da variável), entretanto, esses resultados não foram mostrados. Logo, o *rank* criado pelo gVI está exemplificado na Tabela 8.23 para uma interação do SMS para cada kernel usado. Nota-se que todos os oito SNPs causais foram ordenados adequadamente e foram selecionados pela etapa de corte em todos os *kernels*. De forma distinta, o método de seleção baseado no valor-p bruto do teste qui-quadrado não identificou os SNPs 4, 6, 7 e 8. Além desses quatro SNPs informativos, o método do valor-p corrigido não selecionou também o SNP 3.

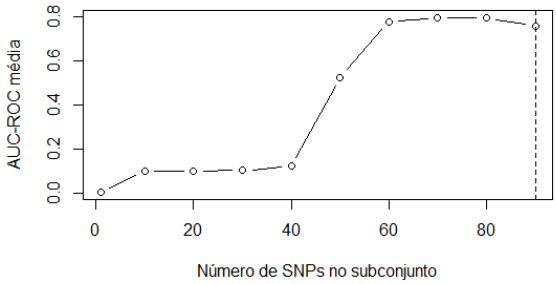
Os SNPs 1 e 2 praticamente não foram selecionados pelo GA nas dez execuçõe do SMS para os *kernels* linear e radiais com $\gamma = 0,001$ e $\gamma = 0,01$, porém foram identificados para os *kernels* radiais com $\gamma = 0,1$ e $\gamma = 1$ em quase todas as dez execuções do SMS conforme indica a Tabela 8.24. Para os outros seis SNPs causais, houve muita instabilidade na seleção para esses três *kernels*, o que mostra que os mesmos são inadequados para selecionar SNPs com essa arquitetura genômica. Por outro lado, os *kernels* radiais com



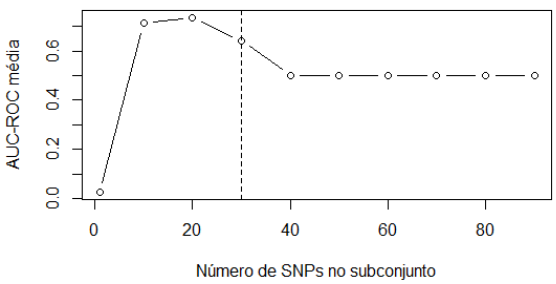
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,01$ na iteração 10.



(c) *Kernel* radial $\gamma = 0,1$ na iteração 2.



(d) *Kernel* radial $\gamma = 1$ na iteração 7.

Figura 8.12 Corte do SVM sobre o *rank* da RF para os *kernels* linear e radial em relação à simulação 6.

Kernels (a) linear, (b) radial com $\gamma = 0,01$, (c) radial com $\gamma = 0,1$ e (d) radial com $\gamma = 1$ em relação ao modelo 6. A linha tracejada indica o ponto de corte.

Tabela 8.23 *Rank* gerado por cada método para os oito SNPs causais para a simulação 6.

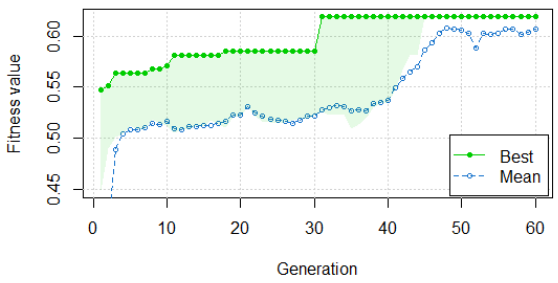
Rank	Iter	Corte	SNPs causais							
			1	2	3	4	5	6	7	8
RF do SMS Linear	1	100	1ª	2ª	3ª	5ª	4ª	17ª	13ª	3ª
RF do SMS Radial $\gamma = 0,001$	3	100	1ª	2ª	3ª	5ª	4ª	15ª	13ª	8ª
RF do SMS Radial $\gamma = 0,01$	2	20	1ª	2ª	3ª	5ª	4ª	16ª	13ª	8ª
RF do SMS Radial $\gamma = 0,1$	1	20	1ª	2ª	3ª	5ª	4ª	16ª	13ª	6ª
RF do SMS Radial $\gamma = 1$	7	30	1ª	2ª	3ª	5ª	4ª	15ª	13ª	7ª
Valor-p bruto	-	4	1ª	2ª	4ª	**60ª**	3ª	**44ª**	**43ª**	**11ª**
Valor-p corrigido	-	3	1ª	2ª	**4ª**	**60ª**	3ª	**44ª**	**43ª**	**11ª**

$\gamma = 0,1$ e $\gamma = 1$ conseguiram selecionar praticamente todos os oitos SNPs informativos nas dez execuções do SMS.

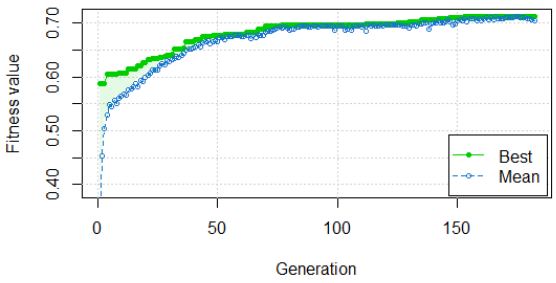
A convergência do GA variou bastante para cada *kernel* como evidencia a Figura 8.13. Os *kernels* linear e radial com $\gamma = 1$ demonstraram maior diversidade nas populações de subconjuntos de SNPs durante as gerações como é mostrado pela área entre as curvas da média da população e do melhor indivíduo do GA.

Tabela 8.24 Frequência da ausência dos oito SNPs causais da simulação 6 nas dez execuções do SMS.

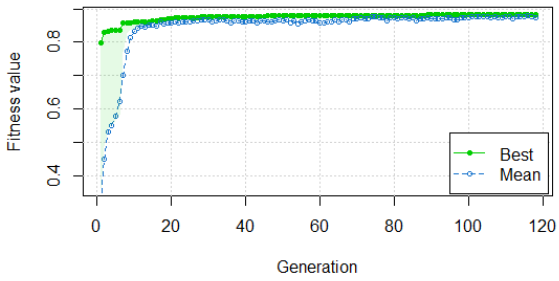
kernel	Etapas do SMS	SNPs causais							
		1	2	3	4	5	6	7	8
Linear	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	**10**	**10**	7	5	8	4	6	5
Radial $\gamma = 0,001$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	**10**	9	3	7	5	7	2	8
Radial $\gamma = 0,01$	Relevância + Corte	0	0	0	0	0	1	0	0
	Refinamento	**10**	**10**	6	2	9	2	3	8
Radial $\gamma = 0,1$	Relevância + Corte	0	0	0	0	0	0	0	0
	Refinamento	0	1	1	0	0	0	0	0
Radial $\gamma = 1$	Relevância + Corte	0	0	0	0	0	2	0	0
	Refinamento	0	0	1	1	0	3	1	1



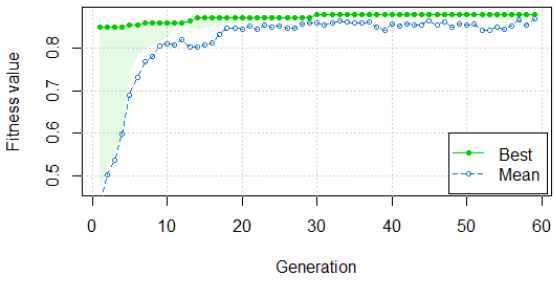
(a) *Kernel* linear na iteração 1.



(b) *Kernel* radial $\gamma = 0,01$ na iteração 10.



(c) *Kernel* radial $\gamma = 0,1$ na iteração 2.



(d) *Kernel* radial $\gamma = 1$ na iteração 7.

Figura 8.13 Convergência da aptidão (AUC média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação à simulação 6.

A Tabela 8.25 mostra a quantidade de SNPs produzida pela união das dez execuções do SMS para cada *kernel*. Assim, o subconjunto selecionado de SNPs ao final do SMS foi adotado como sendo a união das dez execuções do *kernel* radial com $\gamma = 1$, pois o mesmo

não apresentou diferença significativa em relação a média da AUC do *kernel* radial com $\gamma = 0,1$, selecionando 36 SNPs. Esse resultado é consideravelmente inferior ao número de elementos da união do *kernel* radial com $\gamma = 0,1$, o qual é igual a 93. Caso o conjunto interseção fosse adotado como solução do SMS para *kernel* radial com $\gamma = 1$, o número total de SNPs selecionados seria seis, entretanto, somente três são causais, enquanto que a interseção para o *kernel* radial com $\gamma = 0,1$ apresentou 27 SNPs com 6 SNPs informativos. Portanto, esse conjunto de dados é complexo e somente o kernel radial com γ assumindo os valores 0,1 e 1 conseguiu produzir subconjuntos de SNPs mais estáveis nas dez execuções do SMS a ponto de existir interseção não-vazia entre eles como é mostrado na Tabela 8.25.

Tabela 8.25 União e interseção dos SNPs selecionados pelo SMS nas 10 execuções para cada *kernel* em relação à simulação 6.

Kernel	γ	# SNPs[a]	# V[b]
União SMS Linear	-	98	6
União SMS Radial	0,001	99	7
União SMS Radial	0,01	68	5
União SMS Radial	0,1	93	8
União SMS Radial	**1**	**36**	**8**
Interseção SMS Linear	-	0	0
Interseção SMS Radial	0,001	0	0
Interseção SMS Radial	0,01	0	0
Interseção SMS Radial	0,1	27	6
Interseção SMS Radial	**1**	**6**	**3**

[a] Número total de SNPs selecionados pelo SMS.

[a] Número total de SNPs causais selecionados pelo SMS.

O método que mostrou o melhor desempenho para esse conjunto de dados simulados foi o SMS, a partir do critério em relação ao maior número de SNPs causais selecionados, pois o mesmo selecionou todos os oito SNPs verdadeiros-positivos e 28 falsos-positivos. Logo em seguida, o valor-p bruto ficou na segunda colocação com quatro SNPs causais (SNPs 1, 2, 5 e 3 na ordem do método) e três SNPs não-causais. O valor-p corrigido ficou na terceira e última posição com três SNPs verdadeiros-positivos (SNPs 1, 2 e 5 na ordem do método) e nenhum SNP falso-positivo. Percebe-se a dificuldade dos métodos do valor-p bruto e valor-p corrigido para encontrar as interações de ordem 2 e 3, enquanto para o SMS, o obstáculo maior foi a introdução significativa de SNPs falsos-positivos.

8.9 Dados Simulados do QTLMAS 2011

A codificação para os marcadores dos cinco cromossomos foi realizada da seguinte forma: o SNP1 foi alocado na posição 1, o SNP2 na posição 2, e assim sucessivamente até o SNP1998 (1+1997) que indica o término do cromossomo 1. Para o cromossomo 2, o SNP1999 é o marcador inicial e o SNP3396 (1999+1997), o final. No cromossomo 3, o SNP3997 é o marcador inicial e o SNP5994 (3997+1997), o final. Para o cromossomo 4, o SNP5995 é primeiro marcador e o SNP7792 (5995+1997) é o último marcador e, finalmente, no cromossomo 5, o marcador inicial é 7993 e o final é 9990 (7993+1997). A Figura 8.14 mostra essa codificação.

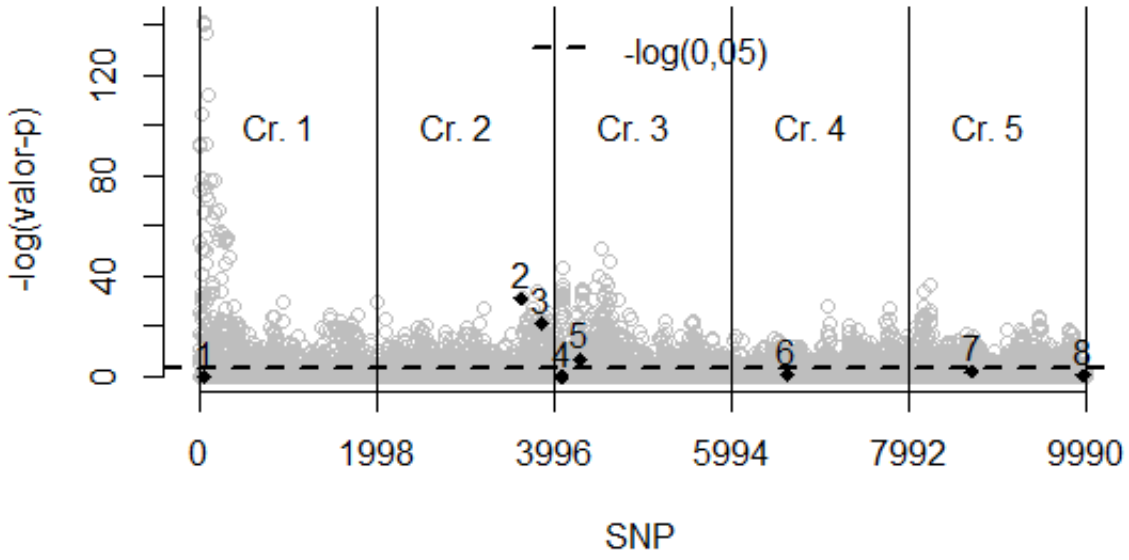


Figura 8.14 Valor-p bruto dos 9.990 SNPs onde a linha tracejada indica o limite inferior $-log(0,05)$ para seleção.

O valor-p bruto foi calculado a partir do teste de hipóteses, com as hipóteses nula e alternativa, $H_0 : \beta_1 = 0$ e $H_a : \beta_1 \neq 0$, onde β_1 é o coeficiente angular da reta estimada por mínimos quadrados baseada nas informações de genótipo-fenótipo de 2.000 indivíduos da amostra. A regressão linear é calculada para cada um dos 9.990 marcadores e plotados na Figura 8.14 para demonstrar que os SNPs 1 (posição 57), 4 (posição 4096), 6 (posição 6635), 7 (posição 8718) e 8 (posição 9976) não serão selecionados pelo limite inferior de

$-log(0,05)$, o qual indica uma significância estastística mínima de 0,05. Cabe destacar que os marcadores 1 e 4, respectivamente, nos cromossomos 1 e 3, não possuem variação alélica alguma, o que implica a impossibilidade de detectar qualquer sinal de associação entre esses SNPs e o fenótipo por qualquer técnica de seleção. Esse fato também foi observado por Fu et al. (2012), portanto, nenhuma técnica consegue detectar os SNPs 1 e 4. Entretanto, os marcadores nas posições 54, 55, 56, 58, 59 e 60, que estão em alto desequilíbrio de ligação com o da posição 57, possuem elevada significância estatística dadas pelos valores-p brutos, respectivamente, iguais a $1,08 \times 10^{-11}$, $5,05 \times 10^{-23}$, $7,25 \times 10^{-7}$, $4,12 \times 10^{-62}$, $4,18 \times 10^{-62}$ e $7,62 \times 10^{-62}$. Logo, esses SNPs marcam uma região no genoma simulado que está associada estatisticamente ao fenótipo. Raciocínio análogo pode ser aplicado aos marcadores 4, 6, 7 e 8 que possuem valor-p bruto superior a 0,05, mas estão em desequilíbrio de ligação com outros SNPs que possuem associação significativa com o fenótipo.

Pela Tabela 8.26, percebe-se que somente os QTLs 2, 3 e 5 podem ser selecionados pelo método do valor-p bruto e que, dentre esses, somente os marcadores 2 e 3 podem ser selecionados pelo valor-p corrigido, onde adotou-se o limite superior para seleção de 0,05. Com isso, o *rank* gerado pela RF parece ser mais adequado neste cenário simulado do que o gerado pelo valor-p da regressão linear de um único *locus* por permitir que o SMS selecione marcadores que estão mais próximos aos oito QTLs (Tabela 8.28).

Tabela 8.26 Valores-p bruto e corrigido por Bonferroni dos oito QTLs.

QTL	Cromossomo	Posição (cM)	Posição do SNP	Valor-p bruto	Valor-p corrigido
1	1	2,85	**57**	**1,00**[a]	$\mathbf{9,99 \times 10^{3}}$
2	2	81,90	3636	$2{,}54\times10^{-14}$	$2{,}54\times10^{-10}$
3	2	93,75	3873	$6{,}39\times10^{-10}$	$6{,}39\times10^{-6}$
4	3	5,00	**4096**	**1,00**[a]	$\mathbf{9,99 \times 10^{3}}$
5	3	15,00	4296	$8{,}38\times10^{-4}$	**8,37**
6	4	32,20	**6638**	$\mathbf{5,91 \times 10^{-1}}$	$\mathbf{5,90 \times 10^{3}}$
7	5	36,30	**8718**	$\mathbf{1,44 \times 10^{-1}}$	$\mathbf{1,44 \times 10^{3}}$
8	5	99,20	**9976**	$\mathbf{6,81 \times 10^{-1}}$	$\mathbf{6,80 \times 10^{3}}$

[a] Adotou-se valor-p igual a 1,00 quando não há variação alélica no SNP.

Os modelos de SVR usados pelo SMS demonstraram que o número de falsos-positivos aumentou de acordo com a diminuição do γ como notado na Tabela 8.27. A partir dessa observação é possível questionar a possível existência de um único γ ótimo no sentido de selecionar o maior número de marcadores verdadeiros e o mínimo de falsos. Usou-se

um raio de 9 cM em relação a cada QTL para considerar o SNP selecionado pelo SMS como marcador verdadeiro, pois como pode ser visto na Tabela 8.29, o método RHM considerou que o marcador na posição 91,05 marca o QTL na posição 99,20, ou seja, um raio de 8,15 cM. Além disso, como comentado por Fu et al. (2012), os SNPs significantes no cromossomo 1 cobrem um grande intervalo entre 0,15 cM e 15,30 cM, o que mostra que o raio de 9 cM adotado é coerente com a estrutura de LD observada nos marcadores simulados, podendo este valor de 9 cM ser considerado um raio limite conservador para a classificação de SNPs verdadeiros-positivos.

Tabela 8.27 Número de SNPs selecionados próximos aos oito QTLs, número de QTLs marcados por pelo menos um SNP selecionado, número de SNPs falso-positivos e total de SNPs selecionados por cada modelo SMS na iteração 1.

Método	$\overline{r}$	#SNPs verdadeiros (1)	#QTLs marcados (2)	#SNPs falsos-positivos (3)	#SNPs selecionados (4) = (1) + (3)
SMS2 Linear	0,53	30	8	35	65
SMS2 Radial $\gamma = 0,001$	0,57	70	8	128	198
SMS2 Radial $\gamma = 0,01$	0,58	29	8	54	83
SMS2 Radial $\gamma = 0,1$	0,52	14	6	14	28
Interseção SMS2	-	5	3	1	6
União SMS2	-	96	8	187	253

Os resultados do SMS apresentados na Tabela 8.28 permitem inferir a robustez do método em encontrar marcadores SNPs próximos aos oito QTLs simulados em todos os modelos analisados, entretanto, as maiores divergências entre os *kernels* ocorreram nos QTLs 3, 5, 6 e 8. Na, iteração 1 do SMS, o *kernel* que demonstrou simultaneamente maior correlação média ($\overline{r} = 0,58$), melhor seleção, maior acurária (menor distância em relação aos 8 QTLs como indicado na Tabela 8.28) e menor número de falsos-positivos foi o radial com $\gamma = 0,01$. Cabe destacar que esse *kernel* foi o único que conseguiu identificar o sinal do segundo QTL no cromossomo 5 com alta precisão, cuja posição real é 99,20 e a posição do marcador selecionado é 99,35. Já nas iterações 2 e 3 do SMS, não houve um *kernel* que teve melhor desempenho em todos os quesitos como o radial com $\gamma = 0,01$ na iteração 1, mas a união dos quatro *kernels* apresentou subconjuntos de SNPs similares nas três iterações (Tabela 8.28), onde os dois QTLs em epistasia foram identificados pelas três soluções do SMS, mostrando estabilidade do SMS no processo de seleção de atributos.

Os resultados obtidos por Demeure et al. (2012) mostram que, dentre os quatro

Tabela 8.28 Comparação dos SNPs mais próximos dos oito QTLs selecionados pelo SMS2 para cada *kernel* utilizado nas iterações 1, 2 e 3. Os números em negrito representam os melhores resultados de cada união por iteração. Os número sublinhados representam os melhores resultados das três iterações.

Método	Iteração	Cr. 1	Cr. 2	Cr. 2	Cr. 3	Cr. 3	Cr. 4	Cr. 5	Cr. 5
Posição real	1	2,85	81,90	93,75	5,00	15,00	32,20	36,30	99,20
SMS2 Linear	1	**2,90**	**81,90**	95,70	**4,80**	12,50	27,65	35,75	90,50
SMS2 $\gamma = 0,001$	1	**2,90**	**81,90**	**93,75**	**4,80**	**15,05**	**32,05**	**36,25**	91,40
SMS2 $\gamma = 0,01$	1	**2,90**	**81,90**	95,70	**4,80**	15,85	**32,05**	**36,25**	**99,35**
SMS2 $\gamma = 0,1$	1	**2,90**	**81,90**	90,65	**4,80**	-	35,95	34,20	-
Interseção SMS2	1	**2,90**	**81,90**	-	**4,80**	-	-	-	-
União SMS2	1	**2,90**	**81,90**	**93,75**	**4,80**	**15,05**	**32,05**	**36,25**	**99,35**
SMS2 Linear	2	**2,90**	**81,90**	95,00	**4,80**	15,85	26,30	**36,25**	91,25
SMS2 $\gamma = 0,001$	2	**2,90**	**81,90**	**93,75**	**4,80**	**14,20**	**32,05**	**36,25**	96,00
SMS2 $\gamma = 0,01$	2	**2,90**	**81,90**	95,00	**4,80**	13,85	26,30	**36,25**	**99,35**
SMS2 $\gamma = 0,1$	2	**2,90**	**81,90**	**93,75**	**4,80**	**14,20**	35,95	**36,25**	-
Interseção SMS2	2	**2,90**	**81,90**	-	**4,80**	-	26,30	**36,25**	-
União SMS2	2	**2,90**	**81,90**	**93,75**	**4,80**	**14,20**	**32,05**	**36,25**	**99,35**
SMS2 Linear	3	**2,90**	**81,90**	93,50	**4,80**	**15,20**	**32,05**	35,60	-
SMS2 $\gamma = 0,001$	3	**2,90**	**81,90**	**93,85**	**4,80**	13,85	**32,05**	**36,25**	95,35
SMS2 $\gamma = 0,01$	3	**2,90**	**81,90**	95,75	**4,80**	**15,20**	**32,05**	**36,25**	99,60
SMS2 $\gamma = 0,1$	3	**2,90**	**81,90**	**93,85**	**4,80**	13,85	**32,05**	-	**99,40**
Interseção SMS2	3	**2,90**	**81,90**	-	**4,80**	-	**32,05**	-	-
União SMS2	3	**2,90**	**81,90**	**93,85**	**4,80**	**15,20**	**32,05**	**36,25**	**99,40**

métodos avaliados (MMA, RHM, GENMIX e BVS), somente o método GENMIX encontrou SNPs próximos aos oito QTLs simulados (Tabela 8.29). Entretanto, quando o método SMS é inserido na comparação com os demais, o mesmo detecta os oito QTLs com uma acurácia superior aos quatro métodos avaliados por Demeure et al. (2012), principalmente, em relação aos QTLs 6 (*imprinting*), 7 e 8 (par de QTLs com epistasia). Isso demonstra o potencial do SMS para buscar interações entre QTLs (epistasia) e efeitos marginais complexos tais como *imprinting* genômico. Em contrapartida, o SMS mostrou a mesma desvantagem observada nos conjuntos de dados simulados pelo SCRIME em relação ao número excessivo de falsos-positivos. Outro ponto a considerar é a estabilidade do SMS nas três iterações, pois a maior diferença foi de 0,85 cM entre os dois SNPs selecionados nas iterações 1 e 2 (posições 15,05 e 14,20) que marcam o segundo QTL do cromossomo 3 (quinta linha da Tabela 8.29). Esse comportamento estável se estendeu também para o número de SNPs falsos-positivos do SMS, sendo 187, 158 e 175, respectivamente, para as iterações 1, 2 e 3.

Outra comparação entre 12 técnicas de seleção de marcadores SNPs foi realizada por Demeure et al. (2012), onde na Tabela 8.30 foi inserido os resultados do SMS para mostrar seu potencial frente aos outros métodos. Em comparação com as técnicas de seleção usadas

Tabela 8.29 Posições (cM) dos QTL identificados com os quatro métodos usados por Dashab et al. (2012) juntamente com as seleções do SMS nas três iterações.

Cr.	Posição (cM)	Método						
		SMS2[b]	**SMS2**[c]	**SMS2**[d]	MMA	RHM	GENMIX	BVS
1	2,85	**2,90**	**2,90**	**2,90**	3,55	2,50	2,70	2,75
2	81,90	**81,90**	**81,90**	**81,90**	**81,90**	-[a]	82,30	83,10
2	93,75	**93,75**	**93,75**	93,85	-[a]	95,95	95,80	**93,75**
3	5,00	4,80	4,80	4,80	4,80	**4,85**	4,80	4,80
3	15,00	**15,05**	14,20	15,20	16,52	14,90	11,10	14,80
4	32,20	**32,05**	**32,05**	**32,05**	-[a]	-[a]	31,70	28,30
5	36,30	**36,25**	**36,25**	**36,25**	36,19	35,95	36,00	35,15
5	99,20	**99,35**	**99,35**	99,40	91,29	91,05	91,20	-[a]
Falsos positivos	-	**187**	**158**	**175**	2	6	4	2
Total de SNPs	-	283	249	261	8	12	12	9

[a] Falsos-negativos.
[b] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS com os quatro métodos avaliados (MMA, RHM, GENMIX e BVS) por Dashab et al. (2012). A solução do SMS é a união das soluções de cada kernel para a iteração 1.
[c] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS com os quatro métodos avaliados (MMA, RHM, GENMIX e BVS) por Dashab et al. (2012). A solução do SMS é a união das soluções de cada kernel para a iteração 2.
[d] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS com os quatro métodos avaliados (MMA, RHM, GENMIX e BVS) por Dashab et al. (2012). A solução do SMS é a união das soluções de cada kernel para a iteração 3.

em Demeure et al. (2012), o SMS foi o único que selecionou SNPs que flanquearam os oito QTLs, além de demonstrar maior acurácia como observado em 8.31. Porém, o SMS selecionou um número de marcadores falso-positivos muito superior a maior parte dos métodos avaliados, sendo essa quantidade igual a 187 (última linha da segunda terceira coluna da Tabela 8.31).

A abordagem de múltiplos modelos usados no SMS2, a qual é permitida pela variação do parâmetro γ no *kernel* radial e pelo uso do *kernel* linear no SVR, possibilita ao SMS2 identificar melhor QTLs com somente efeitos aditivos com γs próximos de 0 e interações entre QTLs com γs mais próximos de 1. O SMS2 nas iterações 1 e 2 superou as outras metodologias de seleção nos quesitos número de QTLs identificados e acurácia na detecção, porém, obteve desempenho inferior em relação à quantidade de SNPs falsos-positivos como pode ser visto na Tabela 8.31.

Como destaca Demeure et al. (2012), o primeiro QTL no cromossomo 5, é detectado por todos os métodos avaliados, com exceção do GBLUP, BayesB e do BayesCπ de Zeng et al. (2012) e do modelo misto aproximado denominado de EMMAX. De maneira contrária, nenhum dos métodos foi capaz de detectar o segundo QTL na posição 99,20 cM, porém, um sinal positivo foi identificado no intervalo 91-92 cM por todas as abordagens usadas

Tabela 8.30 Localização dos QTLs simulados dependendo do método/modelo usado. Adaptado de Demeure et al. (2012).

Método	Referência	Cr. 1	Cr. 2	Cr. 2	Cr. 3	Cr. 3	Cr. 4	Cr. 5	Cr. 5
Posição real	-	2,85	81,90	93,75	5,00	15,00	32,20	36,30[a]	99,20
SMS2 iter 1[c]	-	**2,90**[b]	**81,90**	**93,75**	**4,80**	**15,05**	**32,05**	**36,25**	**99,35**
SMS2 iter 2[d]	-	**2,90**[b]	**81,90**	**93,75**	**4,80**	14,20	**32,05**	**36,25**	**99,35**
SMS2 iter 3[e]	-	**2,90**[b]	**81,90**	93,85	**4,80**	15,20	**32,05**	**36,25**	99,40
GBLUP	Zeng et al. (2012)	2,95	83,00	-	4,75	-	28,20	-	-
BayesB	Zeng et al. (2012)	**2,85**	83,00	93,70	4,75	15,80	27,90	-	-
BayesC	Dashab et al. (2012)	2,75	83,10	93,40	**4,80**	14,80	28,30	35,10	-
BayesCn	Schurink, Janss e Heuven (2012)	1,60	83,10	93,40	2,90	14,80	28,30	35,10	-
BayesCn	Zeng et al. (2012)	2,75	83,00	93,60	4,60	16,60	-	-	-
LASSO1	Usai, Carta e Casu (2012)	**2,90**[b]	**81,90**	95,80	**4,80**	16,10	28,00	35,10	-
LASSO2	Usai, Carta e Casu (2012)	**2,90**[b]	81,80	94,00	**4,80**	16,70	34,90	36,80	-
LASSO3	Usai, Carta e Casu (2012)	**2,90**[b]	81,80	95,80	**4,80**	15,80	28,00	36,80	-
MM single SNP	Dashab et al. (2012)	3,55	82,00	-	**4,80**	16,50	-	36,20	-
MM Haplotype	Dashab et al. (2012)	2,50	-	96,00	**4,80**	14,90	-	35,90	-
MM Phylogeny	Dashab et al. (2012)	2,70	82,30	95,80	**4,80**	11,10	31,70	36,00	-
EMMAX	Fu et al. (2012)	**2,90**[b]	83,10	-	**4,80**	-	-	-	-

[a] O valor original era 36,60 cM no trabalho de Demeure et al. (2012), entretanto, no artigo seminal de Elsen et al. (2012) que descreve a geração dos dados simulados do QTLMAS 2011 , o valor correto é 36,30 cM.

[b] O SNP na posição 2,85 cM não possui variação alélica como observado por Fu et al. (2012), por isso, o marcador na posição 2,90 foi colocado em negrito.

[c] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS é a união das soluções de cada *kernel* na iteração 1.

[d] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS2 é a união das soluções de cada *kernel* na iteração 2.

[e] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS2 é a união das soluções de cada *kernel* na iteração 3.

de modelos mistos lineares (DASHAB et al., 2012). Outro ponto de destaque citado por Demeure et al. (2012) é que o conjunto de dados construído para o QTLMAS 2010 possuía interações similares que foram detectadas por todos os métodos (MUCHA et al., 2011). Estes resultados podem ser explicados pela hipótese de que existe um efeito no primeiro QTL somente se houver o genótipo "1 1"no segundo QTL.

A etapa de corte feita no cromossomo 1 mostra que o MSE médio de cada SVR varia substancialmente com o *kernel*, linear ou radial, onde neste último, sofre mudanças significativas com a variação do γ. Na Figura 8.15, o comportamento do MSE médio do SVR é mais coerente com o *rank* da RF para o *kernel* linear do que para o radial,

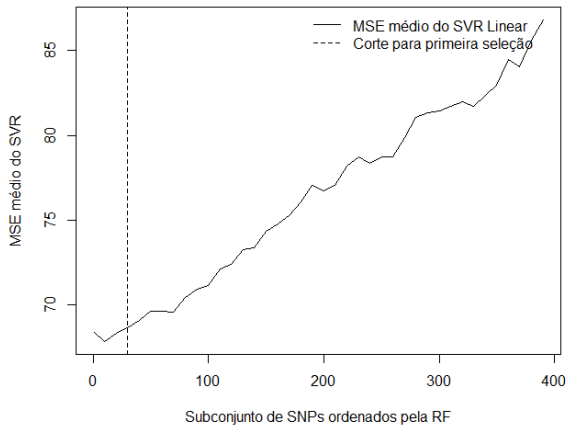
Tabela 8.31 Comparação dos resultados do mapeamento dos oito QTLs (adaptado de Demeure et al. (2012)).

