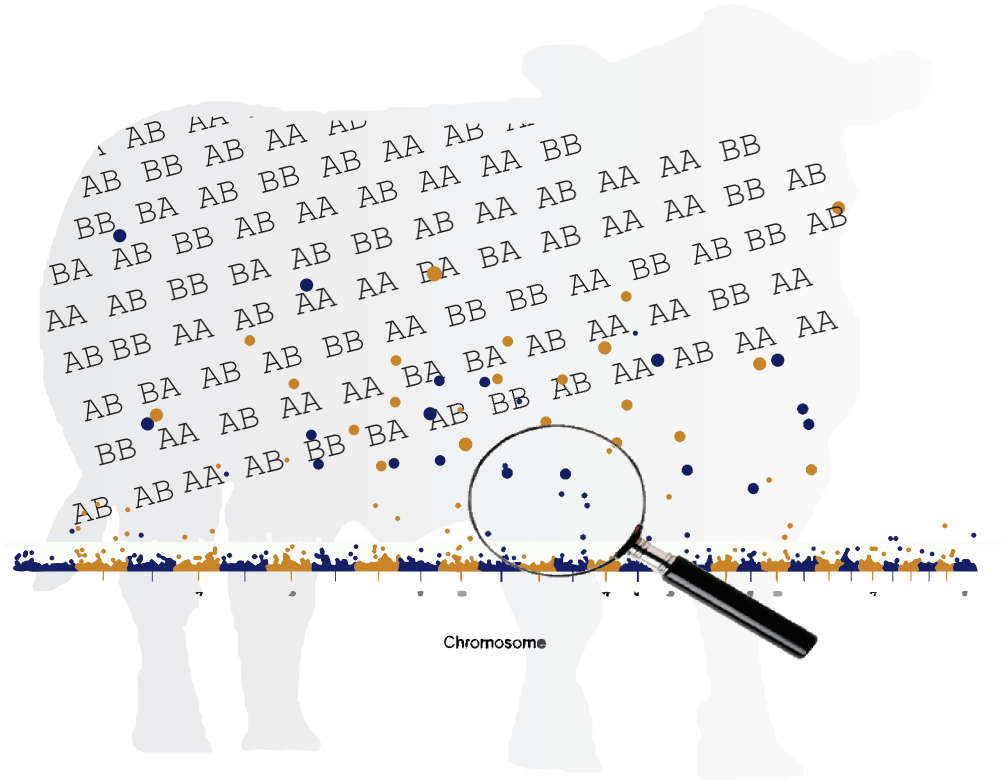
Documentos

ISSN 1982-5390
Dezembro, 2016
152

Aplicação de Métodos Bayesianos em Estudos de Associação Genômica-Ampla para Características Complexas em Animais de Produção



Embrapa

ISSN 1982-5290
Dezembro, 2016

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Pecuária Sul
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 152

Aplicação de Métodos Bayesianos em Estudos de Associação Genômica-Ampla para Características Complexas em Animais de Produção

Bruna Pena Sollero
Fernando Antonio Reimann
Gabriel Soares Campos
Fernando Flores Cardoso

Embrapa Pecuária Sul
Bagé, RS
2016

Exemplares desta publicação podem ser adquiridos na:

Embrapa Pecuária Sul
BR 153, Km 632,9. Caixa postal 242
96401-970 - Bagé – RS
Fax: 55 53 3240-4650
https://www.embrapa.br/pecuaria-sul
www.embrapa.br/fale-conosco/sac

Supervisor editorial: *Fernando Goss*
Revisor de texto: *Fernando Goss*
Normalização bibliográfica: *Graciela Olivella Oliveira*
Editoração eletrônica: *Núcleo de Comunicação Organizacional*
Ilustração da capa: *Bruna Pena Sollero / Núcleo de comunicação Organizacional*

1ª edição
Publicação digitalizada (2016)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Pecuária Sul

Aplicação de métodos Bayesianos em estudos de associação genômica-ampla para características complexas em animais de produção / Bruna Pena Sollero ... [et al.]. — Bagé : Embrapa Pecuária Sul, 2016.
24 p. : il. ; 21 x 15 cm. – (Documentos / Embrapa Pecuária Sul, ISSN 1982-5390 ; 152)

1. Genética quantitativa. 2. Melhoramento genético animal. 3. Gado de corte. I. Sollero, Bruna Pena. II. Embrapa Pecuária Sul. III. Série.