Método	Referência	# QTLs detectados	Falsos positivos	QTL não detectados (a)	Distância média do QTL (cM)
GBLUP	Zeng et al. (2012)	5	**0**	3, 5, 8	1,32
BayesB	Zeng et al. (2012)	6	3	7, 8	1,21
BayesC	Dashab et al. (2012)	7	2	8	0,96
BayesCn	Schurink, Janss e Heuven (2012)	5	**0**	5, 6, 8	0,83
BayesCn	Zeng et al. (2012)	5	1	6, 7, 8	0,45
LASSO1	Usai, Carta e Casu (2012)	7	Inúmeros	6	0,83
LASSO2	Usai, Carta e Casu (2012)	7	Inúmeros	6	0,74
LASSO3	Usai, Carta e Casu (2012)	7	Inúmeros	6	0,29
MM single SNP	Dashab et al. (2012)	5	3	3, 6, 8	0,73
MM Haplotype	Dashab et al. (2012)	5	6	2, 6, 8	0,61
MM Phylogeny	Dashab et al. (2012)	7	5	8	1,07
EMMAX	Fu et al. (2012)	3	1	3, 5, 6, 7, 8	0,43
SMS2 iter 1[a]	-	**8**	187	**0**	**0,09**
SMS2 iter 2[b]	-	**8**	158	**0**	**0,18**
SMS2 iter 3[c]	-	**8**	175	**0**	**0,11**

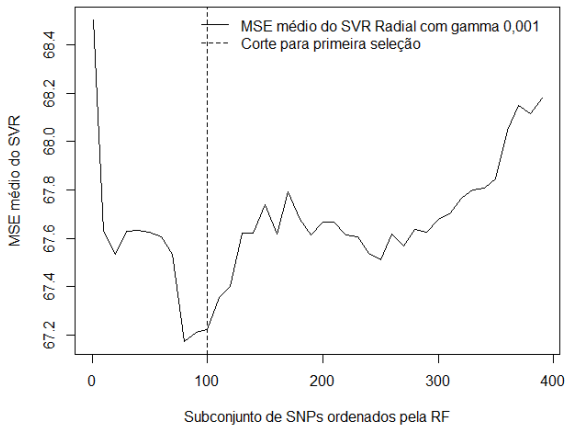
[a] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS é a união das soluções de cada *kernel* na iteração 1.
[b] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS2 é a união das soluções de cada *kernel* na iteração 2.
[c] Inserido posteriormente pelo próprio autor desse estudo para comparação do SMS2 com os 12 métodos avaliados por Demeure et al. (2012). A solução final do SMS2 é a união das soluções de cada *kernel* na iteração 3.

pois o MSE médio é praticamente uma função crescente, o que indica que à medida que seleciona-se marcadores menos relevantes, o erro predito do SVR aumenta. Com relação ao γ do *kernel* radial, somente os $\gamma = 0,01$ e $\gamma = 0,1$ são similares ao *kernel* linear, mas divergem nos primeiros marcadores do *rank*. O gráfico do MSE médio do *kernel* radial com $\gamma = 0,001$ mostrou pouca correlação com os demais, consequentemente, é o *kernel* que selecionou mais marcadores para a etapa de refinamente do GA, logo, o número de final de SNPs selecionados e o número de gerações também são os maiores (Tabelas 8.27 e 8.32). O gráfico (a) da Figura 8.15 mostra que as variações do MSE médio absoluta e relativa estão entre os limites 65 e 90 para o *kernel* linear,respectivamente, próximas de 35 e 54%. Esse fato demonstra que o efeito conjunto dos SNPs no cromossomo 1 nas dados do QTLMAS 2011 é superior aos efeitos simulados nos dados do SCRIME.

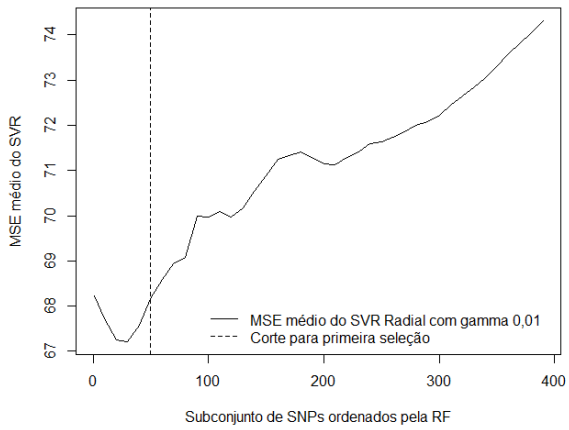
A Tabela 8.32 indica o tempo gasto em cada etapa do SMS para os quatro modelos de SVR avaliados variando somente o γ e o tempo total para construção do conjunto união. Assim, o SMS possui uma significativa restrição de custo computacional para executar a seleção em conjunto de dados com 2.000 ou mais instâncias (indivíduos), que nesses dados, foram gastos aproximadamente 12 dias para a execução dos quatro modelos do SMS e posterior união dos respectivos subconjuntos de marcadores.



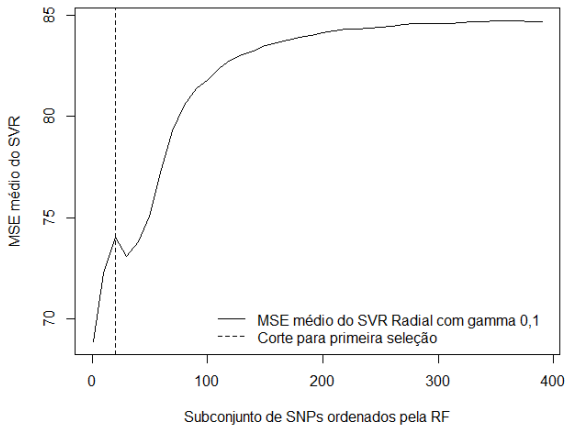
(a) *Kernel* linear.



(b) *Kernel* radial com $\gamma = 0,001$.



(c) *Kernel* radial com $\gamma = 0,01$.



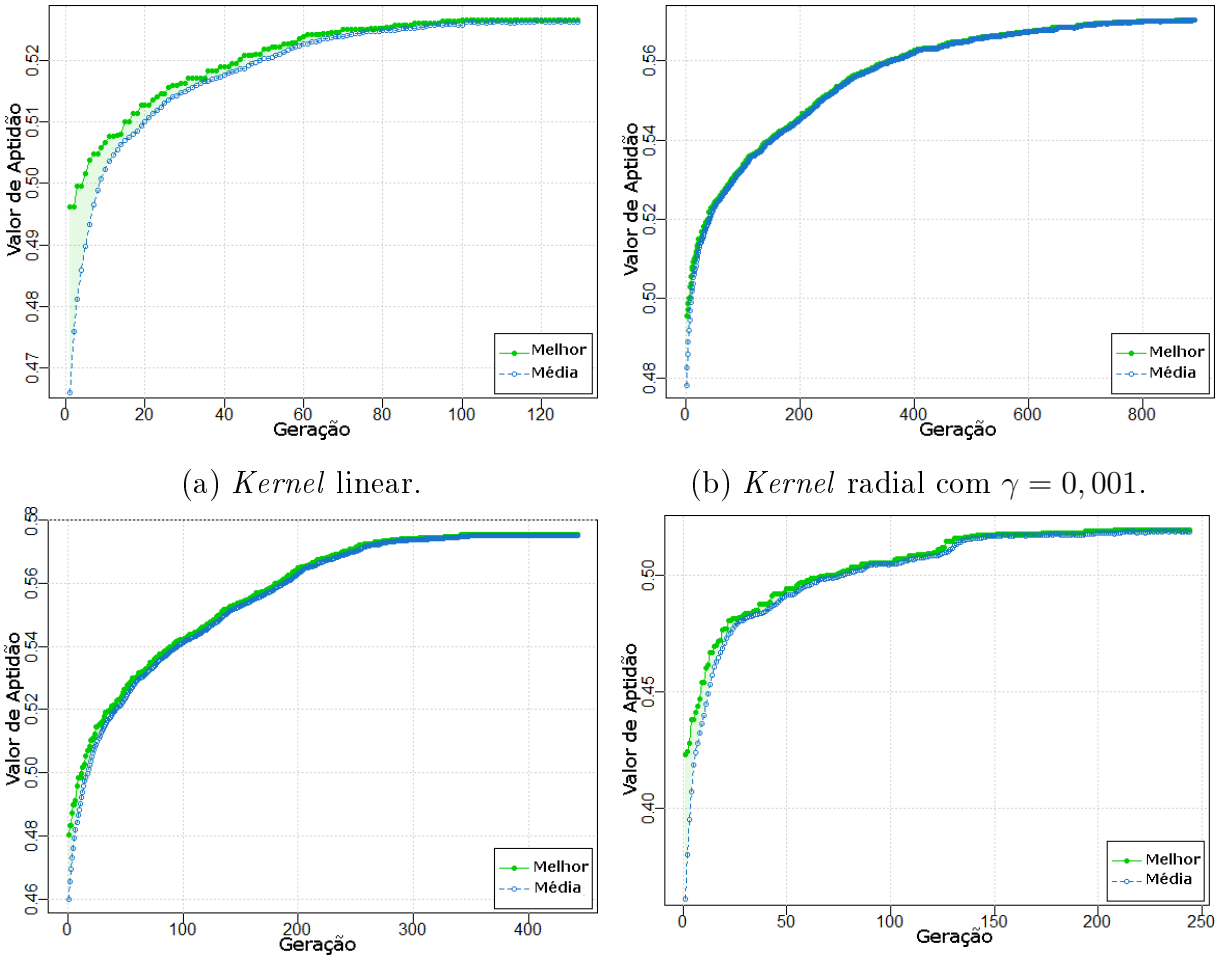
(d) *Kernel* radial com $\gamma = 0,1$.

Figura 8.15 Corte do SVR sobre o *rank* da RF no cromossomo 1 para os *kernels* linear e radial em relação à simulação do QTLMAS 2011.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação aos dados simulados do QTLMAS 2011 para a iteração 1 do SMS2.

Tabela 8.32 Número de gerações do GA, tempo de cada etapa do SMS2 para cada *kernel* avaliado e tempo total do conjunto união do SMS2 na iteração 1.

Método	# gerações do GA	Tempo Relevância (h)	Tempo Corte (h)	Tempo Refinamento (h)	Total (h)
SMS2 Linear	129	5,18	36,72	20,34	62,24
SMS2 Radial $\gamma = 0,001$	890	5,04	5,56	161,54	172,14
SMS2 Radial $\gamma = 0,01$	442	5,14	5,49	23,71	34,34
SMS2 Radial $\gamma = 0,1$	244	5,09	5,67	8,04	18,81
União SMS2 (Total)	-	-	-	-	287,53

A convergência do GA em cada *kernel* ocorreu de maneira similar, porém, com grande diferença para o número de gerações final como pode ser notado na Figura 8.16. Como explicado anteriormente, isso aconteceu pelos diferentes números de marcadores selecionados pela etapa de corte do SMS. O elevado número de gerações do GA referente ao *kernel* radial com $\gamma = 0,001$ foi ocasionado pela divergência do MSE médio predito pelo SVR com o *rank* da RF, o que gerou o maior subconjunto entre os *kernels* adotados no SMS, com 100 SNPs selecionados para a etapa de refinamento do GA.



(a) *Kernel* linear.

(b) *Kernel* radial com $\gamma = 0,001$.

(c) *Kernel* radial com $\gamma = 0,01$.

(d) *Kernel* radial com $\gamma = 0,1$.

Figura 8.16 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação aos dados do QTLMAS 2011.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação aos dados do QTLMAS 2011.

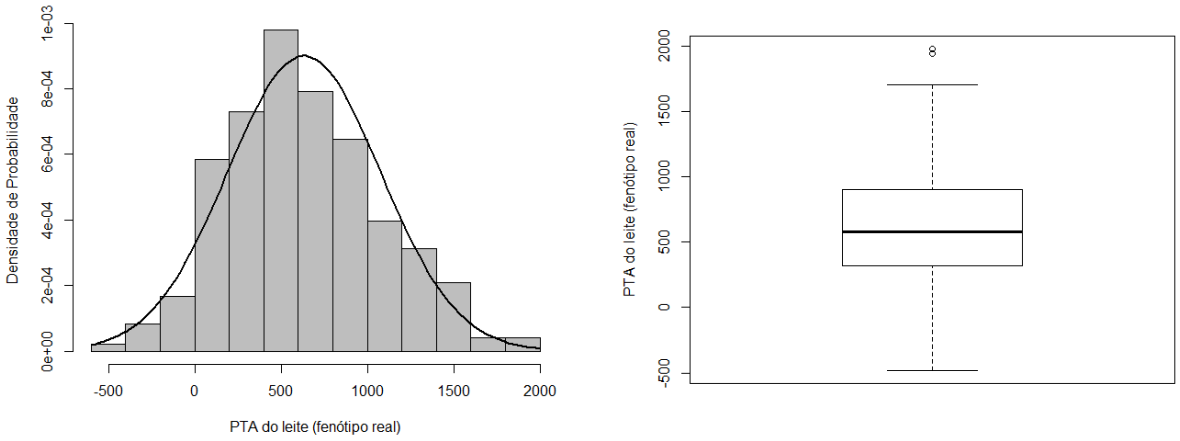
Finalmente, o SMS demonstrou que é capaz de selecionar marcadores que flanqueam QTLs com grande precisão em um cenário simulado com estrutura de LD próxima à encontrada em populações reais de indivíduos relacionados, mesmo não considerando a

informação de *pedigree* da população em seu algoritmo de treinamento. Outro benefício do SMS é a possibilidade de identificar marcadores tanto com ações aditivas quanto não-aditivas, o que mostra seu potencial na presença de interações entre SNPs, todavia, o número de marcadores falsos-positivos selecionados e o tempo consumido na sua execução precisam ser reduzidos para torná-lo mais competitivo frente aos métodos de seleção mais usados atualmente.

8.10 Conjunto de Dados Reais

8.10.1 Resultados do SMS

O fenótipo medido pela PTA do leite demonstrou um amplo espectro de variação como pode ser observado na Figura 8.17. O teste de normalidade de Shapiro-Wilk foi aplicado aos dados amostrais da PTA do leite e concluiu-se que existem evidências de que esse fenótipo não possui distribuição normal, pois o valor-p do teste foi igual a 0,03135, sendo menor que $\alpha = 0,05$. A não-normalidade desse fenótipo indicada pelo teste de Shapiro-Wilk ocorreu pelo tamanho reduzido da amostra juntamente com a ocorrência de 2 valores aberrantes identificados pelo *boxplot*, pois em amostras maiores ou iguais a 1.000 touros, a PTA do leite possui distribuição normal.



(a) Densidade de probabilidade do fenótipo PTA do leite.

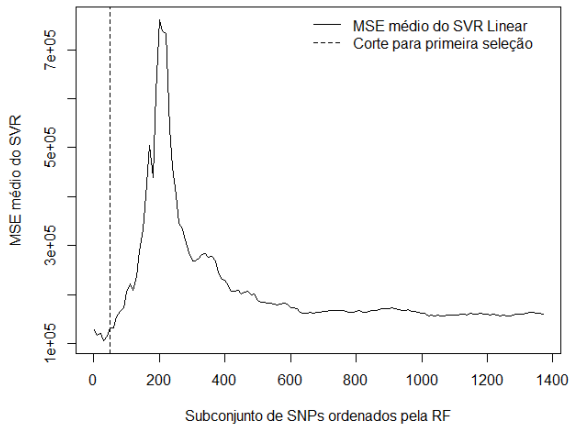
(b) *Boxplot* do fenótipo PTA do leite.

Figura 8.17 Densidade de probabilidade da PTA do leite juntamente com a distribuição normal com média 633,30 (média da PTA do leite) e desvio-padrão 443.09 (desvio-padrão da PTA do leite, e *Boxplot* da PTA do leite.

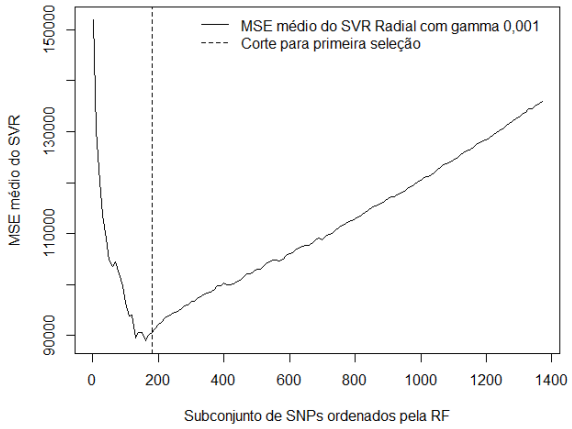
Ao final da execução do SMS com quatro modelos de SVR, baseados nos *kernels* linear e radial, onde neste o γ assumiu os valores 0,001, 0,01 e 0,1. Cada modelo foi executado uma única vez ($iter = 1$) e a Tabela 8.33 mostra o número de SNPs encontrado por cada modelo.

Pela Figura 8.18, nota-se que o SVR sobre o *rank* da RF no cromossomo 1 apresenta comportamento distinto para os *kernels* linear e radial, sendo que no último, sofre também mudanças em função do γ, demonstrando que deve-se escolher um γ adequado

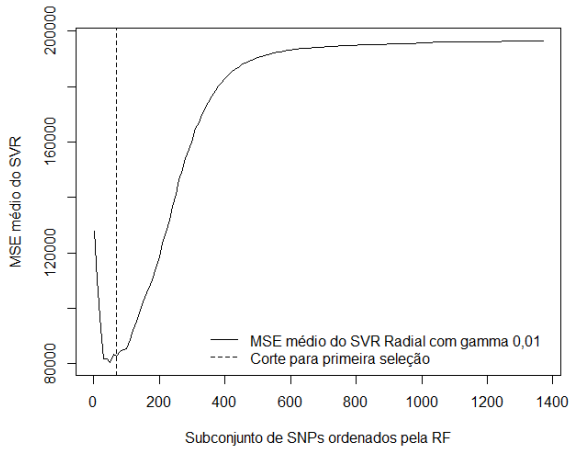
para realizar uma melhor seleção de SNPs. Essa variação é observado em todos os 30 cromossomos que compõem o genoma do *Bos tauros*. Pelo gráfico (a) da Figura 8.18, verifica-se que as variações absoluta e relativa do *kernel* linear são respectivamente próximas de 600.000 e 600%, considerando os limites inferior e superior iguais a 100.000 e 700.000, respectivamente. Essa observação demonstra que o efeito conjunto dos SNPs no cromossomo 1 é muito superior aos efeitos simulados nos dados do SCRIME e do QTLMAS 2011.



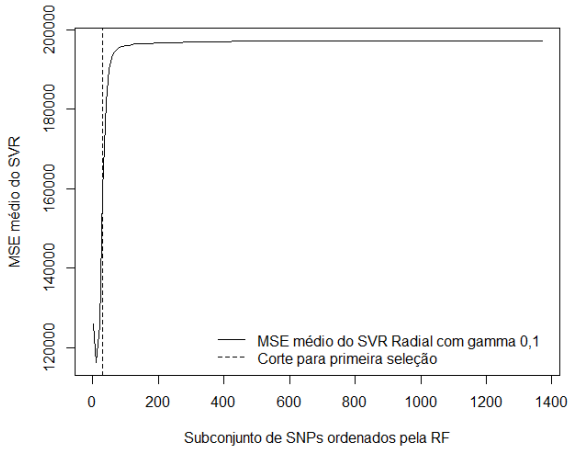
(a) *Kernel* linear.



(b) *Kernel* radial com $\gamma = 0,001$.



(c) *Kernel* radial com $\gamma = 0,01$.

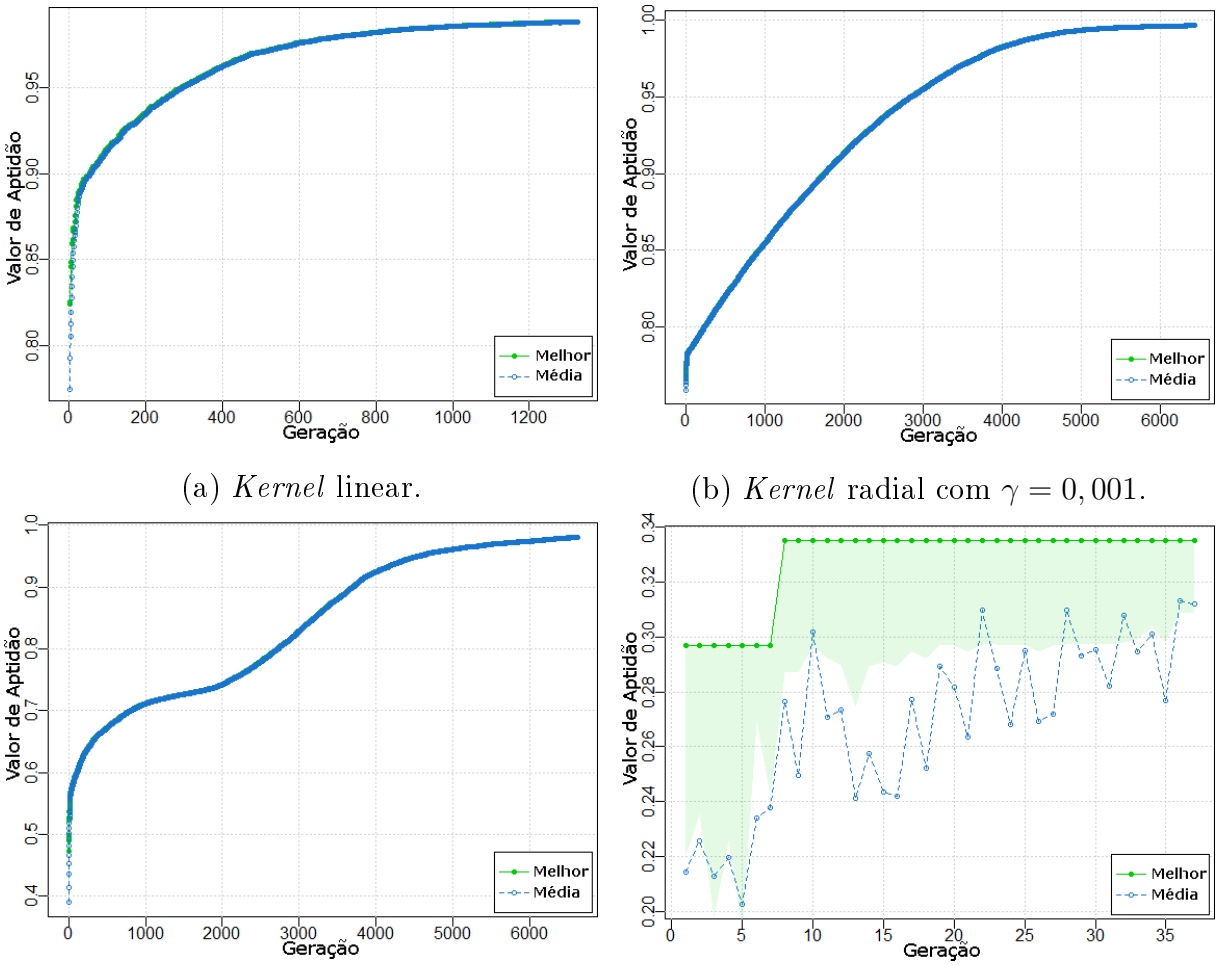


(d) *Kernel* Radial com $\gamma = 0,1$.

Figura 8.18 Corte do SVR sobre o *rank* da RF no cromossomo 1 para os *kernels* linear e radial em relação aos dados reais.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação ao SMS2 para os dados reais. A linha tracejada indica o ponto de corte.

O GA demonstrou comportamento similiar ao ocorrido com os dados simulados do

QTLMAS 2011 conforme a Figura 8.19. Cabe destacar que o número de gerações foi superior aos computados no dados do QTLMAS 2011 e apresentou variação significativa entre os *kernels* avaliados. Outra questão de destaque é que somente no radial com $\gamma = 0,1$, a média da aptidão da população ao longo das gerações foi nitidamente inferior ao melhor indivíduo do GA (melhor subconjunto de SNPs), onde nos outros três *kernels* essa diferença é pequena como pode ser observada na Figura 8.19.



(a) *Kernel* linear.

(b) *Kernel* radial com $\gamma = 0,001$.

(c) *Kernel* radial com $\gamma = 0,01$.

(d) *Kernel* radial com $\gamma = 0,1$.

Figura 8.19 Convergência da aptidão (correlação média em 10-*fold*) do melhor subconjunto de SNPs e da aptidão média da população ao longo das gerações para os *kernels* linear e radial em relação aos dados reais.
Kernels (a) linear, (b) radial com $\gamma = 0,001$, (c) radial com $\gamma = 0,01$ e (d) radial com $\gamma = 0,1$ em relação ao SMS2 para os dados reais.

A Tabela 8.33 mostra que o SMS2 com *kernel* radial e $\gamma = 0,001$ selecionou o maior número de marcadores nas etapas de corte (primeira seleção) e refinamento (segunda seleção), entretanto, o *kernel* radial com $\gamma = 0,1$ teve o menor número de SNPs selecionados no corte, mas a segunda maior quantidade de marcadores selecionada no

refinamento pelo GA. É importante destacar que o número de SNPs selecionados na etapa de corte é o somatório dos SNPs selecionados em cada cromossomo, por exemplo, no *kernel* linear, foram selecionados 1.100 SNPs para os 30 cromossomos (segunda linha e segunda coluna da Tabela 8.33). Isso mostra que a composição de quatro *kernels* distintos permite selecionar marcadores que possuem informações diferentes. O número total de SNPs foi 1.584, porém, a união dos quatro subconjuntos construídos por cada *kernel* adotado no SMS possui 1.265 marcadores, pois alguns SNPs foram selecionados por mais de um *kernel*.

Tabela 8.33 Número de SNPs total no conjunto inicial e número de SNPs selecionados nas duas etapas de seleção do SMS2 para o *kernel* linear e para o *kernel* radial variando-se os γs.

Etapas do SMS2	Kernel			
	Linear	$\gamma = 0,001$	$\gamma = 0,01$	$\gamma = 0,1$
Dados de entrada	22.844	22.844	22.844	22.844
Relevância + corte (RF + SVR)	1.100	4.881	2.640	870
Refinamento (GA + SVR)	294	640	253	397

O *kernel* que consumiu o maior tempo de processamento no SMS foi o radial com $\gamma = 0,001$ com aproximadamente 92 horas pela Tabela 8.34, e o *kernel* com o menor tempo foi o radial com $\gamma = 0,1$, que gastou 3,82 horas conforme 8.34. O tempo total para SMS referente aos quatro *kernels* totalizou 175,02 horas, ou seja, em torno de sete dias.

Tabela 8.34 Tempo do SMS2 por etapa e por *kernel*.

Método	Tempo por etapa do SMS2 (min.)			Total (horas)
	Relevância	Corte	Refinamento	
SMS2 Linear	42,57	531,75	1.050,89	27,09
SMS2 Radial $\gamma = 0,001$	24,89	188,81	5.330,02	92,40
SMS2 Radial $\gamma = 0,01$	24,57	191,59	2.886,67	51,71
SMS2 Radial $\gamma = 0,1$	24,18	194,58	10,50	3,82

O Blasso não conseguiu convergir para uma solução factível com base nos dados reais, pois o seu λ final convergiu para zero e vários coeficientes tiveram variâncias que divergiram para o infinito. Assim, outros parâmetros para as distribuições *a priori* e outras distribuições *a priori* (Beta e Gamma) foram testados na tentativa de resolução desse problema, mas nenhum conjunto de parâmetros com suas respectivas distribuições *a priori* obtiveram êxito, apresentando a mesmo problema inicial.

O métodos valor-p bruto e o valor-p corrigido selecionaram 6.428 e 130 SNPs

respectivamente. Como dito anteriormente, a solução do SMS2 baseada na união dos melhores subconjuntos dos quatro *kernels* adotados selecionou 1.265 SNPs distintos. O SMS1 (primeira versão do SMS) selecionou 3.357 marcadores com o PUK, porém, pode-se observar na Tabela 8.35, que o desempenho da solução desse *kernel*, com correlação média igual a 0,82 e desvio-padrão 0,08, foi inferior em relação às soluções dos *kernels* radiais do SMS2 com $\gamma = 0,001$, $\gamma = 0,01$ e $\gamma = 0,1$, cuja maior correlação média foi igual a 0,997 com desvio-padrão 0,003 para o radial com $\gamma = 0,001$. Essa significativa diferença entre as correlações de Pearson médias pode ser explicada pela maior eficiência do *rank* da RF (SMS2) em relação ao do valor-p bruto da correlação de Spearman (SMS1).

Foi realizado um comparativo do ponto de vista estatístico entre os conjuntos de SNPs selecionados pelos métodos SMS, valor-p bruto e valor-p corrigido a fim de mostrar que a solução do SMS demonstra maior acurácia para a predição do fenótipo com somente a utilização dos marcadores considerados informativos. Assim, os três conjuntos de SNPs foram avaliados a partir de uma validação cruzada com 10-*fold* pelo SVR com *kernels* linear e radial com γ assumindo os valores 0,001, 0,01 e 0,1. Como pode ser notado na Tabela 8.35, as melhores predições, do ponto de vista da correlação média, ocorreram com o *kernel* linear ($\overline{r} = 0,988$) e com o *kernel* radial, obtendo $\overline{r} = 0,997$ e $\overline{r} = 0,980$, respectivamente, referentes ao $\gamma = 0,001$ e $\gamma = 0,01$. Entretanto, cabe ressaltar que ocorreu uma divergência em relação à medida de erro predito MSE, pois os MSEs da solução do radial com $\gamma = 0,001$ e $\gamma = 0,01$ demonstraram médias muito superiores ao do linear. Essa alta correlação mostra que fenótipos com valores elevados tiveram predições elevadas e fenótipos com valores menores, predições menores, entretanto, a distância entre os valores observados e preditos é substancial. O mesmo comportamento foi observado em relação à métrica do MAPE.

Adotou-se a estratégia de utilizar os subconjuntos de SNPs selecionados pelo SMS2 (união), pelo valor-p bruto e pelo valor-p corrigido para verificar se seria possível o uso do Blasso como avaliador desses subconjuntos de SNPs, já que o mesmo não pode ser usado como método de seleção porque não ocorreu a convergência das variâncias dos coeficientes dos SNPs. A Tabela 8.36 mostra que o subconjunto de SNPs selecionado pela união do SMS2 apresentou o melhor resultado para a correlação de Pearson média ($\overline{r}$), MSE médio ($\overline{mse}$) e MAPE médio ($\overline{mape}$). Em relação aos desvios-padrões, a solução da união do SMS2 demonstrou maior desvio-padrão para o MSE e a solução do valor-p corrigido, o

Tabela 8.35 Avaliação dos subconjuntos de marcadores gerados a partir dos métodos SMS1 e SMS2, valor-p bruto e valor-p corrigido, com o uso do SVR com *kernels* linear, radial e PUK.

Método Seleção	Método Avaliação	# SNPs	$\overline{r}$	σ_r	$\overline{mse}$	σ_{mse}	$\overline{mape}$	σ_{mape}
SMS1 Linear[a]	Linear	-[b]	0,820	0,08	-[b]	-[b]	-[b]	-[b]
SMS1 Radial[a]	Radial	-[b]	0,440	0,25	-[b]	-[b]	-[b]	-[b]
SMS1 PUK[a]	PUK	3.357	0,810	0,08	-[b]	-[b]	-[b]	-[b]
SMS2 Linear	Linear	294	0,988	0,005	**8.215**	4.918	**63**	67
SMS2 Radial	Radial $\gamma = 0,001$	640	**0,997**	0,003	26.336	20.714	125	125
SMS2 Radial	Radial $\gamma = 0,01$	253	0,980	0,007	113.301	55.738	264	224
SMS2 Radial	Radial $\gamma = 0,1$	397	0,335	0,238	197.074	80.417	329	267
Valor-p bruto	Linear	6.428	0,811	0,079	68.488	26.187	187	184
	Radial $\gamma = 0,001$	6.428	-[c]	-[c]	197.074	80.418	329	267
	Radial $\gamma = 0,01$	6.428	-[c]	-[c]	197.074	80.418	329	267
	Radial $\gamma = 0,1$	6.428	-[c]	-[c]	197.074	80.418	329	267
Valor-p corrigido	Linear	130	0,279	0,270	197.066	80.416	0,28	0,27
	Radial $\gamma = 0,001$	130	0,279	0,270	197.066	80.416	0,28	0,27
	Radial $\gamma = 0,01$	130	0,279	0,270	197.066	80.416	0,28	0,27
	Radial $\gamma = 0,1$	130	0,279	0,270	197.066	80.416	0,28	0,27

[a] Resultados extraídos de Oliveira et al. (2014b).
[b] Esses valores não foram computados no estudo de Oliveira et al. (2014b).
[c] As correlações médias $\overline{r}$ não foram calculadas devido às predições do SVR de todas as dez partições serem iguais, gerando desvio-padrão igual a zero, o que inviabiliza o cômputo da correlação de Pearson.

menor. Com relação à correlação de Pearson e ao MAPE, a solução do SMS2 teve a maior média e o menor desvio-padrão, indicando sua superioridade na predição dos valores da PTA do leite.

Tabela 8.36 Avaliação dos subconjuntos de SNPs pelo Blasso usando validação cruzada com 10-*fold*, os quais foram selecionados pelos métodos SMS2 (união), valor-p bruto e valor-p corrigido.

Método	# SNPs	$\overline{r}$	σ_r	$\overline{mse}$	σ_{mse}	$\overline{mape}$	σ_{mape}
União SMS2	1.265	**0,932**	**0,031**	**524.103**	389.884,3	**322,91**	**189,10**
Valor-p bruto	6.428	0,801	0,077	1.016.303	380.846,8	400,44	224,49
Valor-p corrigido	130	0,760	0,095	879.634	**172.382,6**	394,93	225,34

Uma importante conclusão é que o Blasso conseguiu ser executado a partir do subconjunto de SNPs selecionado pelo SMS com 1.265 marcadores. Isso mostra que o processo de seleção do SMS superou a dificuldade encontrada pelo algoritmo de seleção do Blasso, ou seja, o SMS eliminou os SNPs que produziram obstáculos para a convergência das variâncias dos coeficientes do Blasso. Portanto, em conjuntos de dados onde o Blasso não converge pode-se usar o SMS como método de seleção e, posteriormente, o Blasso como método de avaliação e predição.

8.10.2 Validação dos Resultados

Na literatura, encontraram-se duas abordagens para descoberta de regiões promissoras no genoma usadas em GWAS: a primeira, mais restritiva, que indica somente o QTL e/ou o gene mais próximo ao SNP selecionado como realizada por Jiang et al. (2010) e Frąszczak e Szyda (2015) e a segunda, que define um raio em torno do SNP selecionado para construir uma vizinhança que contenha QTLs e/ou genes como feito por Mokry et al. (2013). Desta forma, no presente trabalho, deu-se preferência à segunda abordagem, pois consegue-se estipular dois pontos de corte, um acima e outro abaixo da posição do SNP, flexibilizando a possibilidade de encontrar mais de um QTL e/ou gene em regiões que podem ter papel fundamental na determinação do fenótipo considerado. Ressalta-se o fato de que o gene mais próximo ao SNP não necessariamente é o que contém a mutação causal para o fenótipo, logo a escolha pela vizinhança é mais abrangente, aumentando a possibilidade de novas descobertas. Caso a segunda abordagem, que é usada neste estudo, identifique muitas regiões candidatas, pode-se restringir o comprimento do raio para, por exemplo, 10.000 pb diminuindo o número de genes candidatos, ou, no caso extremo, adotar raio igual a zero pares de bases, avaliando somente os SNPs que estejam dentro de algum gene conhecido.

A metodologia empregada para buscar regiões relevantes no genoma do *Bos taurus* para a PTA do leite, a partir dos marcadores do tipo SNP selecionados pelo SMS, foi a mesma empregada por Mokry et al. (2013). Nesse estudo, a metodologia de GWAS foi empregada para selecionar um subconjunto de SNPs associados à qualidade da carne de gado Canchim, medida pela espessura do toucinho. Os autores consideraram a extensão do LD baseado na média geral do r^2 ($\overline{r^2} = 0,12$ em uma distância de 250.000 pb). Com isso, uma janela de 500.000 pb em torno de cada SNP previamente selecionado por um processo de seleção de atributos, baseado em uma regressão com *Stepwise*, foi usada para definir as regiões para a descoberta de genes candidatos.

O ponto de corte usado para verificar se determinado QTL ou gene está marcado por algum SNP selecionado pelo SMS, foi baseado num raio de 250.000 pb, pois esse valor é próximo da distância média dos 30 cromossomos mostrada na última linha da terceira coluna da Tabela 8.37, com a correspondente extensão do LD dada pela média geral r^2 ($\overline{r^2}$), que é igual a 0,094 para todos os 30 cromossomos. É importante destacar a proximidade dos resultados de LD e distância entre pares de SNPs encontrados no presente

trabalho e no artigo de Mokry et al. (2013), apesar do primeiro possuir 22.844 SNPs e o segundo, 708.641, após o controle de qualidade. A relação entre a distância e o LD médios entre SNPs também é próxima à encontrada no estudo de Qanbari et al. (2010) sobre o padrão de LD em gado Holandês alemão, onde o intervalo entre 0,2 e 0,5 cM (intervalo aproximado entre 200.000 e 500.000 pb no mapa físico) possui pares de SNPs com r^2 médio igual a 0,09, considerando que em média 1 cM é equivalente a 1.000.000 pb (QANBARI et al., 2010).

As estimativas de distância e LD foram calculadas pelo *software* Haploview (BARRETT et al., 2005) com parâmetros *default*. Os cálculos de distância física média e LD médio entre os marcadores, computados pelas métricas r^2 e D', estão ilustrados na Tabela 8.37. Note que somente no cromossomo X, essas estimativas apresentam discrepâncias com os demais cromossomos e uma possível explicação para isso é que nesse cromossomo ocorreu uma maior eliminação de SNPs após o controle de qualidade quando comparado aos demais.

O mapa físico das posições dos marcadores foi construído usando o genoma de referência *Bos taurus* UMD 3.1.1. A pesquisa por QTLs e por genes foi realizada no *Cattle Genome Browser* Genome (2015), cujo endereço eletrônico é <http://www.animalgenome.org/cgi-bin/gbrowse/bovine/#search>, com os seguintes filtros selecionados: *Annotated Genes*, *Milk Traits - Association* e *Milk Traits - QTL*, onde o primeiro é referente aos genes anotados e os outros dois, aos QTLs associados ao leite. As informações sobre os genes encontrados foram buscadas no banco de dados do NCBI BioSystems (GEER et al., 2009) com o ID do gene encontrado no *browser* de Genome (2015), enquanto que as referências dos QTLs marcados foram indicadas pelo próprio *Cattle Genome Browser*.

8.10.2.1 QTLs Identificados pelo SMS

A Tabela 8.51 demonstra a seleção de QTLs para cada *kernel* usado no SMS. O maior número de QTLs identificados foi referente ao *kernel* radial com $\gamma = 0,001$, o qual encontrou 229 QTLs, mas isso ocorreu proque esse *kernel* foi o que selecionou o maior número de SNPs. O *kernel* linear identificou 43 QTLs, a menor quantidade dentro os quatro *kernels* avaliados. O cromossomos 6 foi o que apresentou o maior número de QTLs identificados pelos quatro *kernels*, totalizando 165 QTLs, enquanto que os cromossomos

Tabela 8.37 Desequilíbrio de ligação computado pelas medidas r^2 e D' e distância média em pares-base entre SNPs por cromossomo.

Cromossomo	Número de SNPs após CQ	Distância Média (pb)	r^2 médio	D' médio
1	1.450	254.862,94	0,101	0,571
2	1.177	251.442,07	0,109	0,602
3	1.089	249.860,06	0,098	0,559
4	1.030	251.394,37	0,100	0,567
5	886	251.544,17	0,118	0,589
6	1.236	250.993,44	0,019	0,298
7	1.010	246.763,10	0,111	0,591
8	1.028	250.154,66	0,098	0,576
9	914	250.410,22	0,117	0,617
10	919	249.185,82	0,104	0,571
11	945	251.650,82	0,094	0,545
12	709	247.901,75	0,099	0,598
13	726	251.312,74	0,106	0,588
14	778	245.163,02	0,108	0,586
15	746	250.720,06	0,094	0,558
16	764	247.807,13	0,094	0,543
17	741	254.389,61	0,089	0,556
18	603	253.113,67	0,108	0,563
19	545	252.248,95	0,096	0,592
20	707	248.635,33	0,102	0,570
21	613	250.304,58	0,102	0,572
22	612	251.189,02	0,097	0,563
23	499	248.382,80	0,082	0,537
24	570	253.343,51	0,100	0,587
25	427	246.470,00	0,079	0,536
26	479	250.490,33	0,087	0,546
27	427	250.603,26	0,086	0,555
28	429	249.495,89	0,078	0,537
29	448	251.807,62	0,102	0,586
X	50	224.145,48	0,039	0,366
99	287	-	-	-
Total	22.844	249.526,21	0,094	0,554

7, 10, 12, 17 e 21 não identificaram QTL algum.

O SNP que marcou o maior número de QTLs associados a características do leite foi o ARS-BFGL-NGS-53485 (cromossomo 6 e posição 70.612 pb), pois o mesmo mostrou associação com 56 QTLs e cinco genes, a partir de uma vizinhança com diâmetro de 500.000 pb ($SNP \pm 250.000$ pb). Outra observação importante associada a esse marcador, é que seu valor-p bruto é 9,82E-02 e seu valor-p corrigido é 2,24E+03, ambos menores que o valor de corte 0,05, ou seja, esse marcador não seria selecionado pelos métodos monoatributos baseados em regressão linear simples mas comumente usados em GWAS. O segundo marcador em relação ao número de QTLs identificados foi o ARS-RFGL-

Tabela 8.38 QTLs identificados pelos 1.265 SNPs selecionados pelo SMS agrupados por cromossomo para cada *kernel*.

Cr.	Kernel				**Total**
	Linear	Radial $\gamma = 0,001$	Radial $\gamma = 0,01$	Radial $\gamma = 0,1$	
1	4	7	4	3	**18**
2	0	0	1	0	**1**
3	0	4	0	2	**6**
4	1	9	1	1	**12**
5	5	31	6	11	**53**
6	18	91	25	31	**165**
7	0	0	0	0	**0**
8	0	2	0	2	**4**
9	0	3	0	0	**3**
10	0	0	0	0	**0**
11	6	11	5	4	**26**
12	0	0	0	0	**0**
13	1	4	5	2	**12**
14	2	6	1	2	**11**
15	0	0	1	0	**1**
16	0	0	1	0	**1**
17	0	0	0	0	**0**
18	0	5	0	1	**6**
19	1	5	4	1	**11**
20	1	10	3	1	**15**
21	0	0	0	0	**0**
22	1	1	0	0	**2**
23	1	4	4	3	**12**
24	0	12	4	1	**17**
25	0	3	2	0	**5**
26	0	7	1	0	**8**
27	0	0	0	2	**2**
28	0	6	5	2	**13**
29	1	3	1	1	**6**
X	1	5	2	2	**10**
Total	**43**	**229**	**76**	**72**	**420**

NGS-26008 (cromossomo 5 e posição 123.920 pb) com 22 QTLs e 3 genes, porém, seu valor-p bruto é 9,82E-02 e seu valor-p corrigido, 2,24E+03, isto é, esse marcador também não foi selecionado pelos dois métodos baseados no *rank* do valor-p. Isso sugere que o SMS permite capturar marcadores que flanqueam regiões genômicas relevantes para a característica em estudo, mas que estatisticamente não são significativas.

Um exemplo de pesquisa feita pelo buscador *Cattle Genome Browser* (<http://www.animalgenome.org/cgi-bin/gbrowse/bovine/#search>) é mostrado na Figura 8.20, onde o diâmetro de 500 kb em relação ao marcador Hapmap41190-BTA-16578 indica cinco QTLs associados a essa região a saber: percentagem de alfa-caseína no leite (valor-p

$1,00 \times 10^{-9}$), percentagem de beta-caseína no leite (valor-p $7,940 \times 10^{-10}$), percentagem de beta-lactoglobulina no leite (valor-p $5,89 \times 10^{-15}$), indice de caseína no leite (valor-p $4,27 \times 10^{-13}$) e percentagem de kappa-caseína no leite (valor-p $1,23 \times 10^{-8}$). O método usado para identificar esses QTLs e os respectivos valores-p entre parêntesis encontrados estão descritos detalhadamente em Schopen et al. (2011).

Além desses QTLs, foram identificados 10 genes: LOC100337445 (ZIMIN et al., 2009), DGUOK (SONSTEGARD et al., 2002), ACTG2 (VANDEKERCKHOVE; WEBER, 1979; ISHIWATA et al., 2003; HARHAY et al., 2005) , STAMBP (ZIMIN et al., 2009), LOC789284 (ZIMIN et al., 2009), MIR2295 (GLAZOV et al., 2009; GRIFFITHS-JONES et al., 2006), DUSP11 (SONSTEGARD et al., 2002; HARHAY et al., 2005; ZIMIN et al., 2009), TPRKB (ZIMIN et al., 2009), LOC789265 e ALMS1 (ZIMIN et al., 2009). Esses genes poderão ser analisados posteriormente para verificar a possível relação dos mesmos com a produção de leite.

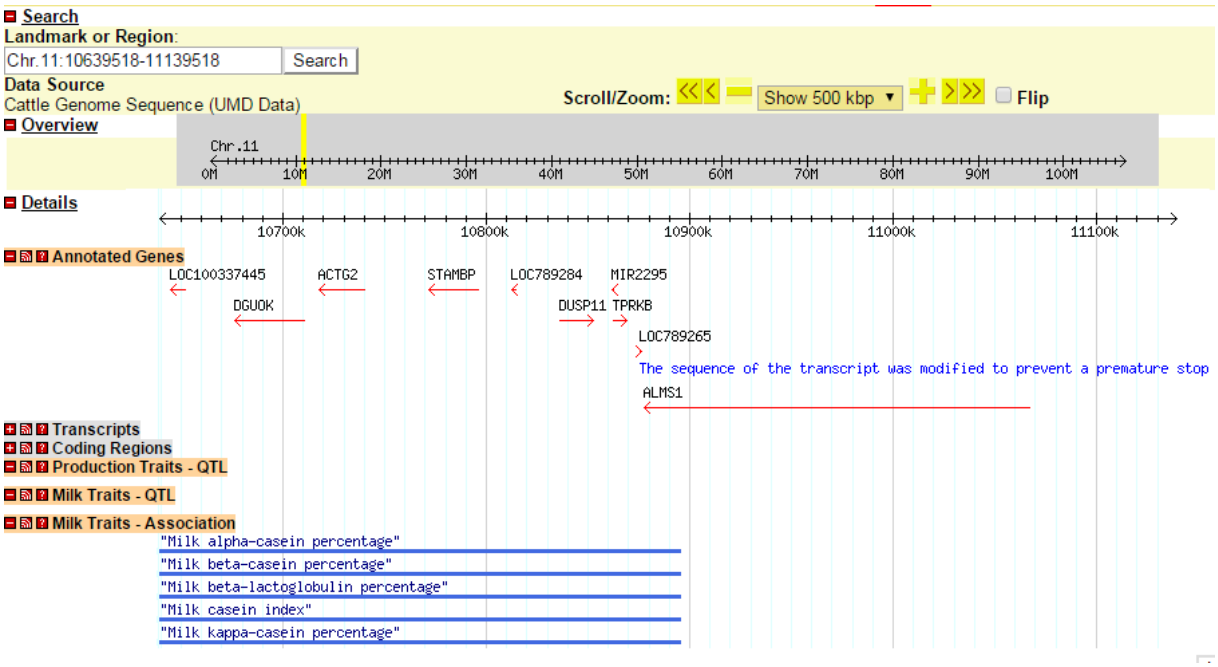


Figura 8.20 Exemplo da saída de uma pesquisa feita no buscador *animalgenome*.

A saída foi baseada na região flanqueada pelo marcador Hapmap41190-BTA-16578 cuja posição é 10.889.518 pb no cromossomo 11 com 500 kb de diâmetro.

A solução do SMS, produzida pela união dos quatro subconjuntos de SNPs selecionados a partir dos *kernels* linear e radial, selecionou 1.265 marcadores SNP, os quais marcam um total de 420 QTLs, agrupados em 54 categorias conforme Tabela 8.39, onde 245 QTLs são distintos, pois alguns marcadores selecionaram o mesmo QTL. Enquanto que a solução

do valor-p corrigido indicou um total de 45 QTLs, onde 35 são diferentes, porque alguns marcadores indicaram o mesmo QTL. Isso demonstra que o SMS foi capaz de encontrar seis vezes mais QTLs associados ao leite do que o método do valor-p corrigido sem considerar que 21 marcadores identificados pelo valor-p corrigido também foram selecionados pelo SMS.

Dos 1.265 marcadores selecionados pelo SMS, 104 não identificaram gene ou QTL associado ao leite, o que representa 8,22% de regiões não-informativas. Por outro lado, dos 130 SNPs selecionados pelo valor-p corrigido, 14 deles marcam regiões que não possuem genes e nem QTLs associados ao leite, o que significa 10,77% referentes a regiões sem informação. Entretanto, essas observações não significam a ausência total de genes ou QTLs na região, pois basta aumentar o raio de abrangência do SNP considerado, para 500.000 pb por exemplo, para possibilitar a ocorrência de novas informações genômicas.

Com o objetivo de evidenciar o percentual de identificação do SMS com relação a todos os QTLs associados ao leite conhecidos atualmente na base de dados do Genome (2015), computou-se o total de 1.026 QTLs referentes ao leite por meio de filtros, onde 245 QTLs foram identificados pelo SMS e 35, pelo valor-p corrigido, com a janela de 500.000 pb adotada para cada SNP. Portanto, aproximadamente 24% dos QTLs conhecidos foram mapeados pelo SMS, o que é um resultado promissor dado que essa seleção de SNPs foi realizada com somente 240 indivíduos (touros Gir). Entretanto, somente 3% do total de QTLs de leite foram marcados pelos SNPs selecionados pelo valor-p corrigido, além disso, 21 SNPs dos 35 encontrados pelo valor-p corrigido também foram selecionados pelo SMS, o que parece mostrar que o SMS seleciona um subconjunto mais geral do que o valor-p corrigido.

A Tabela 8.39 permite inferir que o SMS, baseado na característica da PTA do leite, capturou outras características do leite tais como teor e percentagem de alfa-caseína e de beta-caseína, teor de lactose, percentagem de proteína e gorduna entre outros. Esse fato evidencia que várias regiões cromossômicas não afetam somente a quantidade produzida de leite, mas também sua composição química.

A seleção baseada no valor-p corrigido por Bonferroni menor que 0,05 selecionou 10 SNPs que flanquearam 35 QTLs distintos do leite demonstrados na Tabela 8.40, pois os QTLs de ID 20624, 20618 e 20619 foram selecionados por três marcadores distintos: ARS-BFGL-NGS-16315, ARS-BFGL-NGS-72551 e ARS-BFGL-NGS-8796, que

Tabela 8.39 Descrição dos QTLs do leite flanqueados pelos SNPs selecionados pela união do SMS separados por categorias.

Categoria	Descrição original do QTL do Leite	Descrição traduzida do QTL do Leite	Total
1	*305-day milk yield*	Produção de leite em 305 dias	7
2	*Average daily milk yield*	Produção de leite média diária	3
3	*Curd firming rate*	Taxa de endurecimento da coalhada	1
4	*Milk alpha-casein content*	Teor de alfa-caseína no leite	18
5	*Milk alpha-casein percentage*	Percentagem de alfa-caseína no leite	6
6	*Milk alpha-casein to beta-casein ratio*	Razão entre alfa-caseína e beta caseína	17
7	*Milk alpha-lactalbumin content*	Teor de alfa-lactalbumina	1
8	*Milk alpha-lactalbumin percentage*	Percentagem de alfa-lactalbulmina	11
9	*Milk beta-casein content*	Teor de beta-caseína no leite	18
10	*Milk beta-casein percentage*	Percentagem de beta caseína no leite	1
11	*Milk beta-lactoglobulin percentage*	Percentagem beta-lactoglobulina no leite	3
12	*Milk beta-lactoglobulin protein content*	Teor de proteína beta-lactoglobulina	19
13	*Milk casein index*	Índice de caseína de leite	2
14	*Milk casein percentage*	Percentagem de caseína no leite	1
15	*Milk energy yield*	Rendimento energético no leite	3
16	*Milk fat-to-protein ratio*	Razão entre gorduta e proteína	7
17	*Milk fat percentage*	Percentagem de gordura no leite	21
18	*Milk fat percentage (daughter deviation)*	Percentagem de gordura no leite (desvio da filha)	5
19	*Milk fat percentage (EBV)*	Percentagem de gordura no leite (EBV)	21
20	*Milk fat percentage (PTA)*	Percentagem de gordura no leite (PTA)	1
21	*Milk fat yield*	Produção de gordura no leite	17
22	*Milk fat yield (daughter deviation)*	Produção de gordura no leite (desvio da filha)	13
23	*Milk fat yield (EBV)*	Produção de gordura no leite (EBV)	5
24	*Milk fat yield (PTA)*	Produção de gordura no leite (PTA)	1
25	*Milk fatty acid unsaturated index*	Índice de ácido graxo insaturado	4
26	*Milk kappa-casein content*	Teor de kappa-caseína no leite	18
27	*Milk kappa-casein percentage*	Percentage de kappa-caseína no leite	7
28	*Milk lactose content*	Teor de lactose no leite	1
29	*Milk lactose yield*	Produção de lactose no leite	1
30	*Milk monounsaturated fatty acid content*	Teor de ácido graxo monoinsaturado no leite	2
31	*Milk myristic acid percentage*	Percentagem de ácido mirístico no leite	11
32	*Milk myristoleic acid percentage*	Porcentagem de ácido miristoleico no leite	1
33	*Milk palmitoleic acid percentage*	Porcentagem de ácido palmitoléico no leite	1
34	*Milk protein content*	Teor de proteína no leite	17
35	*Milk protein percentage*	Percentagem de proteína no leite	44
36	*Milk protein percentage (daughter deviation)*	Percentagem de proteína no leite (desvio da filha)	5
37	*Milk protein percentage (EBV)*	Percentagem de proteína no leite (EBV)	4
38	*Milk protein percentage (PTA)*	Percentagem de proteína no leite (PTA)	1
39	*Milk protein yield*	Produção de proteína no leite	31
40	*Milk protein yield (daughter deviation)*	Produção de proteína no leite (desvio da filha)	12
41	*Milk protein yield (EBV)*	Produção de proteína no leite (EBV)	4
42	*Milk rennet clotting time*	Tempo de coagulação do coalho no leite	3
43	*Milk saturated fatty acid content*	Teor de ácido graxo saturado no leite	2
44	*Milk saturated to unsaturated fatty acid ratio*	Razão de ácido graxo saturado e ácido graxo insaturado no leite	2
45	*Milk solids*	Sólidos no leite	1
46	*Milk solids percentage*	Percentagem de sólidos no leite	1
47	*Milk stearic acid percentage*	Percentagem de ácido esteárico no leite	1
48	*Milk vitamin B-12 content*	Teor de vitamina B-12 no leite	4
49	*Milk yield*	Produção de leite	11
50	*Milk yield (daughter deviation)*	Produção de leite (desvio da filha)	7
51	*Milk yield (EBV)*	Produção de leite (EBV)	15
52	*Milk yield (ECM)*	Produção de leite (ECM)	1
53	*Milk yield (PTA)*	Produção de leite (PTA)	1
54	*Milking speed*	Velocidade de ordenha	5
-	**Total**	-	**420**

estão destacados em negrito na Tabela 8.40. Além do mais, 21 QTLs foram selecionados simultaneamente pelos métodos SMS e valor-p corrigido, restando 14 QTLs não detectados

pelo SMS.

Tabela 8.40 SNPs selecionados pelo valor-p corrigido por Bonferroni menor que 0,05 com seus respectivos QTLs do leite identificados a partir do raio de 250.000 pb.

SNP	Cr.	Posição (pb)	Valor-p bruto	Valor-p corrigido	ID	Descrição original do QTL	Presença do QTL na União do SMS
Hapmap46502-BTA-89692	1	44.295.521	1,77E-06	4,05E-02	20878	Razão entre gordura e proteína no leite	**Sim**
BTA-70379-no-rs	4	48.017.690	1,81E-07	4,14E-03	18821	Produção de leite em 305 dias	**Sim**
					18819	Produção de leite média diária	**Sim**
ARS-BFGL-NGS-16315	5	21.813.311	1,43E-06	3,26E-02	**20624**	Percentagem de gordura no leite	**Sim**
					20618	Produção de gordura no leite	**Sim**
					20619	Produção de proteína no leite	**Sim**
					20139	Percentagem de gordura no leite	Não
					20138	Produção de gordura no leite	Não
					20135	Produção de proteína no leite	Não
					20140	Produção de leite	Não
ARS-BFGL-NGS-72551	5	28.390.921	1,89E-08	4,32E-04	**20624**	Percentagem de gordura no leite	**Sim**
					20618	Produção de gordura no leite	**Sim**
					20619	Produção de proteína no leite	**Sim**
ARS-BFGL-NGS-8796	5	29.984.792	1,10E-06	2,52E-02	**20624**	Percentagem de gordura no leite	**Sim**
					20618	Produção de gordura no leite	**Sim**
					20619	Produção de proteína no leite	**Sim**
Hapmap1572-BTA-111558	6	18.432.155	2,21E-07	5,04E-03	**20558**	Teor de alfa-caseína no leite	**Sim**
					20596	Razão de alfa-caseína no leite	**Sim**
					20563	Teor de beta-caseína no leite	**Sim**
					20579	Teor de kappa-caseína no leite	**Sim**
					20594	Teor de proteína no leite	**Sim**
BTB-00264279	7	26.383.062	6,75E-07	1,54E-02	25332	Percentagem de ácido cáprico no leite	Não
					25328	Percentagem de ácido cáprico no leite	Não
					25330	Percentagem de ácido caprílico no leite	Não
					25335	Percentagem de ácido láurico no leite	Não
					25337	Percentagem de ácido linoleico no leite	Não
					25325	Teor de ácido graxo monoinsaturado no leite	Não
					25336	Percentagem de ácido oleico no leite	Não
					25323	Teor de ácidos graxos saturados no leite	Não
					25326	Razão entre ácidos graxos saturados e insaturados no leite	Não
					25324	Teor de ácidos graxos insaturados no leite	Não
UA-IFASA-4213	11	22.057.744	2,64E-07	6,03E-03	**20587**	Teor de proteína beta-lactoglobulina no leite	**Sim**
					20078	Tempo de coagulação do coalho	**Sim**
ARS-BFGL-NGS-20488	20	42.090.174	5,86E-07	1,34E-02	26975	Percentagem de gordura no leite	Não
					26080	Produção de gordura no leite	Não
					26917	Percentagem de proteína no leite	Não
					26451	Produção de proteína no leite	Não
					25675	Produção de leite	Não
ARS-BFGL-NGS-67889	21	51.148.909	8,61255E-07	1,97E-02	31126	Produção de proteína no leite	Não
ARS-BFGL-NGS-36171	23	21.628.531	1,27576E-06	2,91E-02	**19984**	Produção de leite em 305 dias	**Sim**
					19985	Percentagem de proteína no leite	**Sim**

Pela Tabela 8.41, somente dois SNPs obtiveram valor-p corrigido menor que 0,05, o BTA-70379-no-rs (QTLs 18821 e 18819) e o ARS-BFGL-NGS-36171 (QTLs 19984 e 19988). Em relação ao tipo de *kernel*, o linear identificou somente dois SNPs; o radial com $\gamma = 0,001$, 22; o radial com $\gamma = 0,01$, 12 e o radial com $\gamma = 0,1$ selecionou sete SNPs.

O SMS detectou 38 QTLs associados à produção de leite que são divididos em três categorias: produção de leite em 305 dias, produção de leite média diária e produção de leite. No cromossomo 3, foram identificados dois QTLs referentes à produção de leite e, no cromossomo 4 (16156 e 14889), foi identificado o QTL 18821 referente a produção de leite em 305 dias e o QTL 18819 referente a produção de leite média diária (Tabela 8.41). No cromossomo 5, obtiveram-se sete QTLs, onde somente um QTL é referente a produção de leite em 305 dias (19050) (Tabela 8.41). No cromossomo 6, encontraram-se 11 QTLs, sendo o 15947 referente a produção de leite em 305 dias e o 21193 a produção de leite média diária (Tabela 8.41). Nos cromossomos 13, 14, 20 e 23, identificaram-se somente dois QTLs (19984 e 19988) em cada cromossomo, sendo que no cromossomo 23 dois QTLs são relativos à produção de leite em 305 dias (Tabela 8.41). Nos cromossomos 25 e 26, encontrou-se somente um QTL e nos cromossomos 28 e 30 identificaram-se respectivamente, dois e um (Tabela 8.41). Com relação aos *kernels*, o radial com $\gamma = 0,001$ selecionou 22 SNPs, o radial com $\gamma = 0,01$, 13 SNPs, o radial com $\gamma = 0,1$, seis SNPs e o linear, três SNPs, isso mostra que todos os *kernels* identificaram algum tipo de informação genômica importante associado à produção de leite.

A Tabela 8.42 mostra que o SMS detectou 27 QTLs associados à produção de gordura no leite. Ainda na Tabela 8.42, percebe-se que todos os SNPs possuem valores-p maiores que 0,05, ou seja, estatisticamente não significativos. Em relação aos *kernels*, 18 SNPs foram selecionados pelo radial com $\gamma = 0,001$, cinco pelo radial com $\gamma = 0,01$, cinco pelo radial com $\gamma = 0,1$ e dois pelo linear, o que indica que todos os *kernels* foram importantes na seleção de marcadores informativos que estão associados indiretamente com a produção de gordura do leite, pois o fenótipo usado no método SMS foi referente a produção de leite.