CDD 636.0821

Autores

Bruna Pena Sollero
Zootecnista, Doutora em Melhoramento Animal, pesquisadora da Embrapa Pecuária Sul, Caixa Postal 242, BR 153 Km 632,9, CEP 96401-970, Bagé, RS - Brasil

Fernando Flores Cardoso
Médico Veterinário, Doutor em Bioinformática ênfase em Estatística Genômica, pesquisador da Embrapa Pecuária Sul, Caixa Postal 242, BR 153 Km 632,9, CEP 96401-970 Bagé, RS - Brasil

Fernando Antonio Reimann
Médico Veterinário, Mestre em Melhoramento Genético Animal, Ufpel, fernando.reimann@hotmail.com

Gabriel Soares Campos
Engenheiro Agrônomo,Mestre em Zootecnia/Melhoramento Genético Animal - UFRGS
gabrielsoarescampos@hotmail.com

Apresentação

As publicações técnicas da Série Embrapa são importantes veículos de informação, destinada a produtores, técnicos, empresários do agronegócio, pesquisadores, estudantes e público em geral interessados nas tecnologias desenvolvidas pela Empresa e seus colaboradores. Trata-se de publicações com distintas características, objetivos e público alvo, tais como: Boletim de Pesquisa e Desenvolvimento; Documentos; Circular Técnica; Comunicado Técnico; Sistemas de Produção; Livro e outros.

A Embrapa Pecuária Sul utiliza este veículo para comunicar suas tecnologias produzidas, recomendações, práticas agrícolas e resultados de pesquisa e desenvolvimento, direcionando ao público interessado informações ligadas à produção de forrageiras e pastagens, bovinocultura de corte e leite e ovinocultura dos campos sulbrasileiros. É com satisfação que oferecemos mais esta obra, destacando recente trabalho desenvolvido pelo Centro da Embrapa, em Bagé, em benefício à sustentabilidade da pecuária sulina.

Esta publicação relata a importância dos estudos de associação genômica, aplicada à produção animal, e métodos que permitem identificar genes ligados a melhores desempenhos produtivos, reprodutivos, animais mais tolerantes a determinada doença ou parasita, entre outros; resultando na implantação de testes genômicos que auxiliam criadores na seleção de animais com maior potencial genético.

Esperamos que os leitores desfrutem deste Documento e sugerimos que, em caso de maior interesse no tema abordado ou necessidades de esclarecimentos, realizem o contato com nosso setor de atendimento ao cliente (SAC), acessando www.embrapa.br/fale-conosco/sac/ou pelo fone (53) 3240-4650. A Embrapa terá o máximo prazer em atendê-lo.

Atenciosamente,

Alexandre Varella
Chefe-Geral

Sumário

Introdução 08

Abordagem estatística 10

Passo-a-passo de uma análise Bayesiana para associação genômica utilizando o programa GenSel 12

Arquivos necessários 13
Passo-a-passo para as análises 17
a- Estimação de componentes de variâncias 17
b- Correção dos fenótipos 19
c- Variáveis categóricas 21
d- Escolha do método bayesiano para estimação dos efeitos dos marcadores 22
e- Regiões cromossômicas ou janelas e efeitos de marcadores 27

Desequilíbrio de ligação 33

Seleção de marcadores mais informativos 35

Considerações finais 38

Referências..... 39

Anexo.....44

Aplicação de Métodos Bayesianos em Estudos de Associação Genômica-Ampla para Características Complexas em Animais de Produção

Bruna Pena Sollero
Fernando Antonio Reimann
Gabriel Soares Campos
Fernando Flores Cardoso

Introdução

Desde o início deste século, especialmente depois de finalizado o sequenciamento do genoma bovino (ELSIK et al., 2009), os avanços na biologia molecular e as inovações de sequenciamento de DNA (ácido desoxirribonucleico) e de genotipagem resultaram em reduções nos custos e rapidez na prospecção de marcadores do tipo SNPs (Single Nucleotide Polimorphism) e geração de dados genéticos de animais. As inovações biotecnológicas que permitiram a rapidez, automação e reprodutibilidade no acesso a milhares de genótipos favoreceram então, a seleção assistida por marcadores e estudos de associação em escala genômica (GODDARD; HAYES, 2009; MEUWISSEN et al., 2001).