A Tabela 8.43 mostra que o SMS detectou 34 QTLs associados à produção de proteína no leite. Ainda na Tabela 8.43, percebe-se que todos os SNPs possuem valores-p maiores que 0,05, ou seja, estatisticamente não são significativos. Em relação aos *kernels*, 24 SNPs

Tabela 8.41 SNPs que marcam QTLs referentes à produção de leite com raio de 250.000 pb. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
Hapmap49050-BTA-119760	3	11.348.787	9,01E+00	16156[a]	Pimentel et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-25216	3	31901713	9,14E-01	14889[a]	Jiang et al. (2010)	$\gamma = 0,1$
Hapmap39284-BTA-70361	4	47448400	1,43E+00	18821[b]	Clempson et al. (2011)	$\gamma = 0,01$
				18819[c]		$\gamma = 0,001$
BTA-70379-no-rs	4	48017690	4,14E-03	18821[b]	Clempson et al. (2011)	Linear
				18819[c]	Clempson et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-26008	5	123920	1,44E+01	19050[b]	Mullen et al. (2011)	$\gamma = 0,01$
				9998[a]	Awad et al. (2010)	
				6062[a]	Daetwyler et al. (2008)	
				3604[a]	Bennewitz et al. (2003)	
				14137[a]	He et al. (2011)	
				13035[a]	Bonakdar et al. (2010)	
BTB-01320261	5	11417064	1,87E+03	16163[a]	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-53485	6	70612	2,24E+03	15947[b]	Chen et al. (2011)	$\gamma = 0,001$
				21193[c]	Bartonova et al. (2012)	Linear
				3735[a]	Wiener et al. (2000)	$\gamma = 0,001$
				11302[a]	Mei et al. (2009)	$\gamma = 0,001$
				2430[a]	Velmala et al. (1999)	$\gamma = 0,001$
				10042[a]	Khatib et al. (2007)	$\gamma = 0,001$
				3855[a]	Prinzenberg et al. (2003)	$\gamma = 0,001$
				10395[a]	Schnabel et al. (2005)	$\gamma = 0,001$
				15945[a]	Chen et al. (2011)	$\gamma = 0,001$
Hapmap38696-BTA-78192	6	11261546	2,11E+00	16233[a]	Pimentel et al. (2011)	Linear
Hapmap32217-BTC-039812	6	41714788	5,49E+03	31638[a]	Fontanesi et al. (2014)	$\gamma = 0,001$
ARS-BFGL-NGS-101567	6	44820065	1,70E+04			
BFGL-NGS-116163	6	45132015	2,26E+04			
BTA-64031-no-rs	6	87937981	1,93E+03	22666[a]	Veerkamp et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10855071	2,52E+01	16180[a]	Pimentel et al. (2011)	$\gamma = 0,001$
				16136[a]	Sikora et al. (2011)	
BTB-01689254	14	32307494	5,03E+03	14902[a]	Jiang et al. (2010)	$\gamma = 0,001$
Hapmap44166-BTA-94244	14	51919486	1,09E+03	18882[a]	Marques et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-16696	20	11037662	8,29E+03	16263[a]	Pimentel et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-84611	20	12062170	6,96E+03			$\gamma = 0,001$
ARS-BFGL-NGS-20706	20	21974955	2,58E+01	20232[a]	Chamberlain et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-36171	23	21628531	**2,91E-02**	19984[b]	Yang et al. (2012)	$\gamma = 0,1$
				19988[b]		
ARS-BFGL-NGS-4505	23	22345546	9,37E+00			$\gamma = 0,001$
ARS-BFGL-NGS-70555	24	41709373	8,69E+03	25717[a]	Meredith et al. (2012)	$\gamma = 0,001$
Hapmap31341-BTC-065490	25	44906068	5,59E+01	25719[a]	Meredith et al. (2012)	$\gamma = 0,01$
ARS-BFGL-NGS-33804	26	41800926	1,98E+03	25736[a]	Meredith et al. (2012)	$\gamma = 0,01$
Hapmap38041-BTA-64824	28	11182911	1,52E+00	16203[a]	Pimentel et al. (2011)	Linear
ARS-BFGL-NGS-42033	28	11993479	1,91E+02	16203[a]	Pimentel et al. (2011)	$\gamma = 0,01$
BFGL-NGS-110675	28	41638678	1,28E+04	25749[a]	Meredith et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-4441	30	67709	4,30E+03	1540[a]	Harder et al. (2006)	$\gamma = 0,1$
ARS-BFGL-NGS-94397	30	194212	1,34E+01	1540[a]	Harder et al. (2006)	$\gamma = 0,1$

[a] Produção de leite.
[b] Produção de leite em 305 dias.
[c] Produção de leite média diária.

foram selecionados pelo radial com $\gamma = 0,001$, 13 pelo radial com $\gamma = 0,01$, cinco pelo radial com $\gamma = 0,1$ e dois pelo linear, o que implica que todos os *kernels* também foram importantes na seleção de SNPs informativos indiretamente para a produção de proteína do leite, já que o fenótipo avaliado no método SMS foi referente à produção de leite.

A Tabela 8.44 indica que o SMS detectou 35 QTLs associados à porcentagem de gordura no leite. Ainda na Tabela 8.44, percebe-se que todos os SNPs possuem valores-p maiores que 0,05, ou seja, estatisticamente não significativos. Em relação aos *kernels*, 29 SNPs foram selecionados pelo *kernel* radial com $\gamma = 0,001$, dez pelo radial com $\gamma = 0,01$, três pelo radial com $\gamma = 0,1$ e cinco pelo linear, o que indica que todos os *kernels* foram relevantes na seleção de marcadores informativos indiretamente para a produção de proteína do leite, porque o fenótipo usado no método SMS foi baseado na produção de leite.

Tabela 8.42 SNPs que marcam QTLs referentes à produção de gordura no leite com raio de 250.000 pb. Dados extraídos de Genome (2015).

SNP	Cr.	Posição (pb)	Valor-p corrigido	ID	Referência	Kernel
BTA-55326-no-rs	1	12.059.290	2,09E+00	16207	Pimentel et al. (2011)	$\gamma = 0,001$
Hapmap49050-BTA-119760	3	11.348.787	9,01E+00	16157	Pimentel et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-26008	5	123.920	1,44E+01	9999	Awad et al. (2010)	$\gamma = 0,01$
				10364	Schrooten, Bink e Bovenhuis (2004)	
				13039	Bonakdar et al. (2010)	
BTB-01320246	5	11.383.669	1,08E+00	16171	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-01320261	5	11.417.064	1,87E+03			
ARS-BFGL-NGS-72504	5	25.451.573	3,11E+01	20618	Mullen et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-40375	5	37.165.461	1,11E-01			$\gamma = 0,01$
BTB-01342721	5	42.984.907	6,56E+01			Linear
BTA-91090-no-rs	5	43.593.104	2,28E+01			$\gamma = 0,001$
ARS-BFGL-NGS-69056	5	43.639.553	8,44E+01			$\gamma = 0,01$
Hapmap25585-BTA-150007	5	46.866.458	2,65E+01			$\gamma = 0,001$
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	3742	Wiener et al. (2000)	$\gamma = 0,001$
				11303	Mei et al. (2009)	
				1518	Harder et al. (2006)	
				10038	Khatib et al. (2007)	
				9939	Leyva-Baca et al. (2007)	
				10396	Mei et al. (2009)	
				15948	Chen et al. (2011)	
Hapmap38696-BTA-78192	6	11.261.546	2,11E+00	16234	Pimentel et al. (2011)	Linear
BTB-01689254	14	32.307.494	5,03E+03	14910	Jiang et al. (2010)	$\gamma = 0,001$
Hapmap44166-BTA-94244	14	51.919.486	1,09E+03	18851	Marques et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-95752	18	11.344.153	3,84E+00	16253	Pimentel et al. (2011)	$\gamma = 0,001$
Hapmap41864-BTA-42761	18	21.508.536	5,69E+00	30821	Zielke et al. (2013)	Linear
				30822	Zielke et al. (2013)	
ARS-BFGL-NGS-20706	20	21.974.955	2,58E+01	20230	Chamberlain et al. (2012)	$\gamma = 0,001$
BTA-57274-no-rs	24	11.364.121	1,50E+04	16272	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-00882078	24	11.740.116	6,24E+02			Linear
ARS-BFGL-NGS-54408	24	12.006.479	1,55E+04			$\gamma = 0,001$
ARS-BFGL-NGS-70555	24	41.709.373	8,69E+03	26094	Meredith et al. (2012)	$\gamma = 0,001$
Hapmap31341-BTC-065490	25	44.906.068	5,59E+01	26103	Meredith et al. (2012)	$\gamma = 0,01$
ARS-BFGL-NGS-33804	26	41.800.926	1,98E+03	26104	Meredith et al. (2012)	$\gamma = 0,01$
BFGL-NGS-110675	28	41.638.678	1,28E+04	26113	Meredith et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-4441	30	67.709	4,30E+03	10455	Sandor et al. (2006)	$\gamma = 0,1$
ARS-BFGL-NGS-94397	30	194.212	1,34E+01			

A Tabela 8.45 demonstra que o SMS detectou 39 QTLs associados à porcentagem de proteína no leite. Ainda na Tabela 8.45, percebe-se que todos os SNPs possuem valores-p maiores que 0,05, ou seja, estatisticamente não significativos. Em relação aos *kernels*, 25 SNPs foram selecionados pelo *kernel* radial com $\gamma = 0,001$, 12 pelo radial com $\gamma = 0,01$, sete pelo radial com $\gamma = 0,1$ e sete pelo linear, o que demonstra que todos os *kernels* também foram relevantes na seleção de marcadores informativos indiretamente para a percentagem de proteína do leite, pois o fenótipo utilizado no método SMS foi referente à produção de leite.

A Tabela 8.46 mostra que cinco QTLs encontrados estão associados à percentagem alfa-caseína no leite, e dois associados ao teor de alfa-caseína no leite, além de um QTL associado à razão entre alfa-caseína e beta-caseína no leite. Todos os 41 marcadores que identificaram os QTLs associados à alfa-caseína também obtiveram valores-p maiores que 0,05, não mostrando significância estatística. O *kernel* radial com $\gamma = 0,001$ identificou seis QTLs; o radial com $\gamma = 0,01$, três QTLs; o radial com $\gamma = 0,1$, dois QTLs; e o linear, também dois QTLs. Note que na Tabela 8.46, os QTLs 20558 e 20596 foram identificados por marcadores selecionados pelos quatro *kernels*.

Tabela 8.43 SNPs que marcam QTLs referentes à produção de proteína no leite com raio de 250.000 pb. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
BTA-55326-no-rs	1	12.059.290	2,09E+00	16208	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-25216	3	31.901.713	9,14E-01	14920	Jiang et al. (2010)	$\gamma = 0,1$
ARS-BFGL-NGS-26008	5	123.920	1,44E+01	9994	Awad et al. (2010)	$\gamma = 0,01$
				6063	Daetwyler et al. (2008)	
				10436	Bennewitz et al. (2004)	
				14138	He et al. (2011)	
				13308	Bonakdar et al. (2010)	
BTB-01320246	5	11.383.669	1,08E+00	16172	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-01320261	5	11.417.064	1,87E+03	16172	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-72504	5	25.451.573	3,11E+01	20619	Mullen et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-40375	5	37.165.461	1,11E-01			$\gamma = 0,01$
BTB-01342721	5	42.984.907	6,56E+01			Linear
BTA-91090-no-rs	5	43.593.104	2,28E+01			$\gamma = 0,001$
ARS-BFGL-NGS-69056	5	43.639.553	8,44E+01			$\gamma = 0,01$
Hapmap25585-BTA-150007	5	46.866.458	2,65E+01			$\gamma = 0,001$
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	3750	Wiener et al. (2000)	$\gamma = 0,001$
				6066	Daetwyler et al. (2008)	
				10289	Schrooten, Bink e Bovenhuis (2004)	
				3856	Prinzenberg et al. (2003)	
				15949	Chen et al. (2011)	
Hapmap31963-BTC-055529	6	32.262.005	1,47E+01	14922	Jiang et al. (2010)	Linear
Hapmap32217-BTC-039812	6	41.714.788	5,49E+03	31639	Fontanesi et al. (2014)	$\gamma = 0,001$
ARS-BFGL-NGS-101567	6	44.820.065	1,70E+04			
BFGL-NGS-116163	6	45.132.015	2,26E+04	31639	Fontanesi et al. (2014)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16181	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16137	Sikora et al. (2011)	$\gamma = 0,001$
Hapmap47517-BTA-116140	13	21.539.421	8,26E-01	20163	Chamberlain et al. (2012)	$\gamma = 0,01$
BTB-01689254	14	32.307.494	5,03E+03	14929	Jiang et al. (2010)	$\gamma = 0,001$
Hapmap44166-BTA-94244	14	51.919.486	1,09E+03	18874	Marques et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-95752	18	11.344.153	3,84E+00	16256	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-16696	20	11.037.662	8,29E+03	16264	Pimentel et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-84611	20	12.062.170	6,96E+03			$\gamma = 0,001$
ARS-BFGL-NGS-20706	20	21.974.955	2,58E+01	20223	Chamberlain et al. (2012)	$\gamma = 0,001$
BTA-87444-no-rs	23	11.810.893	1,81E+01	16268	Pimentel et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-36171	23	21.628.531	2,91E-02	20240	Chamberlain et al. (2012)	$\gamma = 0,1$
ARS-BFGL-NGS-4505	23	22.345.546	9,37E+00			$\gamma = 0,001$
BTA-57274-no-rs	24	11.364.121	1,50E+04	16273	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-00882078	24	11.740.116	6,24E+02			Linear
ARS-BFGL-NGS-54408	24	12.006.479	1,55E+04			$\gamma = 0,001$
ARS-BFGL-NGS-70555	24	41.709.373	8,69E+03	26476	Meredith et al. (2012)	$\gamma = 0,001$
Hapmap31341-BTC-065490	25	44.906.068	5,59E+01	26478	Meredith et al. (2012)	$\gamma = 0,01$
ARS-BFGL-NGS-33804	26	41.800.926	1,98E+03	26484	Meredith et al. (2012)	$\gamma = 0,01$
Hapmap38041-BTA-64824	28	11.182.911	1,52E+00	16204	Pimentel et al. (2011)	Linear
ARS-BFGL-NGS-42033	28	11.993.479	1,91E+02			$\gamma = 0,01$
BFGL-NGS-110675	28	41.638.678	1,28E+04	26497	Meredith et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-4441	30	67.709	4,30E+03	1539	Harder et al. (2006)	$\gamma = 0,1$
ARS-BFGL-NGS-94397	30	194.212	1,34E+01			

A Tabela 8.47 demonstra que ao todo foram identificados oito QTLs associados à beta-caseína e à beta-lactoglobulina no leite. De forma específica, dois QTLs encontrados estão associados ao teor de beta-caseína no leite, um QTL à percentagem de beta-caseína, dois QTLs ao teor de beta-lactoglobulina e três QTLs à percentagem de beta-lactoglobulina. Todos os 41 marcadores que identificaram os QTLs associados à beta-caseína e à beta-lactoglobulina também obtiveram valores-p maiores que 0,05, mostrando ausência de significância estatística.

A Tabela 8.48 demonstra que ao todo foram identificados três QTLs relacionados à caseína no leite. De forma específica, dois QTLs encontrados estão associados ao índice caseína no leite e um QTL à percentagem de caseína no leite. Todos os três marcadores que identificaram os QTLs associados à beta-caseína e à beta-lactoglobulina também obtiveram valores-p maiores que 0,05, mostrando ausência de significância estatística.

Tabela 8.44 SNPs que marcam QTLs referentes à porcetagem de gordura no leite com raio de 250.000 pb. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
BTA-55326-no-rs	1	12.059.290	2,09E+00	16209	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-152	1	88.956.654	2,22E+04	22574	Maxa et al. (2012)	$\gamma = 0,001$
BTA-40510-no-rs	1	89.653.845	5,36E+03			
Hapmap27145-BTA-116154	2	32.094.742	7,96E+01	14931	Jiang et al. (2010)	$\gamma = 0,001$
ARS-BFGL-NGS-26008	5	123.920	1,44E+01	2717	Olsen et al. (2002)	$\gamma = 0,01$
				11501	Schennink et al. (2009)	
				2717	Olsen et al. (2002)	
BTB-01320246	5	11.383.669	1,08E+00	16164	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-01320261	5	11.417.064	1,87E+03			
ARS-BFGL-NGS-72504	5	25.451.573	3,11E+01	20624	Mullen et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-40375	5	37.165.461	1,11E-01			$\gamma = 0,01$
BTB-01342721	5	42.984.907	6,56E+01			Linear
BTA-91090-no-rs	5	43.593.104	2,28E+01			$\gamma = 0,001$
ARS-BFGL-NGS-69056	5	43.639.553	8,44E+01			$\gamma = 0,01$
Hapmap25585-BTA-150007	5	46.866.458	2,65E+01			$\gamma = 0,001$
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	10214	Georges et al. (1995)	$\gamma = 0,001$
				11305	Mei et al. (2009)	
				10266	Olsen et al. (2005)	
				2595	Viitala et al. (2003)	
				10469	Bovenhuis e Weller (1994)	
				10039	Khatib et al. (2007)	
				10478	Cohen-Zinder et al. (2005)	
				18418	Schnabel et al. (2005)	
Hapmap38696-BTA-78192	6	11.261.546	2,11E+00	16057	Zheng et al. (2011)	Linear
				16235	Pimentel et al. (2011)	
ARS-BFGL-NGS-65089	8	11.828.306	8,88E+02	16239	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-00470654	11	31.864.223	1,26E+04	14935	Jiang et al. (2010)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16182	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16138	Sikora et al. (2011)	$\gamma = 0,001$
BTB-01689254	14	32.307.494	5,03E+03	30778	Wang et al. (2012b)	$\gamma = 0,001$
				14998	Jiang et al. (2010)	
Hapmap44166-BTA-94244	14	51.919.486	1,09E+03	18881	Marques et al. (2011)	$\gamma = 0,1$
Hapmap41864-BTA-42761	18	21.508.536	5,69E+00	20172	Chamberlain et al. (2012)	Linear
ARS-BFGL-NGS-16696	20	11.037.662	8,29E+03	16265	Pimentel et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-84611	20	12.062.170	6,96E+03	16265	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-20706	20	21.974.955	2,58E+01	20238	Chamberlain et al. (2012)	$\gamma = 0,001$
Hapmap32374-BTA-158494	20	30.631.380	5,87E+03	30779	Wang et al. (2012b)	$\gamma = 0,01$
ARS-BFGL-NGS-36171	23	21.628.531	2,91E-02	19989	Yang et al. (2012)	$\gamma = 0,1$
ARS-BFGL-NGS-4505	23	22.345.546	9,37E+00	19989	Yang et al. (2012)	$\gamma = 0,001$
BTA-57274-no-rs	24	11.364.121	1,50E+04	16274	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-00882078	24	11.740.116	6,24E+02			Linear
ARS-BFGL-NGS-54408	24	12.006.479	1,55E+04			$\gamma = 0,001$
ARS-BFGL-NGS-70555	24	41.709.373	8,69E+03	27134	Meredith et al. (2012)	$\gamma = 0,001$
Hapmap31341-BTC-065490	25	44.906.068	5,59E+01	26720	Meredith et al. (2012)	$\gamma = 0,01$
ARS-BFGL-NGS-33804	26	41.800.926	1,98E+03	26721	Meredith et al. (2012)	$\gamma = 0,01$
Hapmap38041-BTA-64824	28	11.182.911	1,52E+00	16205	Pimentel et al. (2011)	Linear
ARS-BFGL-NGS-42033	28	11.993.479	1,91E+02			$\gamma = 0,01$
BFGL-NGS-110675	28	41.638.678	1,28E+04	26722	Meredith et al. (2012)	$\gamma = 0,001$

A Tabela 8.49 apresenta dois QTLs associados à relação entre gordura e proteína no leite. No total foram selecionados sete SNPs, sendo que cinco identificaram o QTL 20878 e dois SNPs marcaram o QTL 20907. Com relação aos *kernels*, os dois QTLs foram selecionados pelo radial com $\gamma = 0,001$, somente o QTL 20878 foi selecionado pelo $\gamma = 0,01$, nenhum para $\gamma = 0,1$ e os dois QTLs pelo linear.

Pela Tabela 8.50, 14 marcadores identificaram 34 QTLs associados a vitaminas, proteínas e ácidos graxos que compõem o leite. Dentre todos os SNPs selecionados, somente o marcador UA-IFASA-4213 teve valor-p significativo igual a 6,03E-03. O *kernel* radial com $\gamma = 0,001$ identificou 18 QTLs, o radial com $\gamma = 0,01$ três QTLs, o radial com $\gamma = 0,1$, 13 QTLs e nenhum QTL foi encontrado com o linear.

Com base nos resultados de QTLs associados à produção e à composição do leite que foram identificados pelos SNPs selecionados pelo SMS com base na PTA do leite,

Tabela 8.45 SNPs que marcam QTLs com raio de 250.000 pb referentes à porcentagem de proteína no leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
ARS-BFGL-NGS-68915	1	21.723.208	5,51E-02	20134	Chamberlain et al. (2012)	Linear
UA-IFASA-9700	1	21.834.588	9,12E-02			
Hapmap39987-BTA-25749	1	21.882.817	1,22E+03			$\gamma = 0,01$
Hapmap49050-BTA-119760	3	11.348.787	9,01E+00	16159	Pimentel et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-26008	5	123.920	1,44E+01	10438	Bennewitz et al. (2004)	$\gamma = 0,01$
				20597	Wang et al. (2012a)	
ARS-BFGL-NGS-28269	5	9.299.499	5,12E-02	15763	Schopen et al. (2011)	$\gamma = 0,1$
Hapmap33365-BTA-142791	5	9.543.667	2,45E+02			Linear
BTB-01320246	5	11.383.669	1,08E+00	16165	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-01320261	5	11.417.064	1,87E+03			
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	3726	Tassell, Ashwell e Sonstegard (2000)	$\gamma = 0,001$
				10210	Freyer et al. (2002)	
				10265	Olsen et al. (2005)	
				10288	Schrooten, Bink e Bovenhuis (2004)	
				10040	Khatib et al. (2007)	
				3857	Prinzenberg et al. (2003)	
				18419	Schnabel et al. (2005)	
				3364	Ashwell e Tassell (1999)	
BFGL-NGS-116085	6	9.346.033	4,16E+01	15768	Schopen et al. (2011)	$\gamma = 0,1$
Hapmap38696-BTA-78192	6	11.261.546	2,11E+00	16236	Pimentel et al. (2011)	Linear
				16057	Schopen et al. (2011)	
Hapmap31963-BTC-055529	6	32.262.005	1,47E+01	15005	Jiang et al. (2010)	Linear
Hapmap32217-BTC-039812	6	41.714.788	5,49E+03	31643	Fontanesi et al. (2014)	$\gamma = 0,001$
ARS-BFGL-NGS-101567	6	44.820.065	1,70E+04			
BFGL-NGS-116163	6	45.132.015	2,26E+04	31643	Fontanesi et al. (2014)	$\gamma = 0,001$
ARS-BFGL-NGS-65089	8	11.828.306	8,88E+02	16240	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-01095261	9	22.232.218	4,05E+00	20155	Chamberlain et al. (2012)	$\gamma = 0,01$
BTB-00382893	9	22.702.729	2,29E+01			
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16183	Pimentel et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	16139	Sikora et al. (2011)	$\gamma = 0,001$
Hapmap47517-BTA-116140	13	21.539.421	8,26E-01	20161	Chamberlain et al. (2012)	$\gamma = 0,01$
BTB-01689254	14	32.307.494	5,03E+03	15012	Jiang et al. (2010)	$\gamma = 0,001$
Hapmap44166-BTA-94244	14	51.919.486	1,09E+03	18859	Marques et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-21847	15	10.263.008	8,17E+02	15781	Schopen et al. (2011)	$\gamma = 0,001$
Hapmap41864-BTA-42761	18	21.508.536	5,69E+00	20174	Chamberlain et al. (2012)	Linear
Hapmap49928-BTA-24568	20	9.993.632	2,30E+01	15785	Schopen et al. (2011)	$\gamma = 0,01$
Hapmap33374-BTA-148421	20	10.276.329	3,03E+01			
ARS-BFGL-NGS-16696	20	11.037.662	8,29E+03			
ARS-BFGL-NGS-20706	20	21.974.955	2,58E+01	20237	Chamberlain et al. (2012)	$\gamma = 0,001$
BTA-87444-no-rs	23	11.810.893	1,81E+01	16269	Pimentel et al. (2011)	$\gamma = 0,01$
ARS-BFGL-NGS-36171	23	21.628.531	2,91E-02	19985	Yang et al. (2012)	$\gamma = 0,1$
				20242	Chamberlain et al. (2012)	
ARS-BFGL-NGS-4505	23	22.345.546	9,37E+00		Chamberlain et al. (2012)	$\gamma = 0,001$
BTA-57274-no-rs	24	11.364.121	1,50E+04	16275	Pimentel et al. (2011)	$\gamma = 0,001$
BTB-00882078	24	11.740.116	6,24E+02			Linear
ARS-BFGL-NGS-54408	24	12.006.479	1,55E+04			$\gamma = 0,001$
Hapmap31341-BTC-065490	25	44.906.068	5,59E+01	26942	Meredith et al. (2012)	$\gamma = 0,01$
ARS-BFGL-NGS-33804	26	41.800.926	1,98E+03	26945	Meredith et al. (2012)	$\gamma = 0,01$
Hapmap38041-BTA-64824	28	11.182.911	1,52E+00	16206	Pimentel et al. (2011)	Linear
ARS-BFGL-NGS-42033	28	11.993.479	1,91E+02			$\gamma = 0,01$
BFGL-NGS-110675	28	41.638.678	1,28E+04	26947	Meredith et al. (2012)	$\gamma = 0,001$
Hapmap32898-BTA-66437	29	9.778.300	1,43E-01	15796	Schopen et al. (2011)	$\gamma = 0,1$
ARS-BFGL-NGS-51329	29	10.447.643	1,15E+04			
BTA-66539-no-rs	29	11.051.399	1,92E+03			$\gamma = 0,001$

conclui-se que esse método avalia implicitamente pelo SVR os efeitos dos SNPs de forma distinta dos métodos dos valores-p bruto e corrigido e, consequentemente, detecta regiões não mapeados por abordagens monoatributos. Apesar do fenótipo avaliado ser a PTA do leite, o SMS identificou vários QTLs associados a outros fenótipos correlacionados com a produçao de leite. Esse fato pode ser explicado pela possível existência de genes em comum que atuam na determinação dos fenótipos encontrados, ou seja, existe o fenômeno da pleiotropia[2]. Portanto, o SMS mostrou maior eficiência na identificação de regiões com QTLs associados à produção e composição do leite do que o método do valor-p corrigido.

[2]São os múltiplos efeitos de um gene, ou seja, um gene afeta múltiplas características (PIERCE, 2013). A fenilcetonúria resulta de um alelo recessivo, onde as pessoas homozigotas para esse alelo, caso não tratadas, apresentam retardo mental, olhos azuis e pele de cor clara (PIERCE, 2013).

Tabela 8.46 SNPs que marcam QTLs com raio de 250.000 pb referentes à alfa-caseína no leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID do QTL	Referência	Kernel
BTA-38496-no-rs	1	9.858.346	1,80E+02	15761[a]	Schopen et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-1296	1	10.394.962	5,11E+03			
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	13574[b]	Heck et al. (2009)	$\gamma = 0,001$
				21535[a]	Schopen et al. (2009)	
BTB-00254654	6	17.935.515	1,97E+02	20558[b]	Huang et al. (2012)	Linear
				20596[c]		
BFGL-NGS-119394	6	18.000.390	2,51E+03	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
Hapmap27235-BTA-163445	6	18.040.265	1,63E+01	20558[b]	Huang et al. (2012)	$\gamma = 0,01$
				20596[c]		
Hapmap1572-BTA-111558	6	18.432.155	5,04E-03	20558[b]	Huang et al. (2012)	$\gamma = 0,1$
				20596[c]		
BTA-94473-no-rs	6	19.649.589	1,49E+04	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
Hapmap31963-BTC-055529	6	32.262.005	1,47E+01	20558[b]	Huang et al. (2012)	Linear
				20596[b]		
Hapmap26885-BTC-055761	6	32.528.610	6,66E+02	20558[b]	Huang et al. (2012)	$\gamma = 0,01$
				20596[c]		
Hapmap29925-BTC-036976	6	33.430.552	1,49E-01	20558[b]	Huang et al. (2012)	$\gamma = 0,1$
				20596[c]		
Hapmap23507-BTC-041133	6	34.663.790	5,10E-01	20558[b]	Huang et al. (2012)	$\gamma = 0,1$
				20596[c]		
Hapmap44513-BTA-107931	6	39.612.323	7,66E+00	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
ARS-BFGL-NGS-23937	6	39.878.785	5,92E+01	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
Hapmap32217-BTC-039812	6	41.714.788	5,49E+03	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
ARS-BFGL-NGS-101567	6	44.820.065	1,70E+04	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
BFGL-NGS-116163	6	45.132.015	2,26E+04	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
Hapmap23279-BTA-158875	6	46.505.294	1,14E+03	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
ARS-BFGL-NGS-58275	6	46.542.372	1,47E+03	20558[b]	Huang et al. (2012)	$\gamma = 0,001$
				20596[c]		
BTB-01201475	6	47.378.342	2,34E+01	20558[b]	Huang et al. (2012)	$\gamma = 0,1$
				20596[c]		
Hapmap41190-BTA-16578	11	10.889.518	1,40E+04	15774[a]	Schopen et al. (2011)	$\gamma = 0,001$
ARS-BFGL-NGS-64857	13	10.855.071	2,52E+01	15779[a]	Schopen et al. (2011)	$\gamma = 0,001$
BTA-55233-no-rs	22	10.369.618	1,04E+02	15788[a]	Schopen et al. (2011)	$\gamma = 0,01$

[a] Percentagem de alfa-caseína no leite.
[b] Teor de alfa-caseína no leite.
[c] Relação entre alfa-caseína e beta-caseína no leite.

Tabela 8.47 SNPs que marcam QTLs com raio de 250.000 pb referentes à beta-caseína no leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	13575[a]	Heck et al. (2009)	$\gamma = 0,001$
				13572[b]		
BFGL-NGS-116085	6	9.346.033	4,16E+01	15766[c]	Schopen et al. (2011)	$\gamma = 0,1$
BTB-00254654	6	17.935.515	1,97E+02	20563[a]	Huang et al. (2012)	Linear
BFGL-NGS-119394	6	18.000.390	2,51E+03			$\gamma = 0,001$
Hapmap27235-BTA-163445	6	18.040.265	1,63E+01			$\gamma = 0,01$
Hapmap1572-BTA-111558	6	18.432.155	5,04E-03			$\gamma = 0,1$
BTA-94473-no-rs	6	19.649.589	1,49E+04			$\gamma = 0,001$
Hapmap31963-BTC-055529	6	32.262.005	1,47E+01			Linear
Hapmap26885-BTC-055761	6	32.528.610	6,66E+02			$\gamma = 0,01$
Hapmap29925-BTC-036976	6	33.430.552	1,49E-01			$\gamma = 0,1$
Hapmap23507-BTC-041133	6	34.663.790	5,10E-01			$\gamma = 0,1$
Hapmap44513-BTA-107931	6	39.612.323	7,66E+00			$\gamma = 0,001$
ARS-BFGL-NGS-23937	6	39.878.785	5,92E+01			$\gamma = 0,001$
Hapmap32217-BTC-039812	6	41.714.788	5,49E+03			$\gamma = 0,001$
ARS-BFGL-NGS-101567	6	44.820.065	1,70E+04			$\gamma = 0,001$
BFGL-NGS-116163	6	45.132.015	2,26E+04			$\gamma = 0,001$
Hapmap23279-BTA-158875	6	46.505.294	1,14E+03			$\gamma = 0,001$
ARS-BFGL-NGS-58275	6	46.542.372	1,47E+03			$\gamma = 0,001$
BTB-01201475	6	47.378.342	2,34E+01			$\gamma = 0,1$
Hapmap41190-BTA-16578	11	10.889.518	1,40E+04	15775[d]	Schopen et al. (2011)	$\gamma = 0,001$
Hapmap41190-BTA-16578	11	10.889.518	1,40E+04	15777[c]		
BTB-01940421	11	18.411.000	3,51E+03	20587[b]	Huang et al. (2012)	$\gamma = 0,001$
BTB-01996428	11	18.506.708	1,81E+01			$\gamma = 0,001$
Hapmap36648-SCAFFOLD255197_18545	11	18.970.267	1,13E+04			$\gamma = 0,001$
Hapmap32683-BTA-88615	11	19.999.912	5,96E+03			$\gamma = 0,001$
BTB-00464946	11	21.193.152	1,91E+04			$\gamma = 0,001$
UA-IFASA-4213	11	22.057.744	6,03E-03			$\gamma = 0,01$
BTA-87562-no-rs	11	22.627.876	4,42E+03			$\gamma = 0,001$
ARS-BFGL-NGS-43804	11	23.025.746	6,93E+03			$\gamma = 0,001$
BFGL-NGS-110642	11	23.066.087	6,18E+03			$\gamma = 0,001$
BFGL-NGS-116483	11	23.125.907	4,99E+00			Linear
BTB-00466317	11	23.288.695	9,45E-01			$\gamma = 0,001$
ARS-BFGL-NGS-62536	11	23.342.934	1,31E+02			$\gamma = 0,001$
BTA-93788-no-rs	11	24.063.543	8,19E+03			$\gamma = 0,001$
BTB-01260153	11	28.879.413	1,85E+04			$\gamma = 0,001$
ARS-BFGL-NGS-15269	11	30.351.845	1,70E+04			$\gamma = 0,001$
BTB-00470654	11	31.864.223	1,26E+04			$\gamma = 0,001$
BTA-122098-no-rs	11	32.375.337	5,44E+01			$\gamma = 0,01$
Hapmap23482-BTA-126430	11	38.578.120	6,78E+03			$\gamma = 0,001$
BTA-18966-no-rs	24	10.699.491	1,21E+04	15789[c]	Schopen et al. (2011)	$\gamma = 0,001$

[a] Teor de beta-caseína no leite.
[b] Teor de proteína beta-lactoglobulina no leite.
[c] Percentagem de beta-lactoglobulina no leite.
[d] Percentagem de beta-caseína no leite.