Uma das grandes vantagens dos marcadores SNP é a fácil conversão dos ensaios entre diferentes plataformas. Diferentes tecnologias para genotipagem de SNPs nessa escala têm sido apresentadas muito rapidamente por diferentes empresas (Illumina – BeadExpress, ABI/BioTrove OpenArrary, ABI-SNPlex, ABI-TaqMan, etc.) (CAETANO, 2009). Dentre os chips disponíveis no mercado para genotipagem de bovinos em larga escala, citamos:

- Illumina High Density Bovine Bead Chip Array (777.962 SNPs);
- Affymetrix Axiom Genome Wide BOS 1 Array (648.874 SNPs);
- BovineSNP50 da Illumina (54.609 SNPs) (MATUKUMALLI et al., 2009);
- BovineLD (6.909 SNPs) (BOICHARD et al., 2012).

A incorporação das informações genômicas na estimação dos valores genéticos resulta em DEPs genômicas, que se referem à diferença esperada no desempenho da progênie de um animal, aprimorada pela genômica, em comparação com a performance média das progênies de todos os animais em avaliação. Melhores acurácias de predição na estimativa destas DEPs vêm sendo almejadas, testando-se diferentes metodologias para estimação de efeitos dos marcadores relacionados a uma determinada característica e modelos de equações preditivas para animais candidatos à seleção.

Por sua vez, os estudos de associação ampla do genoma (GWAS) possibilitam a identificação de regiões cromossômicas e SNPs que explicam parte da variação genética de determinado fenótipo e se baseiam na suposição de que uma mutação causativa para tal está em desequilíbrio de ligação com marcadores adjacentes nas diversas famílias de uma população. De maneira prática, estudos de associação genômica ampla permitem identificar genes ligados a melhores desempenhos produtivos, reprodutivos, animais mais tolerantes a determinada doença ou parasita, entre outros; resultando na implantação de testes genômicos que auxiliam criadores na seleção de animais com maior potencial genético.

Os testes genômicos utilizam marcadores mais informativos na expressão de determinados fenótipos e têm o potencial de detectar quais animais apresentam as melhores formas alélicas relacionadas ao maior ou menor valor de uma característica. Com uma pequena amostra de tecido do animal (pelo, sêmen ou sangue), em qualquer fase de vida, pode-se extrair seu DNA e submetê-lo à avaliação. Contudo, para se propor testes genômicos, estudos de associação que contam com uma significativa amostragem tanto de genótipos, quanto fenótipos, devem ser primeiramente desenvolvidos mediante métodos estatísticos adequados.

Abordagem estatística

É preciso estar atento às limitações impostas por alguns modelos genéticos e métodos estatísticos comumente utilizados no estudo de predisposições genéticas (CAMPOS et al., 2010) ou características quantitativas, associadas a determinado fenótipo. Os modelos de regressão de marcador único (SMR), por exemplo, se destacam como a forma mais comum para estudar a associação entre marcas e QTLs - *Quantitative Trait Loci*, que identifica um segmento de DNA (*locus*) relacionado com a variação em um fenótipo (característica quantitativa), em que o efeito de cada marcador (ou SNP) é estimado individualmente. Desta forma, nas abordagens estatísticas clássicas, a associação entre um marcador e uma determinada característica pode ser testada com o modelo:

$$y=1_n\mu+Xg+e,$$

Em que y é o vetor das observações (valores fenotípicos),

1_n é um vetor de 1s atribuído à média,

X é a matriz relacionando cada valor fenotípico com o marcador (Genótipo),

g é o efeito do marcador,

e é o vetor do resíduo (~ N(0,

Este método normalmente é aplicado quando se assume que apenas poucos genes afetam determinado fenótipo. Portanto, este pode não ser satisfatório na aplicação em muitas características economicamente importantes na pecuária, ou seja, aquelas afetadas por um grande número de genes de pequenos efeitos e pelas interações entre eles.