Tabela 8.48 SNPs que marcam QTLs com raio de 250.000 pb referentes à caseína no leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	13577[a]	Heck et al. (2009)	$\gamma = 0,001$
BFGL-NGS-116085	6	9.346.033	4,16E+01	15767[b]	Schopen et al. (2011)	$\gamma = 0,1$
Hapmap41190-BTA-16578	11	10.889.518	1,40E+04	15778[b]	Schopen et al. (2011)	$\gamma = 0,001$

[a] Percentagem de caseína no leite.
[b] Índice de caseína no leite.

Tabela 8.49 SNPs que marcam QTLs com raio de 250.000 pb referentes à relação entre gordura e proteína no leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor-p corrigido	ID	Referência	Kernel
Hapmap43356-BTA-79201	1	37.627.303	1,32E+03	20878	Tetens et al. (2013)	$\gamma = 0,001$
Hapmap47738-BTA-79246	1	40.492.488	1,67E+03			$\gamma = 0,01$
Hapmap48210-BTA-120730	1	43.043.206	6,88E+03			$\gamma = 0,001$
Hapmap50812-BTA-89669	1	43.185.098	1,75E+00			$\gamma = 0,01$
Hapmap46502-BTA-89692	1	44.295.521	**4,05E-02**			Linear
BTA-62673-no-rs	27	27.560.414	2,29E+00	20907	Tetens et al. (2013)	Linear
BTA-62667-no-rs	27	27.586.592	2,32E+01			$\gamma = 0,001$

Tabela 8.50 SNPs que marcam QTLs com raio de 250.000 pb referentes a vários QTLs associados ao leite. Dados extraídos de Genome (2015).

SNP	Cr.	Posição	Valor p corrigido	ID	Descrição	Artigo	Kernel
BTA-55326-no-rs	1	12.059.290	2,09E+00	23178	Teor de vitamina B-12	Rutten et al. (2013)	$\gamma = 0,001$
BTA-72648-no-rs	4	21.545.578	4,16E+01	20063	Índice de ácido graxo insaturado	Rincon et al. (2012)	$\gamma = 0,001$
				20033	Relação entre ácido graxo saturado e insaturado		
				20041	Teor de ácido graxo monoinsaturado		
BTB-00677925	4	22.280.165	2,73E+03	20063	Índice de ácido graxo insaturado	Rincon et al. (2012)	$\gamma = 0,001$
				20033	Relação entre ácido graxo saturado e insaturado		
				20041	Teor de ácido graxo monoinsaturado		
ARS-BFGL-NGS-26008	5	123.920	1,44E+01	11504	Índice de ácido graxo insaturado	Schennink et al. (2009)	$\gamma = 0,01$
				3459	Velocidade de ordenha	Boichard et al. (2003)	
ARS-BFGL-NGS-53485	6	70.612	2,24E+03	25040	Taxa de endurecimento da coalhada	Poulsen et al. (2013)	$\gamma = 0,001$
				11510	Índice de ácido graxo insaturado	Schennink et al. (2009)	
				11507	Percentagem de ácido mirístico		
				11505	Percentagem de ácido mirístico		
				19930	Percentagem de sólidos	Silva et al. (2011)	
				19928	Sólidos		
				19929	Teor de lactose		
				19927	Produção de lactose		
				1519	Produção de energia	Harder et al. (2006)	
				25034	Tempo de coagulação do coalho no leite	Cecchinato et al. (2012)	
				3437	Velocidade de ordenha	Boichard et al. (2003)	
BTA-64031-no-rs	6	87.937.981	1,93E+03	22754	Velocidade de ordenha	Boichard et al. (2003)	$\gamma = 0,001$
ARS-BFGL-NGS-65089	8	11.828.306	8,88E+02	23189	Teor de vitamina B-12	Rutten et al. (2013)	$\gamma = 0,001$
BTA-28730-no-rs	8	12.361.086	1,29E+04	23189	Teor de vitamina B-12	Rutten et al. (2013)	$\gamma = 0,01$
UA-IFASA-4213	11	22.057.744	**6,03E-03**	20078	Tempo de coagulação do coalho no leite	Cecchinato et al. (2012)	$\gamma = 0,01$
BTA-87562-no-rs	11	22.627.876	4,42E+03	20078	Tempo de coagulação do coalho no leite	Cecchinato et al. (2012)	$\gamma = 0,001$
ARS-BFGL-NGS-47995	19	51.728.559	2,70E+00	23254	Percentagem de ácido mirístico	Bouwman et al. (2014)	$\gamma = 0,1$
				23256	Percentagem de ácido mirístico		
				23246	Percentagem de ácido mirístico		
				23253	Percentagem de ácido mirístico		
				23247	Percentagem de ácido mirístico		
				23248	Percentagem de ácido mirístico		
				23249	Percentagem de ácido mirístico		
				23250	Percentagem de ácido mirístico		
				23251	Percentagem de ácido mirístico		
				23252	Percentagem de ácido mirístico		
				23245	Percentagem de ácido mirístico		
ARS-BFGL-NGS-80071	22	12.461.662	2,49E+03	23211	Produção de energia	Rutten et al. (2013)	$\gamma = 0,001$
ARS-BFGL-NGS-4441	30	67.709	4,30E+03	1541	Produção de energia	Harder et al. (2006)	$\gamma = 0,1$
				10456	Velocidade de ordenha	Boichard et al. (2003)	
ARS-BFGL-NGS-94397	30	194.212	1,34E+01	1541	Produção de energia	Harder et al. (2006)	$\gamma = 0,1$
				10456	Velocidade de ordenha	Boichard et al. (2003)	

8.10.2.2 Genes Candidatos Identificados pelo SMS

Os 1.265 marcadores selecionados pelo SMS identificaram um total de 6.317 genes, sendo que desse total, 90 são genes candidatos relacionados à produção de leite e à mastite [3] que foram identificados em trabalhos anteriores organizados por Ogorevc et al. (2009). Especificamente, 36 genes são relacionados a estudos de expressão do leite, seis a estudos de associação do leite, 17 a estudos de expressão da mastite, dois a estudos de associação de mastite e 29 referentes a estudos do modelo animal baseado no modelo do camundongo.

A Tabela 8.51 demonstra identificação de genes pelos SNPs selecionados por cada *kernel* usado no SMS. Os maiores subconjuntos selecionados foram os referentes ao *kernel* radial com $\gamma = 0,001$ e $\gamma = 0,1$ com, respectivamente, 3.184 e 1.835 genes. Apesar do $\gamma = 0,001$ ser o mais próximo de zero dos γs avaliados, ele selecionou um número consideravelmente superior ao *kernel* linear. Isso mostra que, dentre os três *kernels* radiais avaliados, apesar do $\gamma = 0,001$ ter a maior similaridade com o linear, ele não é suficientemente próximo a zero para demonstrar comportamento análogo a esse *kernel*. Os cromossomos que tiveram as maiores quantidades de genes identificados foram o 15, o 18 e o 25 com respectivamente 423, 448 e 478 genes. Por outro lado, os que tiveram as menores quantidades de genes encontrados foram o 27 e o X com respectivamente 47 e 33 genes.

Na Tabela 8.52, estão indicados os 36 genes candidatos relacionados a estudos anteriores de expressão do leite em gado marcados pelos SNPs selecionados pelo SMS. Note que todos os genes foram unicamente identificados por um único marcador, ou seja, não ocorreu redundância de identificação por tipo de *kernel* adotado. O número de marcadores selecionados por *kernel* foi em ordem decrescente: 20 para o radial com $\gamma = 0,001$; nove para o radial com $\gamma = 0,1$; cinco para o radial com $\gamma = 0,01$ e dois para o linear. Do total de 36 genes, 25 foram marcados por SNPs com valores-p menores que 0,05, o que mostra uma quantidade de marcadores SNP selecionados, que identificaram genes candidatos, simultaneamente pelo SMS e pelo valor-p corrigido superior à quantidade de SNPs do mesmo conjunto interseção para QTLs associados ao leite que pode ser obervado um único

[3] A mastite ocorre quando o úbere fica inflamado, porque os leucócitos (ou células somáticas) são enviados para a glândula mamária, em resposta à invasão do canal da teta, normalmente por bactérias (JONES; BAILEY, 2009). Estas bactérias se multiplicam e produzem toxinas que causam lesão ao tecido secretor do leite e várias vias em toda a glândula mamária (JONES; BAILEY, 2009). Elevada quantidade de leucócitos causa uma redução na produção de leite e altera a composição do leite, que por sua vez afetam adversamente a qualidade e a quantidade de leite (JONES; BAILEY, 2009).

Tabela 8.51 Genes identificados pelos 1.265 SNPs selecionados pelo SMS agrupados por cromossomo para cada *kernel*.

Cr.	Kernel				**Total**
	Linear	Radial $\gamma = 0,001$	Radial $\gamma = 0,01$	Radial $\gamma = 0,1$	
1	33	75	7	43	**158**
2	11	139	7	44	**201**
3	12	117	14	99	**242**
4	5	54	13	37	**109**
5	21	46	26	33	**126**
6	11	209	27	30	**277**
7	35	260	17	33	**345**
8	6	61	21	26	**114**
9	5	89	27	34	**155**
10	5	79	24	39	**147**
11	17	174	28	63	**282**
12	9	54	4	54	**121**
13	38	255	54	41	**388**
14	20	100	19	49	**188**
15	36	273	9	105	**423**
16	41	81	4	65	**191**
17	1	49	43	70	**163**
18	73	170	39	166	**448**
19	22	144	32	116	**314**
20	15	43	46	23	**127**
21	3	85	14	18	**120**
22	4	78	41	122	**245**
23	18	84	26	52	**180**
24	6	90	3	15	**114**
25	38	161	73	206	**478**
26	34	41	25	41	**141**
27	2	15	11	19	**47**
28	11	88	26	42	**167**
29	23	66	56	128	**273**
X	7	4	0	22	**33**
Total	**562**	**3.184**	**736**	**1.835**	**6.317**

SNP (ARS-BFGL-NGS-36171) na Tabela 8.41. Uma observação que deve ser ressaltada é que dois SNPs, Hapmap23279-BTA-158875 no cromossomo 6 e Hapmap25580-BTA-149003 no cromossomo 15, estão dentro dos respectivos genes candidatos identificados SLC34A2 e MMP12 que pode ser visto na Tabela 8.52. Além disso, esses SNPs possuem valores-p significativos, porém, o valor-p do marcador Hapmap23279-BTA-158875 difere apenas 0,01 do ponto de corte 0,05, isto é, por uma diferença ínfima esse marcador não foi eliminado da seleção pelo valor-p corrigido.

Na Tabela 8.53, estão indicados os seis genes candidatos relacionados a estudos anteriores de associação do leite em gado marcados pelos SNPs selecionados pelo SMS. Todos os *kernels* usados identificaram pelo menos um gene candidato associado à produção de leite. Em relação aos SNPs selecionados pelo SMS, somente o ARS-BFGL-NGS-101567 e o ARS-BFGL-NGS-105633 possuem valor-p corrigido maior que 0,05, isto é, não significativos estatisticamente. O marcador ARS-BFGL-NGS-101567 no cromossomo 6 está localizado dentro do gene candidato PPARGC1A associado à produção de leite, entretanto seu valor-p de 7,44E-01 não é significante estatisticamente. De forma contrária, o marcador ARS-BFGL-NGS-76756 no cromossomo 20 está localizado dentro do gene candidato GHR associado à produção de leite, mas seu valor-p de 5,25E-06 é significativo do ponto de vista estatístico.

Tabela 8.52 Genes candidatos a partir de estudos anteriores de expressão do leite para *Bos taurus* segundo a base de dados de Ogorevc et al. (2009) marcados pelos SNPs selecionados pelo SMS com raio de 250.000 pb.

Referência	Gene	Cr.	SNP mais próximo	Posição (pb)	Distância (pb)	Valor-p corrigido do SNP	Kernel
Ron et al. (2007)	PFKL	1	**ARS-BFGL-NGS-33058**	147.418.155	182.655	**8,53E-10**	$\gamma = 0,1$
	CP	1	Hapmap24824-BTA-136168	120.791.279	104.721	1,44E-01	$\gamma = 0,001$
	FAM46C	3	**Hapmap54955-rs29010328**	25.503.180	191.610	**4,27E-02**	$\gamma = 0,001$
	FRRS1	3	**BTA-67733-no-rs**	46.735.459	224.459	**4,99E-02**	$\gamma = 0,001$
	ELOVL1	3	**ARS-BFGL-NGS-31442**	109.683.198	108.098	**7,81E-07**	$\gamma = 0,01$
	SLC34A2	6	**Hapmap23279-BTA-158875**	46.505.294	**0**	**4,99E-02**	$\gamma = 0,001$
	CD320	7	**UA-IFASA-7060**	15.404.137	7.663	**3,09E-04**	Linear
	CD24	9	**BTB-01352842**	50.042.020	42.019	**8,65E-04**	$\gamma = 0,01$
	GK	9	**Hapmap38580-BTA-122605**	922.974	227.974	**7,53E-05**	$\gamma = 0,1$
	ACOT2	10	Hapmap40595-BTA-79304	87.178.746	249.454	1,70E-01	$\gamma = 0,001$
	CEL	11	**ARS-BFGL-NGS-59310**	106.659.331	149.769	**5,08E-03**	$\gamma = 0,001$
	NDRG1	14	**ARS-BFGL-BAC-8730**	7.430.109	79.991	**2,66E-03**	$\gamma = 0,1$
	STARD10	15	**BTB-01371672**	51.736.863	70.937	**1,55E-08**	$\gamma = 0,1$
	MMP12	15	**Hapmap25580-BTA-149003**	4.712.494	**0**	**1,03E-05**	$\gamma = 0,001$
	FXYD2	15	BTA-96062-no-rs	26.718.365	192.835	9,50E-01	$\gamma = 0,001$
	PEX14	16	**BFGL-NGS-114895**	39.941.474	38.874	**3,10E-03**	$\gamma = 0,1$
	LAMB3	16	**ARS-BFGL-NGS-19358**	71.717.864	114.636	**3,52E-02**	$\gamma = 0,001$
	G0S2	16	**ARS-BFGL-NGS-19358**	71.717.864	86.306	**3,52E-02**	$\gamma = 0,001$
	SLC6A2	18	ARS-BFGL-NGS-2512	23.411.398	61.502	3,45E-01	$\gamma = 0,001$
	DPEP3	18	ARS-BFGL-NGS-63965	34.317.638	157.462	3,46E-01	$\gamma = 0,01$
	CEACAM21	18	**BTA-43712-no-rs**	51.991.765	172.688	**4,51E-03**	Linear
	TP53	19	ARS-BFGL-BAC-36717	27.779.765	99.935	9,11E-01	$\gamma = 0,001$
	DAP	20	ARS-BFGL-NGS-71821	66.310.836	90.336	1,50E-01	$\gamma = 0,001$
	ISG20	21	**ARS-BFGL-NGS-28897**	19.433.975	192.375	**2,78E-03**	$\gamma = 0,001$
	CMTM8	22	**BTB-01641930**	6.674.746	114.639	**5,12E-03**	$\gamma = 0,1$
	ATP2B2	22	**ARS-BFGL-NGS-106287**	54.759.837	235.794	**3,96E-02**	$\gamma = 0,01$
	ACLY	23	**BTA-107490-no-rs**	43.489.318	75.518	**1,98E-02**	$\gamma = 0,1$
	ELOVL5	23	**ARS-BFGL-NGS-100422**	26.023.095	83.995	**1,63E-03**	$\gamma = 0,01$
	OSBPL1A	24	ARS-BFGL-NGS-10333	33.512.785	122.915	4,79E-01	$\gamma = 0,001$
	FHOD3	24	ARS-BFGL-NGS-30844	21.592.657	143.957	7,59E-01	$\gamma = 0,001$
	LMAN1	24	Hapmap28681-BTA-150829	60.836.565	165.465	3,27E-01	$\gamma = 0,001$
	FAM20C	25	**UA-IFASA-6068**	42.538.054	155.898	**2,35E-05**	$\gamma = 0,001$
	KIFF11	26	**ARS-BFGL-NGS-95979**	14.720.748	241.548	**1,50E-02**	$\gamma = 0,1$
	ALDH18A1	26	**ARS-BFGL-NGS-17995**	17.434.772	32.528	**5,60E-04**	$\gamma = 0,001$
	MKI67	26	ARS-BFGL-NGS-3095	48.363.642	77.242	6,90E-02	$\gamma = 0,001$
	CAPN6	29	**ARS-BFGL-NGS-18010**	37.709.030	222.030	**1,18E-03**	$\gamma = 0,1$

Tabela 8.53 Genes candidatos a partir de estudos anteriores de associação do leite para *Bos taurus* segundo a base de dados de Ogorevc et al. (2009) marcados pelos SNPs selecionados pelo SMS com raio de 250.000 pb.

Referência	Gene	Cr.	SNP mais próximo	Posição (pb)	Distância (pb)	Valor-p corrigido do SNP	Kernel
Weikard et al. (2005)	PPARGC1A	6	ARS-BFGL-NGS-101567	44.820.065	**0**	7,44E-01	$\gamma = 0,001$
Sanders et al. (2006)	CSN1S2	6	**BTB-01025945**	87.134.718	127.739	**8,23E-03**	$\gamma = 0,001$
Kamiński et al. (2008)	GHR	20	**ARS-BFGL-NGS-76756**	34.124.209	**0**	**5,25E-06**	$\gamma = 0,1$
Khatib, Heifetz e Dekkers (2005)	SERPINA1	21	**Hapmap38841-BTA-52789**	59.421.036	101.136	**5,81E-03**	$\gamma = 0,01$
Brym, Kamiński e Wójcik (2004)	PRL	23	**ARS-BFGL-NGS-108416**	35.659.109	71.309	**9,74E-06**	Linear
Nascimento et al. (2006)	Bola-DRB3	23	ARS-BFGL-NGS-105633	25.522.560	45.675	9,24E-01	$\gamma = 0,001$

Na Tabela 8.54, estão indicados os 17 genes candidatos de estudos anteriores sobre expressão da mastite em gado de leite marcados pelos 16 SNPs selecionados pelo SMS. Desse total de SNPs, 12 apresentaram significância estatística (valor-p $< 0{,}05$). O *kernel* radial com $\gamma = 0,001$ selecionou oito SNPs, o radial com $\gamma = 0,01$ apenas dois SNPs, o radial com $\gamma = 0,1$ sete SNPs e o linear não selecionou marcador algum.

A solução do SMS encontrou dois SNPs que marcam os genes candidatos em estudos anteriores de associação da mastite conforme é apresentado na Tabela 8.55. Os dois genes, IL8RB e IL8RA, estão localizados no cromossomo 2 e são marcados pelo SNP Hapmap53684-rs29025784, o qual possui distâncias em relação aos genes IL8RB e IL8RA

Tabela 8.54 Genes candidatos a partir de estudos anteriores sobre expressão de mastite para *Bos taurus* marcados pelos SNPs selecionados pelo SMS em um raio de 250.000 pb. Dados organizados por Ogorevc et al. (2009).

Referência	Gene	Cr.	SNP mais próximo	Posição (pb)	Distância (pb)	Valor-p corrigido do SNP	Kernel
Zheng, Watson e Kerr (2006)	ETS2	1	**ARS-BFGL-NGS-100423**	154.601.847	**0**	**3,29E-05**	$\gamma = 0,1$
Zheng, Watson e Kerr (2006)	C3	7	ARS-BFGL-NGS-3043	16.628.598	247.598	5,54E-02	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	MAP2K7	7	**ARS-BFGL-NGS-10921**	15.307.310	179.310	**6,69E-04**	$\gamma = 0,001$
Long et al. (2001)	MMP9	13	**ARS-BFGL-NGS-109252**	75.514.523	**0**	**5,64E-04**	$\gamma = 0,01$
Schwerin et al. (2003)	AHCY	13	**ARS-BFGL-NGS-95285**	64.328.410	9.370	**3,00E-05**	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	UCP3	15	ARS-BFGL-BAC-19403	52.908.088	188.312	3,63E-01	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	SELP	16	**Hapmap40684-BTA-16356**	34.441.775	193.225	**4,78E-03**	$\gamma = 0,1$
Goldammer et al. (2004)	TLR2	17	**Hapmap54135-rs29019936**	4.196.626	74.974	**3,98E-03**	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	HP	18	**ARS-BFGL-BAC-31823**	38.124.438	4.062	**2,02E-03**	$\gamma = 0,1$
Zheng, Watson e Kerr (2006)	RELB	18	ARS-BFGL-NGS-37312	52.361.585	98.415	2,75E-01	$\gamma = 0,001$
Pareek et al. (2005)	CCL5	19	**UA-IFASA-7392**	14.075.672	92.972	**1,21E-04**	$\gamma = 0,1$
Schwerin et al. (2003)	TP53	19	ARS-BFGL-BAC-36717	27.779.765	90.635	9,11E-01	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	CALM2	20	**BTB-00771463**	8.591.775	81.505	**2,21E-03**	$\gamma = 0,01$
Zheng, Watson e Kerr (2006)	NFKBIA	21	**ARS-BFGL-NGS-4716**	46.446.647	347	**9,40E-05**	$\gamma = 0,1$
Zheng, Watson e Kerr (2006)	YES1	24	**BTB-01542456**	36.877.049	**0**	**3,29E-04**	$\gamma = 0,001$
Zheng, Watson e Kerr (2006)	SAA2	29	**ARS-BFGL-NGS-12494**	26.618.727	77.999	**4,26E-03**	$\gamma = 0,1$
Zheng, Watson e Kerr (2006)	SAA1	29	**ARS-BFGL-NGS-12494**	26.618.727	49.310	**4,26E-03**	$\gamma = 0,1$

menores que 250.000 pb (raio adotado), mas valor-p não-significativo igual a 2,12E-01.

Tabela 8.55 Genes candidatos a partir de estudos anteriores sobre associação de mastite para *Bos taurus* marcados pelos SNPs selecionados pelo SMS em um raio de 250.000 pb. Dados organizados por Ogorevc et al. (2009).

Referência	Gene	Cr.	SNP mais próximo	Posição (pb)	Distância (pb)	Valor-p corrigido do SNP	Kernel
Youngerman et al. (2004)	IL8RB	2	Hapmap53684-rs29025784	110.555.673	24.727	2,12E-01	$\gamma = 0,001$
Rambeaud e Pighetti (2007)	IL8RA	2	Hapmap53684-rs29025784	110.555.673	59.127	2,12E-01	$\gamma = 0,001$

Por causa de suas inúmeras vantagens como grande quantidade de mutações e técnicas eficientes para mutagênese[4] induzida, precisamente descritas por mudanças fenotípicas, o modelo do camundongo foi usado como uma ferramenta para a identificação da relação genótipo-fenótipo (OGOREVC et al., 2009). A disponibilidade da sequência completa do genoma do camundongo possibilita comparações com outras espécies, logo, a identificação de regiões conservadas (GUÉNET, 2005). Segundo Ogorevc et al. (2009), exitem, atualmente, 143 genes que, quando mutados ou expressos como transgenes em camundongo, resultam em fenótipos associados com glândula mamária.

As informações sobre genes candidatos a partir de experiências com animais foram recuperadas a partir da base de dados *Mouse Genome Informatics* (MGI) (EPPIG et al.,

[4]Mutagênesis são mutações induzidas em células embrionárias (KUHN; WURST, 2014).

2015) cujo endereço eletrônico é <http://www.informatics.jax.org> e as mesmas foram organizadas por Ogorevc et al. (2009). Um total de 29 genes identificados em estudos de *knock-outs*[5] e transgenes[6] para o modelo do camundongo foram marcados por SNPs selecionados pelo SMS e estão evidenciados na Tabela 8.56.

Tabela 8.56 Genes candidatos a partir de estudos anteriores sobre o modelo animal do camundongo para *Bos taurus* marcados pelos SNPs selecionados pelo SMS em um raio de 250.000 pb. Dados organizados por Ogorevc et al. (2009).

Referência	Gene	Cr.	SNP mais próximo	Posição (pb)	Distância (pb)	Valor-p corrigido do SNP	Kernel
Eppig et al. (2015)	ETS2	1	ARS-BFGL-NGS-100423	154.601.847	**0**	3,29E-05	$\gamma = 0,1$
	INHBB	2	ARS-BFGL-NGS-38374	75.644.963	41.163	8,67E-01	$\gamma = 0,001$
	GLI2	2	ARS-BFGL-NGS-38374	75.644.963	215.037	8,67E-01	$\gamma = 0,001$
	PAX3	2	**BTA-49040-no-rs**	114.725.429	235.571	**6,73E-05**	$\gamma = 0,01$
	LEPR	3	**ARS-BFGL-NGS-44176**	85.321.007	238.993	**1,09E-05**	Linear
	GLI3	4	**Hapmap40292-BTA-71565**	81.438.152	141.848	**3,07E-06**	$\gamma = 0,1$
	LEF1	6	**Hapmap1572-BTA-111558**	18.432.155	**0**	**2,21E-07**	$\gamma = 0,1$
	RBPJ	6	**BTB-01201475**	47.378.342	86.342	**1,02E-03**	$\gamma = 0,1$
	CSF1R	7	**ARS-BFGL-NGS-54377**	60.886.865	**0**	**4,38E-02**	$\gamma = 0,01$
	JAK2	8	**ARS-BFGL-NGS-102255**	41.420.729	76.271	**1,37E-02**	$\gamma = 0,001$
	SLC30A4	10	**ARS-BFGL-NGS-108454**	66.709.459	97.459	**6,76E-05**	Linear
	NCOA3	13	**BFGL-NGS-119536**	77.220.639	129.639	**2,91E-06**	Linear
	NCOA2	14	BFGL-NGS-113395	34.087.387	**0**	9,76E-01	$\gamma = 0,001$
	NCOA2	14	**Hapmap41185-BTA-122014**	34.134.717	**0**	**3,69E-03**	Linear
	FOG-2	14	**Hapmap23721-BTC-014248**	57.804.264	74.264	**1,41E-03**	Linear
	CDH1	18	**BFGL-NGS-117369**	34.866.493	147.507	**2,82E-07**	$\gamma = 0,1$
	CDH3	18	**BFGL-NGS-117369**	34.866.493	89.507	**2,82E-07**	$\gamma = 0,1$
	TP53	19	ARS-BFGL-BAC-36717	27.779.765	90.635	9,11E-01	$\gamma = 0,001$
	MSX2	20	**BTB-00769424**	6.737.837	125.037	**5,21E-03**	$\gamma = 0,01$
	PRLR	20	**ARS-BFGL-BAC-33668**	41.487.453	**0**	**1,83E-05**	$\gamma = 0,001$
	OXTR	22	**BTA-53738-no-rs**	18.159.460	118.940	**2,03E-03**	$\gamma = 0,001$
	ATP2B2	22	BFGL-NGS-111253	55.638.951	**0**	5,27E-02	$\gamma = 0,001$
	ATP2B2	22	**ARS-BFGL-NGS-123**	55.969.491	**0**	**4,64E-02**	Linear
	PPARG	22	ARS-BFGL-NGS-10654	58.370.932	**0**	2,84E-01	$\gamma = 0,1$
	PRL	23	**ARS-BFGL-NGS-108416**	35.659.109	71.309	**9,74E-06**	Linear
	VGF	25	**BTA-60128-no-rs**	37.825.957	115.057	**3,13E-04**	Linear
	PTEN	26	**BTA-96427-no-rs**	10.036.302	**0**	**1,77E-02**	$\gamma = 0,001$
	FGFR2	26	ARS-BFGL-NGS-33804	41.800.926	183.074	8,68E-02	$\gamma = 0,01$
	NRG3	28	Hapmap33583-BTA-140120	35.570.103	**0**	8,67E-02	$\gamma = 0,001$
	NRG3	28	ARS-BFGL-NGS-58598	35.915.039	**0**	2,01E-01	$\gamma = 0,001$
	NRG3	28	BTB-00990667	38.418.258	**0**	6,41E-01	$\gamma = 0,001$
	PPYR1	28	BFGL-NGS-110675	41.638.678	151.522	5,59E-01	$\gamma = 0,001$
	CCND1	29	**ARS-BFGL-NGS-101178**	48.657.629	76.171	**5,92E-05**	$\gamma = 0,1$

O gene EEF1D no cromossomo 14, que está localizado entre as posições 2.313.946 e 2.326.721 pb, foi identificado pelo SMS e de acordo com um estudo de Xie et al. (2014), esse gene exibe significantemente a maior expressão em glândula mamária como em outros tecidos para gado de leite. Esse gene também foi identificado por Jiang et al. (2010) e Frąszczak e Szyda (2015), sendo que ambos usaram modelos monoatributos e modelos lienares mistos para selecionar os SNPs mais informativos para identificar genes candidatos.

O método do valor p-corrigido selecionou 130 SNPs, os quais identificaram 834 genes a partir de um raio de 250.000 pb. Alguns desses genes são classificados como genes candidatos para produção de leite e para mastite em trabalhos anteriores e foram

[5]Gene *knockout* é uma técnica genética em que um dos genes de um organismo é tornardo inoperante (TYMMS, 2001).

[6]Um transgene é um gene que tenha sido transferido naturalmente, ou por técnicas de engenharia genética, a partir de um organismo para outro. A introdução de um transgene tem o potencial de alterar o fenótipo de um organismo (TYMMS, 2001).

identificados também pelo SMS, indicados em negrito nas Tabelas 8.52, 8.53, 8.54, 8.55 e 8.56. É importante destacar que o número de SNPs pertencentes ao conjunto interseção dos métodos SMS e valor-p corrigido, identificaram um número muito superior de genes candidatos de estudos anteriores em relação ao número de QTLs associados ao leite como pode ser observado pelos SNPs em negrito nas Tabelas 8.40 sobre QTLs e nas Tabelas 8.52, 8.53, 8.54, 8.55 e 8.56 sobre genes identificados pelo SMS.

A interseção dos quatro subconjuntos de SNPs selecionados por cada *kernel* possui somente seis SNPs, os quais identificaram ao todo 18 genes, conforme Tabela 8.57. É importante salientar que nenhum SNP está no mesmo cromossomo e que somente o marcador ARS-BFGL-NGS-105427 possui valor-p corrigido menor que 0,05, além do marcador BTA-31100-no-rs que não foi atribuído a nenhum dos 30 cromossomos do *Bos taurus*, o que é indicado pelo cromossomo 99.

Tabela 8.57 SNPs selecionados simultaneamente pelos quatro *kernels* avaliados pelo SMS juntamente com os genes candidatos com raio de 250.000 pb.

SNP	Cr.	Posição	Valor-p bruto	Valor-p corrigido	Genes
ARS-BFGL-NGS-105427	3	112.656.223	4,33E-07	**9,89E-03**	CSMD2, HMGB4, LOC100140781 [a]
BTB-01367046	4	112.658.643	3,22E-06	7,36E-02	LOC100337375, CUL1, EZH2
ARS-BFGL-NGS-95285	13	64.328.410	3,00E-05	6,85E-01	ASIP, AHCY, ITCH, DYNLRB1, MAP1LC3A, LOC782136, PIGU
ARS-BFGL-NGS-42430	19	31.015.710	4,79E-06	1,09E-01	SHISA6, DNAH9, LOC50663
BTB-01767954	26	2.615.180	7,68E-03	1,75E+02	LOC100138749[a], ZWINT
BTA-31100-no-rs	99	NA	2,58E-03	5,88E+01	-

[a] Estes registros foram retirados pelo NCBI porque os modelos em que se basearam não foram previstos em anotações posteriores. Para o gene LOC100140781 ver <http://www.ncbi.nlm.nih.gov/gene/?term=LOC100140781> e para o gene LOC100138749 ver <http://www.ncbi.nlm.nih.gov/gene/100138749>.

A forma mais natural de indicar os SNPs mais informativos selecionados pelo SMS é avaliar os marcadores pertencentes à interseção dos quatro *kernels*, haja vista que o SVR não fornece diretamente os efeitos dos SNPs. Portanto, os genes marcados pelos seis SNPs da Tabela 8.57 devem ter preferência em estudos posteriores sobre genes potenciais para produção de leite. Todavia, além do conjunto interseção, sugere-se a avaliação posterior de todos os SNPs pertencentes ao conjunto união dos SNPs selecionados pelos quatro *kernels*.