Por outro lado, estudos em escala genômica consideram relativamente muitas covariáveis no modelo (p), ou marcadores, para poucas observações (n). Este tipo de situação é encontrado em qualquer tipo de aplicação da estatística em que é difícil e/ou oneroso obter observações (fenótipos), mas que, por outro lado, é fácil realizar uma grande quantidade de medições (genótipos). Os *p-valores,* inerentes às análises estatísticas clássicas e amplamente utilizados para inferir associações, possuem determinada limitação de acordo com Stephens e Balding (2009, p. 681):

> é difícil de quantificar com base em um *p-valor* o quão confiável é a inferência de que um dado SNP está verdadeiramente associado a um fenótipo. Além disso, o mesmo *p-valor* computado para diferentes SNPs ou em diferentes estudos pode ter diferentes implicações para uma verdadeira associação plausível, dependendo dos fatores que afetam o poder do teste, como MAF (*minor allele frequency*) e número amostral. Isto porque, a verdadeira associação não depende somente da hipótese nula - de que não existe associação entre o SNP e o QTL (a mesma para todos os testes)- mas também, da hipótese alternativa (que difere entre um teste e outro).

Desta forma, modelos Bayesianos têm sido propostos como alternativas para contornar a complexidade das análises de dados em níveis genômicos. Os métodos bayesianos têm propiciado novas perspectivas para as questões relacionadas à estimação de componentes de variância e de parâmetros genéticos, pois diferentes graus de incerteza relativos a determinado parâmetro podem ser inferidos por meio de um modelo estatístico *a priori*, o qual considera tanto as observações, quanto os parâmetros, como quantidades aleatórias. Os métodos de

Markov Chain Monte Carlo (MCMC), dentre os quais se destaca a Amostragem de Gibbs, podem ser utilizados como uma ferramenta de forma a propiciar a inferência bayesiana. O algoritmo de Gibbs é aplicado para gerar um valor para cada parâmetro desconhecido e apresenta fácil implementação, principalmente quando comparado a algoritmos baseados em processos não derivativos, uma vez que os resultados permitem gerar distribuições *a posteriori* marginais completas, a partir das quais são obtidas as estimativas dos componentes de variância e parâmetros genéticos (FARIA et al., 2007).

Assim, esta abordagem pode auxiliar na busca por genes candidatos, quando aplicados em estudos de associação, permitindo investigar a arquitetura genômica envolvida na expressão de diversos caracteres, bem como na identificação dos efeitos relativos a cada marcador; o que ajuda a predizer o valor genético genômico final do animal.

Passo-a-passo de uma análise Bayesiana para associação genômica utilizando o programa genSel:

Em contraste às estratégias que testam a associação de cada marcador individualmente com a característica de interesse, métodos Bayesianos têm a prerrogativa de testar todos os marcadores simultaneamente como efeitos aleatórios e assim, serem capazes de capturar a maioria da variância genética explicável pelos marcadores.

Análises de associação Bayesianas requerem informações genotípicas e fenotípicas, um modelo que descreva os fatores que influenciam o fenótipo e especificações de distribuições *a priori*. Dentre os métodos do "alfabeto" Bayesiano, Bayes A, Bayes B, Bayes C, Bayes Cπ e Bayesian Lasso que vêm sendo propostos para calcular a predição dos efeitos dos marcadores (HABIER et al., 2011), no presente documento, concentramos nos três primeiros métodos citados, uma vez disponibilizados no programa GenSel para GWAS.