8.10.2.3 Comparativo com Trabalhos Correlatos

Frąszczak e Szyda (2015) usaram quatro abordagens para a seleção de SNPs significantes em um estudo de associação em escala genômica, com 2.601 touros da raça Holandês e

46.267 SNPs para duas características complexas: produção de gordura no leite e produção de leite. Os métodos de seleção usados foram baseados em um modelo de um único SNP (M1), um modelo de um único SNP e um efeito poligênico aleatório (M2), um modelo de regressão CAR *score* proposto por Zuber e Strimmer (2011) e um modelo SNP-BLUP com efeitos aleatórios de todos os SNPs ajustados simultaneamente (M4) discutido em Szyda et al. (2011). Cada modelo foi ajustado para cada fenótipo, perfazendo um total de oito modelos. Os autores não utilizaram a seleção do modelo M1, pois o mesmo demonstra problemas metodológicos e selecionou um número muito superior de SNPs em relação aos modelos M2, M3 e M4. Neste trabalho, foram selecionados 40 SNPs pertencentes aos modelos M2, M3 e M4 simultaneamente, que identificaram 24 genes candidatos, onde todos estão localizados no cromossomo 14. Dos 40 marcadores, 32 estão localizados dentro de algum gene e, os outros oito SNPs tem distância ao gene mais próximo variando entre 71 a 4.472 pb. Entretanto, Frąszczak e Szyda (2015) ressaltaram que alguns genes podem representar associações espúrias resultantes do alto desequilíbrio de ligação com o gene DGAT1, mas aqueles que são localizados em regiões mais distantes do cromossomo 14 são potenciais genes candidatos.

Jiang et al. (2010) realizou um estudo de associação com 2.093 filhas de 14 touros da raça Holandês, sendo 54.001 SNPs genotipados para cada animal, e cinco fenótipos associados à produção e composição do leite: produção de leite, produção de gordura, produção de proteína, percentagem de gordura e percentagem de proteína. Para cada fenótipo, foram construídos dois modelos de seleção de atributos, um denominado de teste de desequilíbrio de transmissão (do inglês,*transmission disequilibrium test*) (KOLBEHDARI et al., 2006) e o outro, modelo misto de análise de regressão de um único *locus*. No total foram selecionados 105 SNPs, os quais identificaram sete QTLs relacionados à produção de leite, 6 QTLs à gordura do leite, 8 QTLs à proteína do leite, oito QTLs à percentagem de gordura no leite e seis QTLs à percentagem de proteína no leite. Em relação aos genes, foram identificados 105 genes associados às quatro características, cujas distâncias aos SNPs selecionados variaram entre 0 a 560.215 pb. Os 105 SNPs selecionados estão distribuídos ao longo dos cromossomos 1, 2, 3, 5, 6, 8, 9, 11, 14, 20, 26 e X, isto é, não foram identificados SNPs nos 18 cromossomos restantes.

O SMS selecionou 1.265 SNPs a partir de um conjunto de dados composto de 240 touros da raça Gir e 56.947 SNPs, sendo identificados 245 QTLs associados ao leite em

54 categorias mostradas na Tabela 8.38 e 90 genes candidatos identificados em estudos de associação anteriores. Somente nos cromossomos 10, 17 e 21, o SMS não identificou QTL algum, porém, em todos os 30 cromossomos foram identificados potenciais genes candidatos. Assim, apesar de apenas 240 touros serem usados no procedimento de seleção do SMS, o mesmo obteve uma seleção mais ampla e mais distribuída ao longo dos cromossomos que as realizadas em Jiang et al. (2010) e Frąszczak e Szyda (2015). A explicação possível para esse fato é que adotou-se a união dos quatro *kernels* usados no SMS, enquanto que em Frąszczak e Szyda (2015) foi adotada a interseção entre três modelos de seleção e em Jiang et al. (2010) foi usada a correção de Bonferroni para múltiplos testes nos dois modelos de seleção, o que restringe severamente a seleção de SNPs. Como as metodologias de validação dos SNPs selecionados são diferentes na busca de regiões genômicas informativas, uma comparação direta é inviável, mas é possível perceber que o SMS capturou mais regiões genômicas potenciais e os outros dois trabalhos demonstraram maior acurácia na identificação de QTLs e genes.

8.10.2.4 Genes Candidatos Não-identificados pelo SMS

No intuito de mostrar que o SMS deixou de identificar genes candidatos associados a produção de leite que foram detectados em vários trabahos anteriores, escolheram-se os genes OPN e o DGAT1. Essa análise permite evidenciar pontos restritivos da metodologia de identificação adotada com raio de 250.000 pb em torno da posição do SNP analisado ou devido ao tamanho reduzido da amostra para GWAS (240 touros Gir), reduzindo o poder de detecção do SMS. Outro ponto importante a ser destacado é que muitos SNPs que marcam os genes não identificados foram excluídos pelos filtros usados no controle de qualidade, logo, a eficiência do SMS precisa ser avaliada posteriormente em estudos futuros sem o uso desses filtros.

O gene da OPN (osteopontina), conhecido como SSP1 (do inglês, *Secreted Phosphoprotein 1*) está localizado no cromossomo 6 de bovinos no intervalo entre 38.120.576 e 38.179.866 medido em pares de bases (NCBI) e está em uma região próxima a um QTL para produção de leite <http://www.cnpgl.embrapa.br/sistemaproducao/4122-genes-candidatos>. Dois alelos foram identificados para esse gene: o alelo C, associado com o aumento na porcentagem de proteína e gordura no leite e o alelo T, associado a um aumento na produção de leite (KHATIB et al., 2007). Estudos

de associação deste gene com características de produção de leite foram realizadas nos trabalhos de Sheehy et al. (2009) e Khatib et al. (2007). No presente trabalho, esse gene foi identificado pelo marcador ARS-BFGL-NGS-90128 selecionado pelo método do valor-p bruto com limite superior de corte 0,05, sendo a distância do marcador ao gene igual a 90.128 pb, a qual está dentro do raio de 250.000 pb. Nem o método do valor-p corrigido e nem o SMS foram capazes de selecionar pelo menos um marcador dentro de um raio de 250.000 pb, entretanto, o SNP mais próximo do gene OPN, selecionado pelo SMS, é o Hapmap44513-BTA-107931 e a distância entre eles é 1.432.457 pb. Assim, flexibilizando o raio da vizinhança para 1.500.000 pb, esse gene seria identificado pelo SMS, contudo, esse raio pode não ser adequado dado a possibilidade de baixo desequilíbrio de ligação entre os SNPs da região em análise.

O gene da DGAT1 (diacilglicerol-aciltransferase 1) está localizado no cromossomo 14 de bovinos no intervalo entre 1.795.425 e 1.804.838 medido em pares de bases (NCBI). Diversos estudos anteriores mostram que o gene DGAT1, que pertence ao cromossomo 14 e está numa região de íntron, possui associação com a produção e composição do leite. Os trabalhos de Grisart et al. (2002), Grisart et al. (2004) e Winter et al. (2002) demonstram que variantes de substituição não-conservativas alteram o teor de gordura do leite e produção de leite. Outro estudo feito por Mach et al. (2012) demonstrou o efeito pleiotrópico do DGAT1 sobre o metabolismo de energia e sobre o sistema imunológico. Neste trabalho, esse gene foi identificado pelo marcador ARS-BFGL-NGS-70821 selecionado pelo método do valor-p bruto com limite superior de corte 0,05, sendo a distância do marcador ao gene igual a 90.128 pb, a qual está dentro do raio de 250.000 pb. Isto mostra que nem o método do valor-p corrigido e nem o SMS foram capazes de selecionar pelo menos um marcador dentro de um raio de 250.000 pb. Entretanto, o SNP mais próximo do gene DGAT1, selecionado pelo SMS, é o ARS-BFGL-NGS-70821 (posição 2.260.066 pb) e a distância entre eles é 455.228 pb. Assim, flexibilizando o raio da vizinhança para 500.000 pb, esse gene seria identificado pelo SMS.

8.11 Considerações Finais

Para os dados simulados, o SMS selecionou SNPs causais com efeitos marginais não-significativos (valores-p maiores que 0,05), mas cujas interações com outros SNPs causais

eram relevantes para a explicação do fenótipo. Nos dados reais, o SMS selecionou vários SNPs não significativos estatisticamente pelo valor-p corrigido, entretanto, informativos do ponto de vista biológico, porque marcaram regiões com QTLs e genes relacionados ao fenótipo da PTA do leite ou à características correlacionados com esse fenótipo. Uma possível explicação é que esses SNPs não contribuem isoladamente para explicar o fenótipo, mas quando são considerados em conjunto, o poder explicativo conjunto é detectado por um ou mais *kernels* do SVM/SVR.

O SMS detectou SNPs muito próximos aos oito QTLs simulados no QTLMAS 2011, entretanto, mais de um SNP foi selecionado por QTL, significando que o SMS não selecionou um subconjunto sem redundância. Outro ponto importante a considerar é o número elevado de falsos-positivos selecionados pelo SMS em relação aos outros métodos, o que mostra a necessidade de melhoria nas etapas do SMS para reduzir a redundância e a quantidade de SNPs não-causais. Entretanto, cabe ressaltar que o critério para classificar o SNP como não-causal no QTLMAS 2011 foi conservador, visto que existem blocos haplótipos com 15 cM de comprimento, usou-se apenas 9 cM como limite, o que gera um número maior de SNPs não-informativos.

Como a PTA do leite é calculada com base em um modelo linear, ou seja, considera que os efeitos dos QTLs associados à produção de leite são aditivos, o *kernel* linear ou *kernels* radiais com γ próximos a zero deveriam apresentar o melhor desempenho. Essa espectativa se comprovou pelas médias e pelos desvios-padrões das correlações desses *kernels* demonstradas na Tabela 8.35. Outro possível desdobramento da linearidade no cálculo da PTA do leite é a ausência de interação ou a presença de interação entre SNPs com pequenos efeitos sobre o fenótipo nos dados reais, porém, isso é mera especulação que precisa ser comprovada posteriormente.

A magnitude dos efeitos isolados dos SNPs nos dados reais podem ser observadas indiretamente pelo grande decréscimo do MSE baseado no SVR sobre o *rank* da RF na etapa de corte do SMS. Isso sugere que os coeficientes dos SNPs causais usados nas simulações do SCRIME e do QTLMAS 2011 são muito inferiores aos dos SNPs considerados informativos pelo SMS no conjunto de dados reais. Esse fato pode explicar o fato do SMS2 reduzir percentualmente o conjunto inicial de SNPs em relação ao subconjunto final, dado pela união dos subconjuntos gerados por cada *kernel*, em 87%, 90%, 81%, 83%, 65%, 64%, 97,47% e 94,47% para os dados das simulações 1, 2, 3, 4, 5,

6, QTMAS 2011 e para os dados reais respectivamente. Ou seja, quanto maior os efeitos isolados e de interações entre SNPs, melhor é o resultado da seleção pelo SMS, resultando numa maior redução percentual do número de SNPs em relação ao conjunto inicial.

Apesar do SMS ser um procedimento de seleção estocástico, devido à RF e ao GA, o mesmo possui um comportamento estável como visto nos testes nos conjuntos simulados do SCRIME e do QTLMAS 2011, com exceção do modelo da simulação 5 com interação de ordem 4, pois o SMS demonstrou instabilidade na seleção dos SNPs causais. Essa instabilidade pode ter ocorrido pelo efeito usado para a interação não ser suficiente para ser captado pelo SMS.

Em todos os conjuntos de dados avaliados, as populações do GA, as quais são compostas por subconjuntos de SNPs, demonstraram baixa diversidade, pois o desempenho dos melhores subconjuntos foram próximos à média da população, o que pode ser notado em todos os gráficos relacionados è evolução do GA. Isso pode ter ocorrido devido ao uso do método de seleção baseado em *wrapper* ou devido ao alto desequilíbrio de ligação entre os SNPs selecionados na etapa de corte para os dados reais e do QTLMAS 2011 ou ao efeito conjunto de ambos.

9 Conclusões

Neste trabalho, foram desenvolvidas estratégias para detecção de SNPs visando a identificação de efeitos aditivos e possíveis interações entre SNPs em cenários com LD entre os marcadores. Com relação aos dados sintéticos, nenhum dos simuladores adotados (SCRIME e LDSO do QTLMAS 2011) utilizam em seus processos ferramentas de RF, SVM/SVR ou GA para geração dos dados de genótipo e fenótipo. Os modelos subjacentes para a construção dos dados nesses simuladores são distintos das técnicas usadas pelo SMS, evitando assim o crime de inversão neste estudo.

A estrutura proposta pelo SMS na segunda versão composta pelas etapas de relevância (RF), corte (SVM/SVR + RF) e refinamento (GA+SVM/SVR) mostrou ser eficiente na identificação de SNPs causais em cenários simulados tanto para fenótipos contínuos quanto para binários. A segunda versão mostrou-se superior à primeira versão, a qual foi baseada no *rank* do valor-p bruto da correlação de Spearman.

O uso de uma abordagem multi-*kernel* que captura diferentes relações entre o genótipo e o fenótipo apresentou eficiência para os conjuntos de dados 1, 2, 3 e 4 simulados pelo SCRIME e para os dados simulados pelo LDSO da competição do QTLMAS 2011. Uma explicação é a possibilidade de várias estruturas de relacionamento entre um grupo de genes e o fenótipo em questão em dados reais, podendo estas serem simultaneamente relações lineares e não-lineares de baixa e alta ordens.

A busca pelo melhor γ no *kernel* radial para a seleção de SNPs com o mínimo de falsos-positivos e o máximo de verdadeiros-positivos é uma tarefa computacionalmente custosa na atual versão do SMS. Além disso, dada a complexidade esperada da relação entre genótipo e fenótipo é provável que não exista um único *kernel* que consiga detectar todos os SNPs informativos para a característica fenotípica avaliada. Logo, a solução usada para os problemas de regressão foi o conjunto união dos marcadores selecionados por cada *kernel* para aumentar o número de SNPs causais selecionados, já que esse é o objetivo primário, mesmo que esta união aumente a probabilidade de selecionar SNPs não-causais.

Uma questão complexa é a escolha da medida usada para avaliar os subconjuntos de marcadores pelo GA do SMS, pois nenhuma das métricas avaliadas (correlação, MSE e

MAPE) conseguiu mostrar que o subconjunto formado apenas pelos SNPs informativos apresenta desempenho superior a qualquer outro conjunto composto tanto por SNPs causais quanto não-causais nos conjuntos de dados simulados do SCRIME. Ou seja, sempre subconjuntos com SNPs causais e não-causais são melhores avaliados que os subconjuntos somente com SNPs causais por todas as métricas adotadas. Além disso, a melhor seleção de SNPs para um dado *kernel* é a que melhor se adapta à relação matemática entre o genótipo e o fenótipo, e não pelo *kernel* que apresenta a maior correlação entre os valores fenotípicos observados e os preditos.

Isso é um indicativo que a métrica adotada para avaliação dos subconjuntos ainda precisa ser melhorada ou até mesmo substituída. Logo, o conjunto união dos subconjuntos selecionados por cada *kernel* é uma solução mais robusta em relação a se contornar esse obstáculo, pois a maior correlação só é usada para escolher a melhor solução dentre as selecionadas por um *kernel* específico, mas não para comparar as correlações entre *kernels* distintos. Essa estratégia minimiza a perda de SNPs informativos, mas pode aumentar o número de SNPs não-informativos.

A determinação do subconjunto de SNPs selecionados pelo SMS apresenta como vantagem a possibilidade de se criar uma lista de prioridade para os SNPs a serem analisados. Assim, considerando quatro *kernels* usados no SMS, o subconjunto inicial poderia ser a interseção dos subconjuntos de marcadores dos quatro *kernels* avaliados, em seguida, as interseções entre três *kernels*, depois, as interseções entre dois *kernels* e, finalmente, a união de todos os *kernels*. Considerar as soluções do SMS sob esse prisma, cria níveis descendentes em redundâncias e ascendentes em quantidades de SNPs selecionados pelo SMS para estudos posteriores, pois o número de elementos dos subconjuntos sugeridos aumentam conforme a operação de pertinência entre subconjuntos seja flexibilizada.

O SMS demonstrou melhor desempenho geral em relação aos métodos do valor-p bruto, do valor-p corrigido e do Blasso tanto nos dados simulados quanto nos dados reais, onde o critério adotado foi o maior número de verdadeiros-positivos selecionados. O SMS pode ser usado em conjunto com o método do valor-p corrigido, já que eles demonstraram significativa similaridade nos dados reais e o valor-p possui uma demanda por recursos computacionais inferior ao SMS. Isso é uma forma de antecipar alguns dos resultados do SMS para o pesquisador, o qual poderá começar suas análises pelos SNPs selecionados

pelo valor-p corrigido enquanto aguarda a solução final do SMS. Outra possibilidade é a realização da união ou interseção entre os subconjuntos de SNPs selecionados pelo valor-p corrigido e pelo SMS, o que pode ampliar o poder de detecção do SMS em conjuntos de dados em que o valor-p corrigido apresente uma acurácia maior na seleção de SNPs.

A característica de se obter falsos-positivos identificados pelo SMS pode ser contornada pela otimização dos parâmetros da RF e do GA, pois o SMS foi utilizado sem um ajuste refinado dos mesmos para possibilitar uma melhor compreensão de suas restrições e de suas potencialidades. Por outro lado, o SMS demonstrou que pode selecionar marcadores que não foram selecionados por outros métodos, apesar de indicarem informação biológica relacionada com o fenótipo da PTA do leite.

O método SMS apresentou resultados promissores para encontrar interações entre pares e trios de SNPs em dados simulados, porém, para quádruplas de marcadores, a RF não demonstrou desempenho satisfatório para os parâmetros e efeitos adotados. Ademais, para conjunto de dados com efeitos aditivos e não-aditivos, o SMS também demonstrou resultados adequados tanto em cenários com baixo LD (conjuntos simulados pelo SCRIME) quanto para cenários com alto LD (QTLMAS 2011).

Apesar do SMS não produzir o efeito de cada SNP ao final da seleção, apresentou eficiência principalmente nos dados do QTLMAS 2011, identificando SNPs muito próximos aos QTLs, e no conjunto de dados reais da Embrapa. Implicitamente, o SMS capturou os efeitos dos SNPs mais relevantes nos dados reais, pois os mesmos possuem, em suas vizinhanças, QTLs e/ou genes associados à produção de leite , os quais foram identificados por outras técnicas em estudos anteriores.

O SMS pode ser usado com variáveis genotípicas (SNPs) e variáveis ambientais contínuas ou categóricas, pois as técnicas RF, SVM/SVR e GA permitem essa possibilidade sem qualquer perda de informação. Todavia, estudos nessa linha precisam ser realizados para verificar a real adaptabilidade do modelo.

Uma característica benéfica do SMS é sua modularidade, o que permite o uso de diferentes técnicas computacionais ou estatísticas nas três etapas denominadas de relevância, corte e refinamento. Permite-se, até mesmo, a união de abordagens distintas para ampliar o poder de detecção de SNPs informativos do SMS. Como exemplo, pode-se citar uma possível união do valor-p bruto, do *rank* da RF e do Blasso para realizar a fase de relevância visando possibilitar que SNPs relevantes que não foram ordenados

adequadamente por uma abordagem, mas possam ter sido por outra, passem para a fase de refinamento, aumentando a probabilidade de serem selecionados ao final do processo.

10 Trabalhos Futuros

A medida de importância de cada marcador dada pela RF é usada como filtro inicial para ordenar os SNPs mais promissores na explicação do fenótipo. Essa métrica permite, até certo ponto, capturar algum nível de interação entre os marcadores, devido à própria estrutura das árvores construídas na floresta. Entretanto, para interações entre SNPs com efeito conjunto significante, porém com efeitos marginais pequenos o suficiente para serem superados por efeitos marginais espúrios de marcadores falso-positivos, apresentam grande probabilidade de não serem adequadamente ranqueados pela RF. Consequentemente, é de suma importância construir medidas adequadas para que a RF consiga medir a importância conjunta de pares de variáveis como é realizado por Bureau et al. (2005) para problemas de classificação em GWAS.

Um controle de qualidade, alternativo e mais flexível do que o realizado no conjunto de dados reais, baseado em análise de componentes principais e clusterização, proposto por Pongpanich, Sullivan e Tzeng (2010), pode apresentar benefícios na manipulação principalmente dos dados reais. Outro ponto a considerar é o uso da RF como uma alternativa para imputação de valores ausentes para os SNPs como realizado por Schwarz et al. (2009). A RF gera uma medida de similaridade que pode ser usada para imputar genótipos de SNPs com valores ausentes.

Em relação à função de aptidão do GA, é necessário um estudo comparativo entre o MSE e a correlação de Pearson em vários conjuntos de dados, pois quando minimiza-se o MSE no GA, na base de dados real, o SMS tende a reduzir o número de marcadores selecionados, mas também aumenta a correlação média final (resultados não apresentados). Quando maximiza-se a correlação média de Pearson no GA, em relação ao conjunto real, o MSE médio do subconjunto selecionado não reduz na mesma proporção que a correlação média no caso anterior. Essa observação mostra que quando o MSE é o objetivo do GA, a quantidade de SNPs selecionada é inferior à quantidade de SNPs da solução do GA quando a correlação média é usada como função de aptidão. Portanto, é importante analisar a quantidade e qualidade dos SNPs selecionados pelo SMS nas métricas do MSE e correlação de Pearson para avaliar o impacto das mesmas em diferentes conjuntos de dados.

O ponto mais crítico do SMS é o custo computacional da fase de treinamento do SVM/SVR com 10-*fold*, pois o mesmo cresce proporcionalmente com o número de indivíduos. Este custo computacional é intensificado no GA, quando o SVM/SVR tem que ser aplicado em populações com número maior de indivíduos para o treinamento e teste necessário no cálculo de suas aptidões em cada geração do GA. Assim, como foi visto no conjunto de dados do QTLMAS2011, onde o conjunto de treinamento possuía 2.000 indivíduos com um número de marcadores SNP, relativamente pequeno para GWAS, igual a 9.990, o tempo de processamento foi muito superior quando comparado com os métodos do valor-p e do Blasso. Portanto, uma possibilidade de melhoria no desempenho do SVM/SVR é o uso de processamento paralelo com o uso de unidades gráficas de processamento (do inglês, *graphics processing units* - GPU). O pacote *rpud* implementa o SVM/SVR e pode reduzir em até 10 vezes o tempo de treinamento dessa técnica (YAU, 2012). Outra melhoria que pode ser adotada em conjunto com o processamento paralelo por GPU é o uso de parâmetros C e ϵ do SVR, determinados analiticamente em função do conjunto de treinamento conforme descrito por Cherkassky e Ma (2004).

Outra melhoria seria o uso do PUK (*Pearson Universal kernel*) que pode apresentar resultados equivalentes ou melhores que o *kernel* radial em diversos conjuntos de dados como discutido por Ünstü, Melssen e Buydens (2006) e em estudos de associação em escala genômica como esplanado por Oliveira et al. (2014b) na primeira versão do SMS (SMS1). A determinação dos parâmetros adequados para melhorar o desempenho do SVR com PUK pode ser feita como sugere Ünstü, Melssen e Buydens (2005). Outros *kernels* podem ser avaliados como o polinomial (ordem 2, 3 ou 4), o *kernel* sigmóide, o *kernel* tangente hiperbólico, *kernel* baseado na função de Bessel e o kernel de base radial ANOVA que tem bom desempenho para problemas de regressão multidimensionais (KARATZOGLOU et al., 2004).

A similaridade genética medida pelo complemento da distância de Rogers modificada (ou MDR)(WRIGHT, 1978; GOODMAN; STUBER, 1983) tem sido usada para construir *kernels* que não contém parâmetros como em Maenhout et al. (2007). Isso oferece uma grande vantagem computacional, pois o ajuste de parâmetros é evitado (LONG et al., 2011). Os autores avaliaram ϵ-SVR no desempenho de predição do fenótipo de híbridos não testados em uma base de dados real de milho. Usou-se o MDR para calcular a dissimilaridade (d_{kl}) entre híbridos k e l, e computar seu complemento (s_{kl}) como elemento

(k, l) do *kernel*, onde $s_{kl} = 1 - d_{kl}$. A Expressão 10.1 apresenta o cálculo de s_{kl} e de d_{kl}, onde s é o número de *loci*, n_i é o número de alelos para o *locus* i e p_{ij}^k, p_{ij}^l representam a frequência alélica do jth alelo no *locus* i para os híbridos k e l, respectivamente. Maenhout et al. (2007) comparou a predição do *kernel* s_{kl} com o *kernel* radial e obtiveram resultados similares. Como sugere Long et al. (2011), que *kernels* específicos podem ser boas alternativas para *kernels* comumente usados, além de possuir significado biológico.

$$d_{kl} = \frac{1}{\sqrt{2s}} \sqrt{\sum_{i=1}^{s} \sum_{j=1}^{n_i} (p_{ij}^k - p_{ij}^l)^2} \tag{10.1}$$

No GA, pode ser usada uma nova configuração que busca a seleção de atributos simultaneamente com a determinação de forma adaptativa da parametrização de modelos, levando a obtenção de melhores subconjuntos de atributos. Para ser realizada tal tarefa, os cromossomos do GA teriam que capturar tanto o subconjunto de marcadores quanto os parâmetros do SVM/SVR a serem otimizados. A parte referente às variáveis seriam codificadas como binárias e a parte restante, referente aos parâmetros (C, γ e ϵ para o *kernel* radial) do SVM/SVR, seriam codificadas como variáveis reais. Em conjunto às possibilidades anteriores, outra opção é a variação do *kernel* por meio da matriz *kernel* ao invés da execução de toda a otimização necessária do modelo dual para encontrar os vetores suporte durante a fase de treinamento do SVM/SVR. Esse procedimento conjunto do GA mais o SVR é detalhado em Perolini (2012).

Como o método baseado no valor-p bruto selecionou o gene DGAT1, que é um gene candidato para a produção de leite em gado, e selecionou o SNP6 no conjunto simulado do SCRIME do modelo 1, o qual não foi selecionado pelo SMS e nem mesmo pela RF em nenhuma das 10 execuções do SMS, sugere-se que seja realizado duas fases de corte: uma com o *rank* da RF e outra com o valor-p bruto. Ao final o conjunto união dos dois cortes por cromossomo será inserido no GA para a fase de refinamento. Acredita-se que essa estratégia possa permitir a entrada de SNPs que não foram selecinados pelo *rank* da RF. Todavia, antes de realizar essa análise, precisa-se otimizar os parâmetros da RF para verificar se a não selecão do SNP6 e do gene DGAT1 é devido ao método de ordenação subjacente à RF ou aos parâmetros fixos e não-otimizados *mtry* (número de variáveis selecionado para cada nó por árvore) e *ntree* (número de árvores na floresta usados na RF).

Pretende-se que o SMS seja aplicado em outros conjuntos de dados simulados que apresentem outras complexidades para relação genótipo-fenótipo, como por exemplo, nos dados do QTLMAS 2008, 2009, 2010 e 2012. Além desses conjuntos de dados simulados, podem ser gerados novos dados com os simuladores *QTL Cartographer* (BASTEN et al., 2004) (<http://statgen.ncsu.edu/qtlcart/>) e *QMSim* (SARGOLZAEI; SCHENKEL, 2009) (<http://www.aps.uoguelph.ca/~msargol/qmsim/>). O simulador do *QTL Cartographer* permite gerar diversos tipos de relações entre o genótipo e o fenótipo tais como aditiva, aditiva com dominância, somente epistasia, epistasia com dominância e, em casos extremos de complexidade, relações com, simultaneamente, efeitos aditivos, epistáticos e dominantes. Em relação ao QMSim, sua construção foi projetada para simular dados de genotipagem em larga escala com múltiplos e complexos *pedigrees* de animais (SARGOLZAEI; SCHENKEL, 2009). Conforme Sargolzaei e Schenkel (2009), o QMSim é um simulador de base familiar, que também pode levar em conta as características evolutivas pré-definidas, como LD, mutação, os estrangulamentos e expansões. Com o uso do QMSim, o SMS pode ser avaliado em diversos cenários com herdabilidades distintas e estruturas complexas de *pedigrees*.

A base de dados *Kyoto Encyclopedia of Genes and Genome* (KEGG) (KANEHISA; GOTO, 2000) deverá ser usada para uma melhor compreensão dos processos biológicos envolvidos nos genes identificados pelos SNPs selecionados por meio do SMS. Outra avaliação biológica possível é atrelar os estudos sobre arranjos de expressão gênica de tecidos relevantes aos estudos de associação em escala genômica a fim de facilitar a interpretação biológica (OLAZAR, 2013). Essa etapa de validação funcional é de suma importância para o aumento da compreensão da dinâmica biológica do genótipo na determinação do fenótipo.

Como o SMS é um método de seleção que não define a forma da interação e nem a quantidade de SNPs interagindo, ele pode ser usado como um método inicial, e depois, pode-se usar métodos que busquem somente por interações entre pares de SNPs como o método construído por Olazar (2013). O método de busca exaustiva criado por Neto (2013) para detectar interação entre trios de SNPs também pode ser aplicado sobre o subconjunto de SNPs gerado pelo SMS. Da mesma forma, o SMS pode ser usado como método de seleção inicial para o *adaptive mixed* LASSO elaborado por Wang, Eskridge e Crossa (2011), que estima os coeficientes de todos pares de SNPs que podem

interagir. Essa abordagem pode ser útil, pois o número de combinações dois a dois tomados no conjunto inicial de SNPs cresce rapidamente quando o número de SNPs aumenta, podendo tornar o *adaptive mixed* LASSO inviável na presença de milhares de SNPs. Essa possibilidade de aplicação se deve pela redução considerável, feita pelo SMS, do conjunto inicial de SNPs para o subconjunto final de SNPs, composto na sua maior parte de marcadores informativos.

No caso de fenótipos contínuos, a técnica de programação genética pode ser aplicada ao conjunto de SNPs selecionados pelo SMS para identificar a função matemática que mais se aproxima da relação real entre o genótipo e o fenótipo. Para fenótipos binários, a programação genética poderia ser usada para criar árvores baseada em operadores lógicos sobre os SNPs para compreensão mais profunda sobre a variação do fenótipo, como realizado em Nunkesser et al. (2007). Isso poderia permitir extrapolar regras construídas por técnicas de aprendizado de máquina para melhor entendimento do processo biológico subjacente.

Outra opção é usar o SMS para selecionar os SNPs informativos e, posteriormente, utilizá-los na predição do valor genético genômico, com o intuito de aumentar o poder preditivo de técnicas empregadas na seleção genômica de animais e plantas em programas de melhoramento genético. Como demonstram Weigel et al. (2009) e Morser, Hayes e Raadsma (2010), estudos nessa direção já tem sido desenvolvidos explorando a possibilidade de uso de *chips* de genotipagem de SNPs com baixa densidade, logo, com custo reduzido e simultaneamente com aumento da acurácia na predição do fenótipo de interesse.

Outro forma de utilização do SMS é a seleção de marcadores SNP a partir da técnica denominada meta-análise, a qual é baseada em vários estudos de escala genômica para uma mesma característica fenotípica. Essa abordagem aumenta a amostra e, simultaneamente, reduz o custo e o tempo do estudo.

Os genes relacionados à produção de leite compilados no trabalho de Singh et al. (2014) podem ser usados como referência para os SNPs selecionados pelo SMS. Com isso, poderiam ser encontrados outros genes relacionados à produção de leite que não foram indicados por Ogorevc et al. (2009), o que permite refinar a validação dos resultados do SMS.

Uma questão relevante é o fato de que o potencial genético de um animal é resumido

como seu valor genético estimado, o qual é derivado a partir do seu próprio desempenho, bem como do desempenho de indivíduos relacionados (parentes) (EKINE et al., 2014). Ekine et al. (2014) mostraram que a informação de todos os parentes medidos é a principal fonte de falsos-positivos em GWAS em um estudo de simulação. Logo, o estudo de associação em escala genômica com o fenótipo medido pela produção de leite bruta e individual de vacas da raça Gir deverá ser realizado para comparação com o estudo desenvolvido no presente trabalho, o qual foi baseado na PTA do leite de touros Gir, onde essa caracterítica é uma medida indireta da produção de leite da prole feminina de cada touro, a fim de verificar a possível diminuição de seleção de SNPs falsos-positivos pelo SMS e por outros métodos de seleção.

REFERÊNCIAS

ALBERTS, B. et al. *Biologia molecular da célula*. 5. ed. Porto Alegre, BR: Artmed, 2010. 1396 p. ISBN 978-85-363-2066-3.

AMUNDADOTTIR, L. et al. Genome-wide association study identifies variants in the abo locus associated with susceptibility to pancreatic cancer. *Nature genetics*, Nature Publishing Group, v. 41, n. 9, p. 986–990, 2009.

ARBEX, W. et al. *Modelos computacionais para estabelecimento de meios e procedimentos metodológicos para análise de dados em bioinformática – MCBio*. Juiz de Fora, 2010. 36 p. Projeto de pesquisa e desenvolvimento submetido à Chamada 03/2010 da Embrapa no âmbito do Macroprograma 5 - Desenvolvimento Institucional.

ARBEX, W. A. *Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino*. Tese (Doutorado) — COPPE - UFRJ, 2009.

ASHWELL, M.; TASSELL, C. V. Detection of putative loci affecting milk, health, and type traits in a us holstein population using 70 microsatellite markers in a genome scan. *Journal of dairy science*, Elsevier, v. 82, n. 11, p. 2497–2502, 1999.

AWAD, A. et al. Confirmation and refinement of a qtl on bta5 affecting milk production traits in the fleckvieh dual purpose cattle breed. *Animal genetics*, Wiley Online Library, v. 41, n. 1, p. 1–11, 2010.

BAKER, J. E. Adaptive selection methods for genetic algorithms. In: HILLSDALE, NEW JERSEY. *Proceedings of an International Conference on Genetic Algorithms and their applications*. [S.l.], 1985. p. 101–111.

BALA, A.; CHANA, I. A survey of various workflow scheduling algorithms in cloud environment. In: *2nd National Conference on Information and Communication Technology (NCICT)*. [S.l.: s.n.], 2011. p. 26–30.

BAN, H.-J. et al. Identification of type 2 diabetes-associated combination of snps using support vector machine. *BMC Genetics*, v. 11, p. 11–26, 2010.

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

BARTONOVA, P. et al. Association between csn3 and bco2 gene polymorphisms and milk performance traits in the czech fleckvieh cattle breed. *Genet Mol Res*, v. 11, n. 2, p. 1058–1063, 2012.

BASTEN, C. J. et al. Qtl cartographer, version 1.17. *Department of Statistics, North Carolina State University, Raleigh, NC*, 2004.

BATESON, W. *Mendel's principles of heredity*. [S.l.]: University Press, 1909.

BEN-HUR, A.; WESTON, J. A user's guide to support vector machines. Citeseer, 2007.

BENNEWITZ, J. et al. Combined analysis of data from two granddaughter designs: A simple strategy for qtl confirmation and increasing experimental power in dairy cattle. *Genetics Selection Evolution*, Paris: Elsevier, c1989-, v. 35, n. 3, p. 319–338, 2003.

BENNEWITZ, J. et al. Multiple quantitative trait loci mapping with cofactors and application of alternative variants of the false discovery rate in an enlarged granddaughter design. *Genetics*, Genetics Soc America, v. 168, n. 2, p. 1019–1027, 2004.

BIAU, D. J.; JOLLES, B. M.; PORCHER, R. P value and the theory of hypothesis testing: An explanation for new researchers. *Clinical Orthopaedics and Related Research®*, Springer, v. 468, n. 3, p. 885–892, 2010.

BLACK, P. E. *Dictionary of algorithms and data structures*. [S.l.]: National Institute of Standards and Technology, 2004.

BOICHARD, D. et al. Detection of genes influencing economic traits in three french dairy cattle breeds. *Genetics Selection Evolution*, Paris: Elsevier, c1989-, v. 35, n. 1, p. 77–102, 2003.

BONAKDAR, E. et al. Igf-i gene polymorphism, but not its blood concentration, is associated with milk fat and protein in holstein dairy cows. *Genetics and Molecular Research*, v. 9, n. 3, p. 1726–1734, 2010.

BOUWMAN, A. C. et al. Fine mapping of a quantitative trait locus for bovine milk fat composition on bos taurus autosome 19. *Journal of dairy science*, Elsevier, v. 97, n. 2, p. 1139–1149, 2014.

BOVENHUIS, H.; WELLER, J. I. Mapping and analysis of dairy cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. *Genetics*, Genetics Soc America, v. 137, n. 1, p. 267–280, 1994.

BREIMAN, L. Bagging Predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996. Disponível em: <http://citeseer.ist.psu.edu/breiman96bagging.html>.

BREIMAN, L. Bias, variance, and arcing classifiers. Tech. Rep. 460, Statistics Department, University of California, Berkeley, CA, USA, 1996.

BREIMAN, L. *Out-of-bag estimation*. [S.l.], 1996.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CRC press, 1984.

BROOKES, A. J. The essence of snps. *Gene*, v. 234, n. 2, p. 177 – 186, 1999. ISSN 0378-1119. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037811199900219X>.

BRYM, P.; KAMIŃSKI, S.; WÓJCIK, E. Nucleotide sequence polymorphism within exon 4 of the bovine prolactin gene and its associations with milk performance traits. *Journal of Applied Genetics*, v. 46, n. 2, p. 179–185, 2004.

BUREAU, A. et al. Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, v. 28, n. 2, p. 171–182, 2005. Disponível em: <http://online.liebertpub.com/doi/abs/10.1089/cmb.1995.2.275>.

BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, v. 8, n. 12, p. 1–11, 2012. Disponível em: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822>.

CAETANO, A. R. Marcadores snp: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. *Revista Brasileira de Zootecnia*, Sociedade Brasileira de Zootecnia, v. 38, n. 8, p. 64–71, 2009.

CAMPOS, G. D. L. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, Genetics Soc America, v. 182, n. 1, p. 375–385, 2009.

CAMPOS, G. de los; PÉREZ, P. Blr: bayesian linear regression. r package version 1.2. *Institute for Statistics and Mathematics, Wien, Austria. http://cran. r-project. org/web/packages/BLR/index. html (accessed 26 Apr. 2011)*, 2010.

CANTOR, R. M.; LANGE, K.; SINSHEIMER, J. S. Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, Elsevier, v. 86, n. 1, p. 6–22, 2010.

CARLBORG, Ö.; HALEY, C. S. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, Nature Publishing Group, v. 5, n. 8, p. 618–625, 2004.

CASTELLANO, P. *Validació creuada*. 2014. <http://ca.wikipedia.org/wiki/Validaci%C3%B3_creuada>. Acessado em: 27/04/2015.

CATLETT, J. On changing continuous attributes into ordered discrete attributes. In: SPRINGER. *Machine learning—EWSL-91*. [S.l.], 1991. p. 164–178.

CECCHINATO, A. et al. Short communication: Effects of β-lactoglobulin, stearoyl-coenzyme a desaturase 1, and sterol regulatory element binding protein gene allelic variants on milk production, composition, acidity, and coagulation properties of brown swiss cows. *Journal of dairy science*, Elsevier, v. 95, n. 1, p. 450–454, 2012.

CHAMBERLAIN, A. et al. Validation of single nucleotide polymorphisms associated with milk production traits in dairy cattle. *Journal of dairy science*, Elsevier, v. 95, n. 2, p. 864–875, 2012.

CHANDRASEKHAR, A.; RAGHUVEER, K. Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers. In: IEEE. *Computer Communication and Informatics (ICCCI), 2013 International Conference on*. [S.l.], 2013. p. 1–7.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 2, n. 3, p. 27, 2011.

CHEN, R. et al. Polymorphisms of the il8 gene correlate with milking traits, scs and mrna level in chinese holstein. *Molecular biology reports*, Springer, v. 38, n. 6, p. 4083–4088, 2011.

CHERKASSKY, V.; MA, Y. Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, v. 17, p. 113–126, 2004. Disponível em: <http://www.ece.umn.edu/users/cherkass/N2002-SI-SVM-13-whole.pdf>.

CLEMPSON, A. et al. Evidence that leptin genotype is associated with fertility, growth, and milk production in holstein cows. *Journal of dairy science*, Elsevier, v. 94, n. 7, p. 3618–3628, 2011.

COHEN-ZINDER, M. et al. Identification of a missense mutation in the bovine abcg2 gene with a major effect on the qtl on chromosome 6 affecting milk yield and composition in holstein cattle. *Genome Research*, Cold Spring Harbor Lab, v. 15, n. 7, p. 936–944, 2005.

CONGDON, C. B. *A comparison of genetic algorithms and other machine learning systems on a complex classification task from common disease research*. Tese (Doutorado) — The University of Michigan, 1995.

CORDELL, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, Oxford Univ Press, v. 11, n. 20, p. 2463–2468, 2002.

CORNFIELD, J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, v. 11, n. 6, p. 1269–1275, 1951.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, v. 20, n. 3, p. 273–297, 1995.

COSGUN, N. A. L. E.; DUARTE, C. W. High-dimensional pharmacogenetic prediction of a continuos trait using machine learning tecniques with application to warfarin dose prediction in african americans. *Bioinformatics*, v. 27, n. 10, p. 1384–1389, 2011.

CRISTIANINI, N.; SHAWE-TAYLOR, J. *An introduction to support vector machines*. [S.l.]: Cambridge University Press Cambridge, 2000.

DAETWYLER, H. D. et al. A genome scan to detect quantitative trait loci for economically important traits in holstein cattle using two methods and a dense single nucleotide polymorphism map. *Journal of dairy science*, Elsevier, v. 91, n. 8, p. 3225–3236, 2008.

DASH, M.; LIU, H. Consistency-based search in feature selection. *Artificial intelligence*, Elsevier, v. 151, n. 1, p. 155–176, 2003.

DASHAB, G. R. et al. Comparison of linear mixed model analysis and genealogy-based haplotype clustering with a bayesian approach for association mapping in a pedigreed population. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S4.

DEMEURE, O. et al. Comparison of the analyses of the xv^{th} qtlmas common dataset ii: Qtl analysis. *BMC Proceedings*, v. 6, n. 2, p. 1–5, 2012. Disponível em: <http://www.biomedcentral.com/1753-6561/6/S2/S2>.

DEVLIN B E RISCH, N. A comparação das medidas de desequilíbrio de ligação para mapeamento em escala fina. v. 29, n. 2, p. 311–322, 1995.

DIMITRIADOU, E. et al. Package 'e1071'. *R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html*, 2009.

DRUCKER, H. et al. Support vector regression machines. *Advances in neural information processing systems*, Morgan Kaufmann Publishers, v. 9, p. 155–161, 1997.

EASTON, D. F.; EELES, R. A. Genome-wide association studies in cancer. *Human Molecular Genetics*, v. 17, p. 109–115, 2008.

EASTON, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, Nature Publishing Group, v. 447, n. 7148, p. 1087–1093, 2007.

EKINE, C. C. et al. Why breeding values estimated using familial data should not be used for genome-wide association studies. *G3: Genes| Genomes| Genetics*, Genetics Society of America, v. 4, n. 2, p. 341–347, 2014.

ELSEN, J.-M. et al. XVth QTLMAS: simulated dataset. *BMC Proceedings*, v. 6, n. 2, p. 1–5, 2012. Disponível em: <http://www.biomedcentral.com/1753-6561/6/S2/S1>.

EPPIG, J. et al. *Mouse Genome Database (MGD) at the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine*. 2015. <http://www.informatics.jax.org>. Acessado em: 30/07/2015.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2011.

FALCONER, D. S.; MACKAY, T. F.; FRANKHAM, R. Introduction to quantitative genetics (4th edn). *Trends in Genetics*, [Amsterdam, The Netherlands: Elsevier Science Publishers (Biomedical Division)], c1985-, v. 12, n. 7, p. 280, 1996.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FISHER, R. A. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, Cambridge Univ Press, v. 52, n. 02, p. 399–433, 1919.

FONTANESI, L. et al. A candidate gene association study for nine economically important traits in italian holstein cattle. *Animal genetics*, Wiley Online Library, v. 45, n. 4, p. 576–580, 2014.

FOULKES, A. S. *Applied statistical genetics with R: for population-based association studies*. [S.l.]: Springer, 2009. ISBN 0387895531.

FRANKEL, W. N.; SCHORK, N. J. Who's afraid of epistasis? *Nature genetics*, Nature Publishing Group, v. 14, n. 4, p. 371–373, 1996.

FRĄSZCZAK, M.; SZYDA, J. Comparison of significant single nucleotide polymorphisms selections in gwas for complex traits. *Journal of applied genetics*, Springer, p. 1–7, 2015.

FREEMAN, S.; HERRON, J. C. *Análise evolutiva*. [S.l.]: Artmed, 2009.

FREITAS, A. A. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, Springer, v. 16, n. 3, p. 177–199, 2001.

FREYER, G. et al. Multiple qtl on chromosome six in dairy cattle affecting yield and content traits. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 119, n. 2, p. 69–82, 2002.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

FRIEDMAN, J. H.; KOHAVI, R.; YUN, Y. Lazy decision trees. In: *AAAI/IAAI, Vol. 1*. [S.l.: s.n.], 1996. p. 717–724.

FU, W.-X. et al. Genome-wide association analyses of the 15th qtl-mas workshop data using mixed model based single locus regression analysis. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S5.

GABRIEL, S. B. et al. The structure of haplotype blocks in the human genome. *Science*, American Association for the Advancement of Science, v. 296, n. 5576, p. 2225–2229, 2002.

GEER, L. Y. et al. The ncbi biosystems database. *Nucleic acids research*, Oxford Univ Press, p. gkp858, 2009.

GENOME, B. C. *Cattle Genome UMD3.1*. 2015. Disponível em: <http://www.animalgenome.org/cgi-bin/gbrowse/bovine/>.

GEORGES, M. et al. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*, Genetics Soc America, v. 139, n. 2, p. 907–920, 1995.

GIANOLA, D. et al. Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. *BMC genetics*, BioMed Central Ltd, v. 12, n. 1, p. 87, 2011.

GIANOLA, D.; PEREZ-ENCISO, M.; TORO, M. A. On marker–assisted prediction of genetic value: beyond the ridge. *Genetics*, v. 163, n. 1, p. 445–455, 2003.

GLAZOV, E. A. et al. Repertoire of bovine mirna and mirna-like small regulatory rnas expressed upon viral infection. *PloS one*, Public Library of Science, v. 4, n. 7, p. e6349, 2009.

GOLDAMMER, T. et al. Mastitis increases mammary mrna abundance of β-defensin 5, toll-like-receptor 2 (tlr2), and tlr4 but not tlr9 in cattle. *Clinical and diagnostic laboratory immunology*, Am Soc Microbiol, v. 11, n. 1, p. 174–185, 2004.

GOLDBERG, D. E. *Genetic algorithms in search, optimization and machine learning*. [S.l.]: Addison-Wesley, 1989. ISBN 0201157675.

GOLDSTEIN, B. A. et al. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics*, BioMed Central Ltd, v. 11, n. 1, p. 49+, 2010. ISSN 1471-2156. Disponível em: <http://dx.doi.org/10.1186/1471-2156-11-49>.

GOLDSTEIN, B. A.; POLLEY, E. C.; BRIGGS, F. Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, v. 10, n. 1, p. 1–34, 2011.

GONDRO, C.; WERF, J. Van der; HAYES, B. *Genome-wide association studies and genomic prediction*. [S.l.]: Humana Press, 2013.

GOODMAN, M. M.; STUBER, C. W. Races of maize. 6: Isozyme variation among races of maize in bolivia. *Maydica*, v. 28, p. 169–187, 1983.

GRAVETTER, F.; WALLNAU, L. *Essentials of statistics for the behavioral sciences*. [S.l.]: Cengage Learning, 2013.

GRIFFITHS-JONES, S. et al. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic acids research*, Oxford Univ Press, v. 34, n. suppl 1, p. D140–D144, 2006.

GRISART, B. et al. Positional candidate cloning of a qtl in dairy cattle: identification of a missense mutation in the bovine dgat1 gene with major effect on milk yield and composition. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 2, p. 222–231, 2002.

GRISART, B. et al. Genetic and functional confirmation of the causality of the dgat1 k232a quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, v. 101, n. 8, p. 2398–2403, 2004.

GUÉNET, J. L. The mouse genome. *Genome Research*, Cold Spring Harbor Lab, v. 15, n. 12, p. 1729–1740, 2005.

GUJARATI, D. *Econometria básica*. 4. ed. [S.l.]: Elsevier, 2006.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, JMLR. org, v. 3, p. 1157–1182, 2003. Disponível em: <www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HAMON, J. *Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale*. Tese (Doutorado) — Université des Sciences et Technologie de Lille-Lille I, 2013.

HARDER, B. et al. Mapping of quantitative trait loci for lactation persistency traits in german holstein dairy cattle. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 123, n. 2, p. 89–96, 2006.

HARHAY, G. P. et al. Characterization of 954 bovine full-cds cdna sequences. *BMC genomics*, BioMed Central Ltd, v. 6, n. 1, p. 166, 2005.

HARRIS, B.; JOHNSON, D. The impact of high density snp chips on genomic evaluation in dairy cattle. *Interbull Bulletin*, p. 40, 2010.

HARTL, D. L.; CLARK, A. G. *Principles of population genetics*. 4. ed. [S.l.]: Sinauer associates Sunderland, 2006. 659 p.

HE, Y. et al. Association of bovine cd4 and stat5b single nucleotide polymorphisms with somatic cell scores and milk production traits in chinese holsteins. *Journal of dairy research*, Cambridge Univ Press, v. 78, n. 02, p. 242–249, 2011.

HECK, J. et al. Effects of milk protein variants on the protein composition of bovine milk. *Journal of dairy science*, Elsevier, v. 92, n. 3, p. 1192–1202, 2009.

HEROLD, C. et al. Intersnp: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, Oxford Univ Press, v. 25, n. 24, p. 3275–3281, 2009.

HIGA, R. H. et al. Estudo de associação genômica ampla utilizando *Random Forest*: estudo de caso em bovinos de corte. In: _____. *Talking About Computing and Genomics (TACG): vol.I: Modelos e Métodos Computacionais em Bioinformática*. [S.l.]: Embrapa, 2014. cap. 3, p. 71–99. ISBN 978-85-7035-382-5.

HUANG, W. et al. Association between milk protein gene variants and protein composition traits in dairy cattle. *Journal of dairy science*, Elsevier, v. 95, n. 1, p. 440–449, 2012.

İLHAN, İ.; TEZEL, G. A genetic algorithm–support vector machine method with parameter optimization for selecting the tag snps. *Journal of biomedical informatics*, Elsevier, v. 46, n. 2, p. 328–340, 2013.

ILLUMINA. *"TOP/BOT"Strand and "A/B"Alele*. [S.l.], 2014. Disponível em: <http://res.illumina.com/documents/products/technotes/technote_topbot.pdf>.

ISHIWATA, H. et al. Characterization of gene expression profiles in early bovine pregnancy using a custom cdna microarray. *Molecular reproduction and development*, Wiley Online Library, v. 65, n. 1, p. 9–18, 2003.

JEFFREYS, A. J.; KAUPPI, L.; NEUMANN, R. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature genetics*, Nature Publishing Group, v. 29, n. 2, p. 217–222, 2001.

JIANG, L. et al. Genome wide association studies for milk production traits in chinese holstein population. *PloS one*, v. 5, n. 10, p. e13661, 2010.

JIANG, R. et al. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC bioinformatics*, BioMed Central Ltd, v. 10, n. Suppl 1, p. S65, 2009.

JOHN, G. H. et al. Irrelevant features and the subset selection problem. In: *Machine Learning: Proceedings of the Eleventh International Conference*. [S.l.: s.n.], 1994. p. 121–129.

JONES, G. M.; BAILEY, T. L. Understanding the basics of mastitis. Virginia Cooperative Extension, 2009.

KAMIŃSKI, S. et al. Towards an integrated approach to study snps and expression of candidate genes associated with milk protein biosynthesis. *Russian Journal of Genetics*, Springer, v. 44, n. 4, p. 459–465, 2008.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford Univ Press, v. 28, n. 1, p. 27–30, 2000.

KARARGYRIS, A.; BOURBAKIS, N. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *Biomedical Engineering, IEEE Transactions on*, IEEE, v. 58, n. 10, p. 2777–2786, 2011.

KARATZOGLOU, A.; SMOLA, A.; HORNIK, K. kernlab – an s4 package for kernel methods in r. *Journal Statistical Software*, v. 11, n. 9, p. 1–20, 2004.

KARATZOGLOU, A. et al. kernlab-an s4 package for kernel methods in r. Institut für Statistik und Mathematik, WU Vienna University of Economics and Business, 2004.

KHATIB, H.; HEIFETZ, E.; DEKKERS, J. Association of the protease inhibitor gene with production traits in holstein dairy cattle. *Journal of dairy science*, Elsevier, v. 88, n. 3, p. 1208–1213, 2005.

KHATIB, H. et al. The association of bovine ppargc1a and opn genes with milk composition in two independent holstein cattle populations. *Journal of dairy science*, Elsevier, v. 90, n. 6, p. 2966–2970, 2007.

KIM, J. et al. Snp selection in genome-wide association studies via penalized support vector machine with max test. *Hindawi*, Hindawi Publishing Corporation, v. 2013, p. 1–8, 2013. Disponível em: <http://www.hindawi.com/journals/cmmm/2013/340678/>.

KINGSMORE, S. F. et al. Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews Drug Discovery*, Nature Publishing Group, v. 7, n. 3, p. 221–230, 2008.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1-55860-363-8. Disponível em: <http://dl.acm.org/citation.cfm?id=1643031.1643047>.

KOHAVI, R.; JOHN, G. The wrapper approach. In: _____. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. [S.l.]: Kluwer Academic Publishers, 1998. cap. 3, p. 33–47. ISBN 978-1-4613-7622-4.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, p. 273–324, 1997.

KOHAVI, R.; KUNZ, C. Option decision trees with majority votes. In: CITESEER. *ICML*. [S.l.], 1997. v. 97, p. 161–169.

KOLBEHDARI, D. et al. Transmission disequilibrium test for quantitative trait loci detection in livestock populations. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 123, n. 3, p. 191–197, 2006.

KUHN, R.; WURST, W. *Gene knockout protocols*. [S.l.]: Springer, 2014.

LAIRD, N. M.; LANGE, C. *The fundamentals of modern statistical genetics*. [S.l.]: Springer, 2011. 223 p. (Statistics for Biology and Health).

LAL, T. N. et al. Embedded methods. In: *Feature extraction*. [S.l.]: Springer, 2006. p. 137–165.

LEYVA-BACA, I. et al. Identification of single nucleotide polymorphisms in the bovine ccl2, il8, ccr2 and il8ra genes and their association with health and production in canadian holsteins. *Animal genetics*, Wiley Online Library, v. 38, n. 3, p. 198–202, 2007.

LIAW, A. et al. Package "randomforest". *Retrieved December*, v. 12, p. 2009, 2009.

LINDEN, R. *Algoritmos Geneticos (2a edicao)*. [S.l.]: Brasport, 2008.

LINK, W. A.; EATON, M. J. On thinning of chains in mcmc. *Methods in Ecology and Evolution*, Wiley Online Library, v. 3, n. 1, p. 112–115, 2012.

LIPSCHUTZ, S. *Probabilidade*. [S.l.]: McGraw-Hill, 1972.

LIU, F. et al. Automated fiber type specific cross-sectional area assessment and myonuclei counting in skeletal muscle. *Journal of Applied Physiology*, Am Physiological Soc, 2013.

LONG, E. et al. Escherichia coli induces apoptosis and proliferation of mammary cells. *Cell death and differentiation*, v. 8, n. 8, p. 808–816, 2001.

LONG, N. et al. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theoretical and applied genetics*, Springer, v. 123, n. 7, p. 1065–1074, 2011.

MA, J.; SONG, A.; XIAO, J. A robust static decoupling algorithm for 3-axis force sensors based on coupling error model and ε-svr. *Sensors*, Molecular Diversity Preservation International, v. 12, n. 11, p. 14537–14555, 2012.

MA, L. et al. *epiSNP: A computer package of serial computing programs for epistasis testing in genome-wide association studies, user manual version 2.0*. 2008.

MACH, N. et al. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (dgat1) in the mammary gland tissue of dairy cows. *Journal of dairy science*, Elsevier, v. 95, n. 9, p. 4989–5000, 2012.

MAENHOUT, S. et al. Support vector machine regression for the prediction of maize hybrid performance. *Theoretical and Applied Genetics*, Springer, v. 115, n. 7, p. 1003–1013, 2007.

MARQUES, E. et al. Identification of candidate markers on bovine chromosome 14 (bta14) under milk production trait quantitative trait loci in holstein. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 128, n. 4, p. 305–313, 2011.

MARTINEZ, M. *Técnicas de biologia molecular aplicadas à produção animal*. [S.l.]: Mimeo, 1998.

MARTINS, G. d. A. *Estatística geral e aplicada*. [S.l.]: Atlas, 2002.

MAXA, J. et al. Genome-wide association mapping of milk production traits in braunvieh cattle. *Journal of dairy science*, Elsevier, v. 95, n. 9, p. 5357–5364, 2012.

MCKAY, S. D. et al. Whole genome linkage disequilibrium maps in cattle. *BMC genetics*, BioMed Central Ltd, v. 8, n. 1, p. 74, 2007.

MEI, G. et al. Fine mapping quantitative trait loci affecting milk production traits on bovine chromosome 6 in a chinese holstein population. *Journal of Genetics and Genomics*, Elsevier, v. 36, n. 11, p. 653–660, 2009.

MENG, Y. A. et al. Performance of random forest when snps are in linkage disequilibrium. *BMC bioinformatics*, BioMed Central Ltd, v. 10, n. 1, p. 78, 2009.

MEREDITH, B. K. et al. Genome-wide associations for milk production and somatic cell score in holstein-friesian cattle in ireland. *BMC genetics*, BioMed Central Ltd, v. 13, n. 1, p. 21, 2012.

MEUWISSEN, T. et al. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, Genetics Soc America, v. 157, n. 4, p. 1819–1829, 2001.

MEYER, D.; WIEN, F. T. Support vector machines. *The Interface to libsvm in package e1071*, 2014.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013.

MILLER, D. J. et al. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics*, Oxford Univ Press, v. 25, n. 19, p. 2478–2485, 2009.

MITCHELL, M. *An introduction to genetic algorithms*. [S.l.]: MIT press, 1998.

MITTAG, F. et al. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. *Human Mutation*, v. 33, n. 12, p. 1708–1718, 2012.

MOKRY, F. B. et al. Genome-wide association study for backfat thickness in canchim beef cattle using random forest approach. *BMC genetics*, BioMed Central Ltd, v. 14, n. 1, p. 47, 2013. Disponível em: <http://www.biomedcentral.com/1471-2156/14/47>.

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

MOORE, J. H. et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology*, Elsevier, v. 241, n. 2, p. 252–261, 2006.

MOORE, J. H.; RITCHIE, M. D. The challenges of whole-genome approaches to common diseases. *Jama*, American Medical Association, v. 291, n. 13, p. 1642–1643, 2004.

MOORE, J. H.; WILLIAMS, S. M. New strategies for identifying gene-gene interactions in hypertension. *Annals of medicine*, Informa UK Ltd UK, v. 34, n. 2, p. 88–95, 2002.

MOORE, J. H.; WILLIAMS, S. M. Epistasis and its implications for personal genetics. *Am. J, Hum. Genet.*, v. 85, p. 309–320, 2009.

MOORE, J. H.; WILLIAMS, S. M. *Epistasis: Methods and Protocols*. [S.l.]: Springer New York, 2015.

MORSER, G.; HAYES, B. J.; RAADSMA, H. W. Accuracy of direct genomic values in holstein bulls and cows using subsets of snp markers. *Generics Selection Evolution*, v. 42, n. 37, p. 1–15, 2010.

MORSER, G. et al. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. *Generics Selection Evolution*, v. 41, n. 1, p. 41–56, 2009.

MOTSINGER-REIF, A. A. et al. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC research notes*, BioMed Central Ltd, v. 1, n. 1, p. 65, 2008.

MUCHA, S. et al. Comparison of analyses of the qtlmas xiv common dataset. ii: Qtl analysis. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2011. v. 5, n. Suppl 3, p. S2.

MULLEN, M. P. et al. Single nucleotide polymorphisms in the insulin-like growth factor 1 (igf-1) gene are associated with performance in holstein-friesian dairy cattle. *Frontiers in genetics*, Frontiers Media SA, v. 2, 2011.

NADAF, J. et al. Effect of the prior distribution of snp effects on the estimation of total breeding value. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S6.

NASCIMENTO, C. S. d. et al. Association of the bovine major histocompatibility complex (bola) bola-drb3 gene with fat and protein production and somatic cell score in brazilian gyr dairy cattle (bos indicus). *Genetics and Molecular Biology*, SciELO Brasil, v. 29, n. 4, p. 641–647, 2006.

NETER, J. et al. *Applied linear statistical models*. [S.l.]: Irwin Chicago, 1996. v. 4.

NETO, J. O. de O. A. *Troost - Busca de interações entre trios de SNPs em estudos de associação de genoma inteiro*. Tese (Doutorado) — Programa Interunidades em Bioinformática da USP, Novembro 2013.

ÜNSTü, B.; MELSSEN, W.; BUYDENS, L. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Anal. Chim. Acta*, v. 504, p. 292–305, 2005.

ÜNSTü, B.; MELSSEN, W.; BUYDENS, L. Facilitating the application of support vector regression by using a universal pearson vii function based kernel. *Chemometrics and Intelligent Laboratory Systems*, v. 81, p. 29–40, 2006.

NUNKESSER, R. et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, Oxford Univ Press, v. 23, n. 24, p. 3280–3288, 2007.

OGOREVC, J. et al. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal genetics*, Wiley Online Library, v. 40, n. 6, p. 832–851, 2009.

OKSER, S. et al. Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young finns study. *PLoS genetics*, Public Library of Science, v. 6, n. 9, p. e1001146, 2010.

OLAZAR, M. R. R. *Uma metodologia para a descoberta de marcadores genéticos*. Tese (Doutorado) — Programa de Engenharia Elétrica da COPPE - UFRJ, Maio 2013.

OLIVEIRA, F. C. de et al. Metodologia para seleção de marcadores com máquina de vetores suporte com regressão. In: _____. *Talking About Computing and Genomics (TACG): vol.I: Modelos e Métodos Computacionais em Bioinformática*. [S.l.]: Embrapa, 2014. cap. 4, p. 101–126. ISBN 978-85-7035-382-5.

OLIVEIRA, F. C. de et al. Snps selection using support vector regression and genetic algorithms in gwas. *BMC Genomics*, v. 15 (Suppl 7), p. 15, 2014. Disponível em: <http://www.biomedcentral.com/1471-2164/15/S7/S4/abstract>.

OLSEN, H. et al. A genome scan for quantitative trait loci affecting milk production in norwegian dairy cattle. *Journal of dairy science*, Elsevier, v. 85, n. 11, p. 3124–3130, 2002.

OLSEN, H. G. et al. Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics*, Genetics Soc America, v. 169, n. 1, p. 275–283, 2005.

ORRU, L. et al. Characterization of a snps panel for meat traceability in six cattle breeds. *Food Control*, Elsevier, v. 20, n. 9, p. 856–860, 2009.

PACKARD, N. H. A genetic learning algorithm for the analysis of complex data. *Complex Systems*, v. 4, n. 5, p. 543–572, 1990.

PAGALLO, G.; HAUSSLER, D. Boolean feature discovery in empirical learning. *Machine learning*, Springer, v. 5, n. 1, p. 71–99, 1990.

PANG-NING, T. et al. Introduction to data mining. In: *Library of Congress*. [S.l.: s.n.], 2006. p. 74.

PAREEK, R. et al. Immunorelevant gene expression in lps-challenged bovine mammary epithelial cells. *J Appl Genet*, v. 46, n. 2, p. 171–177, 2005.

PARK, T.; CASELLA, G. The bayesian lasso. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 482, p. 681–686, 2008.

PÉREZ, P. et al. Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in r. *The Plant Genome*, Crop Science Society of America, v. 3, n. 2, p. 106–116, 2010.

PEROLINI, A. Genetic algorithms and kernel matrix-based criteria combined approach to perform feature and model selection for support vector machines. *Word Academy of Science, Engineering and Technology*, 2012.

PHUONG, T. M.; LIN, Z.; ALTMAN, R. B. Choosing snps using feature selection. In: IEEE. *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*. [S.l.], 2005. p. 301–309.

PIERCE, B. A. *Genetics: A conceptual approach*. [S.l.]: Macmillan, 2010.

PIERCE, B. A. *Genética: um enfoque conceitual*. 3rd. ed. Rio de Janeiro, RJ, Brasil: Guanabara Koogan, 2013. ISBN 978-85-277-1664-2.

PIMENTEL, E. et al. Exploration of relationships between production and fertility traits in dairy cattle via association studies of snps within candidate genes derived by expression profiling. *Animal genetics*, Wiley Online Library, v. 42, n. 3, p. 251–262, 2011.

PONGPANICH, M.; SULLIVAN, P. F.; TZENG, J.-Y. A quality control algorithm for filtering snps in genome-wide association studies. *Bioinformatics*, Oxford Univ Press, v. 26, n. 14, p. 1731–1737, 2010.

POULSEN, N. A. et al. The occurrence of noncoagulating milk and the association of bovine milk coagulation properties with genetic variants of the caseins in 3 scandinavian dairy breeds. *Journal of dairy science*, Elsevier, v. 96, n. 8, p. 4830–4842, 2013.

PRINZENBERG, E.-M. et al. Polymorphism of the bovine csn1s1 promoter: linkage mapping, intragenic haplotypes, and effects on milk production traits. *Journal of dairy science*, Elsevier, v. 86, n. 8, p. 2696–2705, 2003.

PURCELL, S. et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, Elsevier, v. 81, n. 3, p. 559–575, 2007.

QANBARI, S. et al. The pattern of linkage disequilibrium in german holstein cattle. *Animal genetics*, Wiley Online Library, v. 41, n. 4, p. 346–356, 2010.

RADDING, C. M. Homologous pairing and strand exchange in genetic recombination. *Annual review of genetics*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 16, n. 1, p. 405–437, 1982.

RAMBEAUD, M.; PIGHETTI, G. Differential calcium signaling in dairy cows with specific cxcr1 genotypes potentially related to interleukin-8 receptor functionality. *Immunogenetics*, Springer, v. 59, n. 1, p. 53–58, 2007.

RIDLEY, M. *Evolução*. [S.l.]: Artmed, 2006.

RINCON, G. et al. Polymorphisms in genes in the srebp1 signalling pathway and scd are associated with milk fatty acid composition in holstein cattle. *Journal of Dairy Research*, Cambridge Univ Press, v. 79, n. 01, p. 66–75, 2012.

RON, M. et al. Combining mouse mammary gland gene expression and comparative mapping for the identification of candidate genes for qtl of milk production traits in cattle. *BMC genomics*, BioMed Central Ltd, v. 8, n. 1, p. 183, 2007.

ROTHMAN, K. J.; GREENLAND, S.; WALKER, A. M. Concepts of interaction. *American journal of epidemiology*, Oxford Univ Press, v. 112, n. 4, p. 467–470, 1980.

RUTTEN, M. J. et al. Genetic variation in vitamin b-12 content of bovine milk and its association with snp along the bovine genome. 2013.

SAHANA, G. et al. A new powerful method for genome-wide association mapping using local genealogies in a mixed model. *Local Genealogies in a Linear Mixed Model for Genome-Wide Association Mapping in Complex Pedigreed Populations. PLoS ONE*, v. 11, n. 6, p. e27061, 2011.

SANDERS, K. et al. Characterization of the dgat1 mutations and the csn1s1 promoter in the german angeln dairy cattle population. *Journal of dairy science*, Elsevier, v. 89, n. 8, p. 3164–3174, 2006.

SANDOR, C. et al. Linkage disequilibrium on the bovine x chromosome: characterization and use in quantitative trait locus mapping. *Genetics*, Genetics Soc America, v. 173, n. 3, p. 1777–1786, 2006.

SARGOLZAEI, M.; SCHENKEL, F. S. Qmsim: a large-scale genome simulator for livestock. *Bioinformatics*, v. 25, n. 5, p. 680–681, 2009. Disponível em: <http://bioinformatics.oxfordjournals.org/content/25/5/680.abstract>.

SCHENNINK, A. et al. Effect of polymorphisms in the fasn, olr1, ppargc1a, prl and stat5a genes on bovine milk-fat composition. *Animal Genetics*, Wiley Online Library, v. 40, n. 6, p. 909–916, 2009.

SCHNABEL, R. D. et al. Fine-mapping milk production quantitative trait loci on bta6: analysis of the bovine osteopontin gene. *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, v. 102, n. 19, p. 6896–6901, 2005.

SCHOPEN, G. et al. Whole genome scan to detect quantitative trait loci for bovine milk protein composition. *Animal genetics*, Wiley Online Library, v. 40, n. 4, p. 524–537, 2009.

SCHOPEN, G. et al. Whole-genome association study for milk protein composition in dairy cattle. *Journal of dairy science*, Elsevier, v. 94, n. 6, p. 3148–3158, 2011.

SCHROOTEN, C.; BINK, M.; BOVENHUIS, H. Whole genome scan to detect chromosomal regions affecting multiple traits in dairy cattle. *Journal of dairy science*, Elsevier, v. 87, n. 10, p. 3550–3560, 2004.

SCHÜPBACH, T. et al. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, Oxford Univ Press, v. 26, n. 11, p. 1468–1469, 2010.

SCHURINK, A.; JANSS, L. L.; HEUVEN, H. C. Bayesian variable selection to identify qtl affecting a simulated quantitative trait. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S8.

SCHWARZ, D. F. et al. Evaluation of single-nucleotide polymorphism imputation using random forests. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2009. v. 3, n. Suppl 7, p. S65.

SCHWENDER, H.; FRITSCH with a contribution of A. *scrime: Analysis of High-Dimensional Categorical Data such as SNP Data*. [S.l.], 2013. R package version 1.3.3. Disponível em: <http://CRAN.R-project.org/package=scrime>.

SCHWERIN, M. et al. Application of disease-associated differentially expressed genes-mining for functional candidate genes for mastitis resistance in cattle. *Genetics Selection Evolution*, INRA/EDP SCIENCES, v. 35, p. S19–S34, 2003.

SCRUCCA, L. Ga: a package for genetic algorithms in r. *Journal of Statistical Software*, v. 53, n. 4, 2012.

SHAH, S. C.; KUSIAK, A. Data mining and genetic algorithm based gene/snp selection. *Artificial intelligence in medicine*, Elsevier, v. 31, n. 3, p. 183–196, 2004.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, p. 591–611, 1965.

SHAWE-TAYLOR, J.; CRISTIANINI, N. *Kernel methods for pattern analysis*. [S.l.]: Cambridge university press, 2004.

SHEEHY, P. et al. A functional genomics approach to evaluate candidate genes located in a qtl interval for milk production traits on bta6. *Animal genetics*, Wiley Online Library, v. 40, n. 4, p. 492–498, 2009.

SIKORA, K. M. et al. Dna sequence polymorphisms within the bovine guanine nucleotide-binding protein gs subunit alpha (gsα)-encoding (gnas) genomic imprinting domain are associated with performance traits. *BMC genetics*, BioMed Central Ltd, v. 12, n. 1, p. 4, 2011.

SILVA, A. et al. Quantitative trait loci affecting lactose and total solids on chromosome 6 in brazilian gir dairy cattle. *Genetics and Molecular Research*, v. 10, n. 4, p. 3817–3827, 2011.

SILVA, F. F. e. *Seleção Genômica no R*. [S.l.], 2013.

SILVA, M. V. et al. The development of genomics applied to dairy breeding. *Livestock Science*, Elsevier, v. 166, p. 66–75, 2014.

SILVA, M. V. G. B. da. *Utilização de modelos aleatórios no mapeamento de Quantitative Trait Loci em famílias de irmãos completos e de meio-irmãos*. Tese (Doutorado) — Programa de Pós-graduação em Genética e Melhoramento - UFV, 2002.

SINGH, U. et al. Molecular markers and their applications in cattle genetic research: A review. *Biomarkers and Genomic Medicine*, Elsevier, v. 6, n. 2, p. 49–58, 2014.

SIVANANDAM, S.; DEEPA, S. *Introduction to genetic algorithms*. [S.l.]: Springer Science & Business Media, 2007.

SMITH, B. J. et al. boa: an r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, American Statistical Association, v. 21, n. 11, p. 1–37, 2007.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, n. 3, p. 199–222, 2004.

SONSTEGARD, T. et al. Analysis of bovine mammary gland est and functional annotation of the bos taurus gene index. *Mammalian Genome*, Springer-Verlag, v. 13, n. 7, p. 373–379, 2002. ISSN 0938-8990. Disponível em: <http://dx.doi.org/10.1007/s00335-001-2145-4>.

STAŃCZYK, U.; JAIN, L. C. *Feature Selection for Data and Pattern Recognition*. [S.l.]: Springer, 2015. v. 584.

SZYDA, J. et al. Fitting and validating the genomic evaluation model to polish holstein-friesian cattle. *Journal of applied genetics*, Springer, v. 52, n. 3, p. 363–366, 2011.

SZYMCZAK, S. et al. Machine learning in genome-wide association studies. *Genetic Epidemiology*, v. 33, p. 51–57, 2009.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.]: Addison Wesley, ISBN 0-321-32136-7, 2006.

TASSELL, C. V.; ASHWELL, M.; SONSTEGARD, T. Detection of putative loci affecting milk, health, and conformation traits in a us holstein population using 105 microsatellite markers. *Journal of dairy science*, Elsevier, v. 83, n. 8, p. 1865–1872, 2000.

TEAM, R. C. *R: A language and environment for statistical computing*. 2013. Disponível em: <http://www.Rproject.org/>. Acesso em: 3.7.2013.

TETENS, J. et al. Whole-genome association study for energy balance and fat/protein ratio in german holstein bull dams. *Animal genetics*, Wiley Online Library, v. 44, n. 1, p. 1–8, 2013.

THOMAS, D. *Statistical methods in genetic epidemiology*. [S.l.]: Oxford University Press, 2004.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 267–288, 1996.

TYMMS, M. J. *Gene knockout protocols*. [S.l.]: Springer Science & Business Media, 2001. v. 158.

UHMN, S. et al. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. *Expert Systems*, Wiley Online Library, v. 26, n. 1, p. 60–69, 2009.

USAI, M. G.; CARTA, A.; CASU, S. Alternative strategies for selecting subsets of predicting snps by lasso-lars procedure. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S9.

ÜSTÜNKAR, G. et al. Selection of representative snp sets for genome-wide association studies: a metaheuristic approach. *Optimization Letters*, Springer, v. 6, n. 6, p. 1207–1218, 2012.

UTSUNOMIYA, Y. T. et al. Genome-wide mapping of loci explaining variance in scrotal circumference in nellore cattle. *PloS one*, Public Library of Science, v. 9, n. 2, p. e88561, 2014.

VALENTE, J. et al. *Melhoramento genético de bovinos de leite*. [S.l.]: Embrapa Gado de Leite, 2001.

VANDEKERCKHOVE, J.; WEBER, K. The complete amino acid sequence of actins from bovine aorta, bovine heart, bovine fast skeletal muscle, and rabbit, slow skeletal muscle: a protein-chemical analysis of muscle actin differentiation. *Differentiation*, Elsevier, v. 14, n. 1, p. 123–133, 1979.

VANRADEN, P.; WIGGANS, G. Derivation, calculation, and use of national animal model information. *Journal of Dairy Science*, Elsevier, v. 74, n. 8, p. 2737–2746, 1991.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995. v. 1.

VEERKAMP, R. et al. Genome-wide associations for feed utilisation complex in primiparous holstein–friesian dairy cows from experimental research herds in four european countries. *Animal*, Cambridge Univ Press, v. 6, n. 11, p. 1738–1749, 2012.

VELMALA, R. et al. A search for quantitative trait loci for milk production traits on chromosome 6 in finnish ayrshire cattle. *Animal Genetics*, Wiley Online Library, v. 30, n. 2, p. 136–143, 1999.

VERNEQUE, R. da S. et al. *Programa de Melhoramento Genético da Raça Girolando – Sumário de Touros – Resultado do Teste de Progênie*. [S.l.], 2012.

VIITALA, S. M. et al. Quantitative trait loci affecting milk production traits in finnish ayrshire dairy cattle. *Journal of dairy science*, Elsevier, v. 86, n. 5, p. 1828–1836, 2003.

VILLELA, S. M.; LEITE, S. de C.; NETO, R. F. Seleção de marcadores genômicos com busca ordenada e um classificador de larga margem. In: _____. [S.l.]: Embrapa, 2014. p. 127–181. ISBN 978-85-7035-382-5.

WAN, X. et al. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, Elsevier, v. 87, n. 3, p. 325–340, 2010.

WAN, X. et al. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, Oxford Univ Press, v. 26, n. 1, p. 30–37, 2010.

WANG, D.; ESKRIDGE, K. M.; CROSSA, J. Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of agricultural, biological, and environmental statistics*, Springer, v. 16, n. 2, p. 170–184, 2011.

WANG, N. et al. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *The American Journal of Human Genetics*, Elsevier, v. 71, n. 5, p. 1227–1234, 2002.

WANG, W. Y. et al. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, Nature Publishing Group, v. 6, n. 2, p. 109–118, 2005.

WANG, X. et al. Short communication: Association of an olr1 polymorphism with milk production traits in the israeli holstein population. *Journal of dairy science*, Elsevier, v. 95, n. 3, p. 1565–1567, 2012.

WANG, X. et al. Identification and dissection of four major qtl affecting milk fat content in the german holstein-friesian population. *PloS one*, Public Library of Science, v. 7, n. 7, p. e40711, 2012.

WANG, Y. et al. Antepiseeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC research notes*, BioMed Central Ltd, v. 3, n. 1, p. 117, 2010.

WASAN, P. S. et al. Application of statistics and machine learning for risk stratification of heritable cardiac arrhythmias. *Expert Systems with Applications*, v. 10, n. 7, p. 1384–1389, 2012.

WEI, Z. et al. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, v. 5, n. 10, p. 1–11, 2009.

WEIGEL, K. et al. Predictive ability of direct genomic values for lifetime net merit of holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of dairy science*, Elsevier, v. 92, n. 10, p. 5248–5257, 2009.

WEIKARD, R. et al. The bovine ppargc1a gene: molecular characterization and association of an snp with variation of milk fat synthesis. *Physiological Genomics*, Am Physiological Soc, v. 21, n. 1, p. 1–13, 2005.

WELTER, D. et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, Oxford Univ Press, v. 42, n. D1, p. D1001–D1006, 2014.

WESTON, J. et al. Consistency-based search in feature selection. *Advances in Neural Information Processing Systems*, MIT Press, v. 12, p. 526—-532, 2000.

WIENER, P. et al. Testing for the presence of previously identified qtl for milk production traits in new populations. *Animal Genetics*, Wiley Online Library, v. 31, n. 6, p. 385–395, 2000.

WINHAM, S. J. et al. Snp interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, BioMed Central Ltd, v. 13, n. 1, p. 164, 2012.

WINTER, A. et al. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-coa: diacylglycerol acyltransferase (dgat1) with variation at a quantitative trait locus for milk fat content. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 99, n. 14, p. 9300–9305, 2002.

WIRGIN, A. The inverse crime. *arXiv preprint math-ph/0401050*, 2004.

WONG, C.; BERNARDO, R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theoretical and Applied Genetics*, Springer, v. 116, n. 6, p. 815–824, 2008.

WRIGHT, S. *Vol. 4: Variability within and among natural populations*. [S.l.]: Chicago [etc.]: University of Chicago Press, 1978.

XIE, Y. et al. Identification and expression pattern of two novel alternative splicing variants of eef1d gene of dairy cattle. *Gene*, Elsevier, v. 534, n. 2, p. 189–196, 2014.

YANG, C. et al. Snpharvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, Oxford Univ Press, v. 25, n. 4, p. 504–511, 2009.

YANG, Y. et al. Three novel single-nucleotide polymorphisms of complement component 4 gene (c4a) in chinese holstein cattle and their associations with milk performance traits and ch50. *Veterinary immunology and immunopathology*, Elsevier, v. 145, n. 1, p. 223–232, 2012.

YAU, C. R tutorial with bayesian statistics using open-bugs. *URL http://www. r-tutor. com/conten 'dr-tutorial-ebook. Indice de instrucciones*, 2012.

YOON, Y. et al. Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Clinical chemistry and laboratory medicine*, v. 41, n. 4, p. 529–534, 2003.

YOUNGERMAN, S. et al. Association of cxcr2 polymorphisms with subclinical and clinical mastitis in dairy cattle. *Journal of dairy science*, Elsevier, v. 87, n. 8, p. 2442–2448, 2004.

YTOURNEL, F. et al. LDSO: a program to simulate pedigrees and molecular information under various evolutionary forces. *Journal of Animal Breeding and Genetics*, v. 129, n. 5, p. 417–421, out. 2012. ISSN 1439-0388. Disponível em: <http://dx.doi.org/10\%2e1111/j\%2e1439\%2d0388\%2e2011\%2e00986\%2ex>.

ZENG, J. et al. Genomic breeding value prediction and qtl mapping of qtlmas2011 data using bayesian and gblup methods. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2012. v. 6, n. Suppl 2, p. S7.

ZHANG, X. et al. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, Oxford Univ Press, v. 26, n. 12, p. i217–i227, 2010.

ZHANG, X. et al. Coe: a general approach for efficient genome-wide two-locus epistasis test in disease association study. In: SPRINGER. *Research in Computational Molecular Biology*. [S.l.], 2009. p. 253–269.

ZHANG, X.; ZOU, F.; WANG, W. Fastanova: an efficient algorithm for genome-wide association study. In: ACM. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2008. p. 821–829.

ZHANG, Y.; LIU, J. S. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, Nature Publishing Group, v. 39, n. 9, p. 1167–1173, 2007.

ZHENG, G.; FREIDLIN, B.; GASTWIRTH, J. L. Comparison of robust tests for genetic association using case-control studies. *Lecture Notes-Monograph Series*, JSTOR, p. 253–265, 2006.

ZHENG, J.; WATSON, A. D.; KERR, D. E. Genome-wide expression analysis of lipopolysaccharide-induced mastitis in a mouse model. *Infection and immunity*, Am Soc Microbiol, v. 74, n. 3, p. 1907–1915, 2006.

ZHENG, X. et al. Single nucleotide polymorphisms, haplotypes and combined genotypes of lap3 gene in bovine and their association with milk production traits. *Molecular biology reports*, Springer, v. 38, n. 6, p. 4053–4061, 2011.

ZIEGLER, A.; KOENIG, I. *A statistical approach to genetic epidemiology*. [S.l.]: Wiley-VCH, 2007.

ZIELKE, L. G. et al. Impact of variation at the fto locus on milk fat yield in holstein dairy cattle. 2013.

ZIMIN, A. V. et al. A whole-genome assembly of the domestic cow, bos taurus. *Genome Biol*, v. 10, n. 4, p. R42, 2009. Disponível em: <http://www.biomedcentral.com/content/pdf/gb-2009-10-4-r42.pdf>.

ZUBER, V.; STRIMMER, K. High-dimensional regression and variable selection using car scores. *Statistical Applications in Genetics and Molecular Biology*, v. 10, n. 1, p. 1–27, 2011.

APÊNDICE A - Termo de Uso dos Dados e Publicações

A.1 Termo de Uso dos Dados

O conjunto de dados reais utilizados nesta tese são bases de dados de genótipo e de fenótipo, ambas de propriedade da Embrapa Gado de Leite, e são partes integrantes do projeto MCBio — Modelos Computacionais para Estabelecimento de Meios e Procedimentos Metodológicos para Análise de Dados em Bioinformática, registrado no Sistema Embrapa de Gestão (SEG), sob o código 05.10.03.006.00.00, e do Programa Nacional de Melhoramento Genético do Gir Leiteiro.

O referido conjunto de dados foi utilizado exclusivamente para o cumprimento dos estudos referentes a esta tese e sem que fosse identificado qualquer indivíduo cujo os dados se encontrassem na mesma.

Para o desenvolvimento destes estudos, o autor comprometeu-se a manter total reserva em relação a quaisquer dados ou informações da Embrapa que venha porventura ter acesso em razão de sua presença no âmbito desta Empresa, não os utilizando para interesse próprio ou de terceiros, nem os repassando a terceiros sob qualquer forma ou pretexto, independentemente de se tratar ou não de informação reservada, confidencial ou sigilosa, mesmo após a sua conclusão.

A.2 Publicação no periódico *BMC Genomics* referente ao congresso X-Meeting 2013

de Oliveira *et al. BMC Genomics* 2014, **15**(Suppl 7):S4
http://www.biomedcentral.com/1471-2164/15/S7/S4

BMC Genomics

RESEARCH Open Access

SNPs selection using support vector regression and genetic algorithms in GWAS

Fabrízzio Condé de Oliveira[1], Carlos Cristiano Hasenclever Borges[1], Fernanda Nascimento Almeida[2,4], Fabyano Fonseca e Silva[3], Rui da Silva Verneque[4], Marcos Vinicius GB da Silva[4], Wagner Arbex[1,4*]

From 9th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2013)
Recife, Brazil. 3-6 November 2013

Abstract

Introduction: This paper proposes a new methodology to simultaneously select the most relevant SNPs markers for the characterization of any measurable phenotype described by a continuous variable using Support Vector Regression with Pearson Universal kernel as fitness function of a binary genetic algorithm. The proposed methodology is multi-attribute towards considering several markers simultaneously to explain the phenotype and is based jointly on statistical tools, machine learning and computational intelligence.

Results: The suggested method has shown potential in the simulated database 1, with additive effects only, and real database. In this simulated database, with a total of 1,000 markers, and 7 with major effect on the phenotype and the other 993 SNPs representing the noise, the method identified 21 markers. Of this total, 5 are relevant SNPs between the 7 but 16 are false positives. In real database, initially with 50,752 SNPs, we have reduced to 3,073 markers, increasing the accuracy of the model. In the simulated database 2, with additive effects and interactions (epistasis), the proposed method matched to the methodology most commonly used in GWAS.

Conclusions: The method suggested in this paper demonstrates the effectiveness in explaining the real phenotype (PTA for milk), because with the application of the wrapper based on genetic algorithm and Support Vector Regression with Pearson Universal, many redundant markers were eliminated, increasing the prediction and accuracy of the model on the real database without quality control filters. The PUK demonstrated that it can replicate the performance of linear and RBF kernels.

Background

Single nucleotide polymorphisms (SNPs) are an abundant form of genomic variation, which differ from rare variants [1] and the basic assumption for wide association studies (GWAS) is that the evaluated characteristic can be explained from this type of marker. Thus, it is considered that there are SNPs in the genotype with high Linkage Disequilibrium (LD) compared to Quantitative Trait Locus (QTL). So, the traditional approach is to evaluate which markers that have a high association with the phenotype through the p-value of beta linear regression between each SNP and the phenotype. After this step, the most relevant SNPs are analyzed for proximity to some region that is associated with that feature or other features that can be indirectly correlated with the phenotype in question. So far, the prediction of disease risk in humans based on validated SNPs based on this methodology showed little predictive power [2], although these SNPs indicate highly significant association with the phenotypic trait. This fact can be explained due the variance of the most significant markers have low explanatory power in relation to the phenotypic variance [3]. Therefore, an alternative approach is to increase the number of markers, considering also those with small correlations on the trait. But, this fact creates two problems: the number of markers is high and many of them are correlated. According to [4], such analysis requires the use of statistical methods that

* Correspondence: wagner.arbex@ufjf.edu.br
[1]Federal University of Juiz de Fora - UFJF, Juiz de Fora, Minas Gerais, Brasil
Full list of author information is available at the end of the article

A.3 Publicação no congresso CISTI 2014

Decision Support in Attribute Selection with Machine Learning Approach

Wagner Arbex
Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil
wagner.arbex@embrapa.br

Fabrízzio Condé de Oliveira
Federal University of Juiz de Fora — UFJF
Juiz de Fora, MG, Brazil

Fabyano Fonseca e Silva
Federal University of Viçosa — UFV
Viçosa, MG, Brazil

Luis Varona
University of Zaragoza — UNIZAR
Zaragoza, Spain

Marcos Vinícius Gualberto Barbosa da Silva
Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil

Rui da Silva Verneque
Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil

Carlos Cristiano Hasenclever Borges
Federal University of Juiz de Fora — UFJF
Juiz de Fora, MG, Brasil

Abstract—**This paper proposes a method to simultaneously select the most relevant single nucleotide polymorphisms (SNPs) markers — the attributes — for the characterization of any measurable phenotype described by a continuous variable using support vector regression (SVR) with Pearson VII Universal Kernel (PUK). The proposed study is multiattribute towards considering several markers simultaneously to explain the phenotype and is based jointly on a statistical tools, machine learning and computational intelligence.**

Keywords-decision support; attribute selection; machine learning; SVR; computational modeling

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are an abundant form of genomic variation, which differ from rare variants [1] and the basic assumption for genome-wide association studies (GWAS) is that the evaluated characteristic can be explained from this type of marker.

The traditional approach is to evaluate which markers that have a high association with the phenotype through the *p-value* of beta linear regression between each SNP and the phenotype. After this step, the most relevant SNPs are analyzed for proximity to some region that is associated with that feature or other features that can be indirectly correlated with the phenotype in question.

Therefore, an alternative approach is to increase the number of markers, considering also those with small correlations on the trait. But, this fact creates two problems: the number of markers is high and many of them are correlated. According to [2], such analysis requires the use of statistical methods that consider the selection of covariates – i. e., the multicollinearity problem -- and the regularization of the estimation process – i.e., the problem of dimensionality.

Other regression techniques were created to address this problem as ridge regression and partial least squares regression [3]. On the other hand, machine learning algorithms such as support vector machine (SVM) in GWAS considering multiple markers in classification problems, have demonstrated satisfactory performance as in [4], [5] and [6].

This study aims to propose a method that can simultaneously evaluate several SNPs in relation to the phenotype described by a continuous variable, unlike case-control dichotomous phenotypes addressed to the majority of GWAS studies. With this, there are two immediate benefits relative to standard methodology: one relating to the various levels of the phenotypes and the other by complex simultaneous interactions that may occur between the various markers.

To demonstrate the proposed method was used a sample of 343 samples (bulls) genotyped provided by the Brazilian Agricultural Research Corporation (Embrapa), and only 244 animals have female offspring, allowing the measurement of the phenotype evaluated.

A.4 Capítulo 4 do livro Talking About Computing and Genomics - Volume 1

4

Metodologia para seleção de marcadores com máquina de vetores de suporte com regressão

Fabrízzio Condé de Oliveira
Fernanda Nascimento Almeida
Fabyano Fonseca e Silva
Marcos Vinícius Gualberto Barbosa da Silva
Carlos Cristiano Hasenclever Borges
Wagner Arbex

ESTE TRABALHO propõe uma nova metodologia para selecionar simultaneamente os marcadores SNPs mais relevantes para a caracterização de qualquer fenótipo mensurável descrito por uma variável contínua, usando SVR com o Pearson Universal Kernel. A metodologia proposta é multiatributo no sentido de considerar vários marcadores simultâneos para a explicação do fenótipo e baseia-se conjuntamente em um ferramental estatístico, técnicas de aprendizado de máquina e de inteligência computacional. Atualmente, a maioria dos estudos de associação em escala genômica, chamados de GWAS (Genome-wide Association Studies), quantificam o impacto médio de cada marcador sobre o fenótipo por meio de regressões lineares simples entre um marcador e o fenótipo (modelos monoatributos), com o intuito de indicar os marcadores mais

A.5 Descrição do SMS

– Descrição de Metodologia –

SMS – SNP Markers Selector

Wagner Arbex[1]

A metodologia ***SNP Markers Selector – SMS*** é capaz de selecionar os marcadores SNPs, do inglês *single nucleotide polymorphisms* (polimorfismos de base única), mais relevantes para a caracterização de qualquer fenótipo mensurável descrito por uma variável contínua e sua abordagem multiatributo permite considerar vários marcadores simultaneamente para a explicação do fenótipo. A metodologia aplica, principalmente, SVR, isto é, *support vector machines* (máquinas de vetores de suporte com regressão), fazendo uso do Pearson Universal Kernel, além de utilizar outras técnicas de aprendizado de máquina e de inteligência computacional apoiadas em recursos estatísticos. Os estudos de associação em escala genômica ou GWAS, isto é, *genome-wide association studies*, em geral, quantificam o impacto médio de cada marcador sobre o fenótipo por meio de regressões lineares simples entre um marcador e o fenótipo (modelos monoatributos), com o intuito de indicar os marcadores mais significativos em relação à característica fenotípica em questão. O uso de tais métodos pressupõem que os efeitos de cada marcador sobre o fenótipo são somente aditivos, desconsiderando a possível ocorrência de interações complexas como epistasia e dominância entre os marcadores. Contudo, diferentemente da abordagem clássica, abordagem multiatributo da SMS em conjunto com os recursos e técnicas de modelagem computacional utilizados em seu desenvolvimento, permitem que tais interações sejam capturadas. A metodologia foi desenvolvida no âmbito do projeto *SEG 05.10.03.006.00.00, Modelos Computacionais para Estabelecimento de Meios e Procedimentos Metodológicos para Análise de Dados em Bioinformática – MCBio* (ARBEX et al., 2010), sendo parte dos trabalhos de doutorado de Fabrízzio Condé de Oliveira, um dos integrantes do referido projeto. No escopo do MCBio, a metodologia em questão era parte de seus resultados e foi entregue como o capítulo de livro Metodologia para Seleção de Marcadores com Máquina de Vetores de Suporte com Regressão (OLIVEIRA et al., 2014).

Referências

ARBEX, Wagner; HIGA, Roberto Hirosh; SILVA JÚNIOR, Orzenil Bonfim da; TOGAWA, Roberto Coiti. Modelos computacionais para estabelecimento de meios e procedimentos metodológicos para análise de dados em bioinformática – MCBio. Juiz de Fora, 2010. 36 p. Projeto de pesquisa e desenvolvimento submetido à Chamada 03/2010 da Embrapa no âmbito do Macroprograma 5 - Desenvolvimento Institucional.

OLIVEIRA, Fabrízzio Condé de; ALMEIDA, Fernanda Nascimento; SILVA, Fabyano Fonseca; SILVA, Marcos Vinícius G. Barbosa da; BORGES, Carlos Cristiano Hansenclever; ARBEX, Wagner. Metodologia para seleção de marcadores com máquina de vetores de suporte com regressão. In: Wagner Arbex; Natália Florêncio Martins; Marta Fonseca Martins. (Ed.). Talking about computing and genomics (TACG): modelos e métodos computacionais em bioinformática. 1ed. Brasília: Embrapa, 2014, v. 1, p. 101-126. ISBN 978-85-7035-382-5.

1 Analista da Embrapa Gado de Leite

A.6 Registro do SMS

Embrapa
Gado de Leite

DECLARAÇÃO

Declaramos para os fins que se fizerem necessários, que o empregado ***Wagner Arbex*** publicou, durante o ano de 2014, em co-autoria, a metodologia **SMS – SNP Markers Selector**, como descrita em:

OLIVEIRA, FABRÍZZIO CONDÉ; ALMEIDA, FERNANDA NASCIMENTO; SILVA, FABYANO FONSECA; SILVA, MARCOS VINÍCIUS G. BARBOSA; BORGES, CARLOS CRISTIANO HANSENCLEVER; ARBEX, WAGNER. Metodologia para seleção de marcadores com máquina de vetores de suporte com regressão. In: ARBEX, WAGNER; MARTINS, NATÁLIA FLORÊNCIO; MARTINS, MARTA FONSECA (Ed.). **Talking about computing and genomics (TACG):** modelos e métodos computacionais em bioinformática. 1. ed. Brasília: Embrapa, 2014, v. 1, cap. 4, p. 101-126. ISBN 978-85-7035-382-5

Juiz de Fora, 27 de janeiro de 2015

William Fernandes Bernardo
Chefe-adjunto de Transferência de Tecnologia

Empresa Brasileira de Pesquisa Agropecuária
Centro Nacional de Pesquisa de Gado de Leite
Ministério da Agricultura, Pecuária e Abastecimento
Rua Eugênio do Nascimento, 610 – Bairro Dom Bosco – 36038-330 Juiz de Fora/MG
Telefone: (32)3311.7475 Fax: (32)3311.7401
cnpgl.chtt@embrapa.br