Arquivos necessários

O programa GenSel é extensivamente utilizado para estudos de associação em que informações de genótipos e fenótipos geram estimativas de efeitos dos SNP bem como valores genéticos moleculares. O programa foi escrito por Rohan Fernando e implementado por Dorian J. Garrick no projeto *Bioinformatics to Implement Genomic Selection* (BIGS), em *Iowa State University*.

O programa é controlado por um arquivo de parâmetros (cartão de parâmetros) e requer que os dados sejam organizados em formatos e arquivos específicos para cada tipo de informação (fenótipos, genótipos, mapas etc.). A seguir descrevemos o conteúdo e estrutura para preparação dos arquivos necessários.

O **arquivo de genótipos** deve ser preparado como mostra a Tabela 1. A primeira linha de cabeçalho contém as informações do conteúdo de cada coluna do arquivo de dados, em que a primeira coluna especifica as identificações dos animais, e nas demais colunas os nomes dos SNPs consecutivamente. Os valores das covariáveis (genótipos) são identificados por -10, 0 ou 10 para genótipos AA, AB ou BB, respectivamente. Genótipos perdidos são identificados pelo número 5.

Tabela1. Exemplo de um arquivo de genótipos no formato do GenSel, representando genótipos codificados de 4 animais (primeira coluna) para 4 marcadores SNPs (primeira linha) representantes do genoma bovino.

ID.bt	HAPMAP30204-BTA-124882	ARS-BFGL-NGS-1705	ARS-BFGGL-NGS5057	ARS-BFGL-NGS-103851
0000208FSB-M	-10	10	0	5
0006008FSB-F	0	-10	10	0
0006808FSB-M	-10	-10	10	0
0010008FSB-F	0	0	5	10

ID.bt indica a linha de nomes dos marcadores.

No programa GenSel, o item *markerFileName* das opções de comandos especificadas no arquivo de extensão ".inp" (arquivo/cartão de parâmetros ou *input*) é utilizado para identificar o nome do arquivo de genótipos preparado. O nome do **arquivo com o mapa** das localizações genômicas dos marcadores (Tabela 2) é identificado por *mapOrderFile-Name*. É importante destacar que o nome da primeira coluna do arquivo de mapa deve ser igual ao nome da primeira coluna no arquivo de genótipos para referenciar corretamente as informações nas análises (ver coluna "ID.bt", Tabelas 1 e 2).

Tabela 2. Exemplo de um arquivo de mapa do GenSel representando 4 marcadores SNPs localizados no cromossomo 1 (BTA1) de bovinos. ID.bt indica a coluna de nomes dos marcadores; BTA_chr indica o número do cromossomo que o marcador pertence; BTA_pos indica a posição exata em pares de bases do marcador no cromossomo; chip# indica a sequência de ordem de cada marcador utilizado.

ID.bt	BTA_chr	BTA_pos	chip#
Hapmap43437-BTA-101873	1	135098	1
ARS-BFGL-NGS-16466	1	267940	2
ARS-BFGL-NGS-105096	1	353745	3
Hapmap34944-BES1_Contig627_1906	1	393248	4

ID.bt indica a coluna de nomes dos marcadores.

O GenSel ajusta o modelo misto que inclui os coeficientes de regressão aleatórios não correlacionados para cada loci representado no arquivo de genótipos. Todos os outros efeitos, exceto o residual, são assumidos como fixos e são especificados por meio do cabeçalho no **arquivo de fenótipo**. O modelo indicará os fatores a serem incluídos na análise, como apresentaremos a seguir.

No cabeçalho do arquivo de fenótipos, símbolos são usados no título para definir o tratamento que será dado pelo programa aos dados daquela coluna. O símbolo # logo após a palavra de um cabeçalho de coluna indica que os dados desse parâmetro não serão considerados. O símbolo $ significa que a coluna corresponde a variáveis classificatórias (valores alfanuméricos). Os fenótipos podem ser medidas diretas ou indiretas das características de interesse, com, por exemplo, valores genéticos estimados (EBV) ou derregredidos (DEBV) dos indivíduos genotipados a serem considerados nas análises de associação.

O arquivo de fenótipos deve conter a identificação de cada animal na primeira coluna e os dados da característica na segunda coluna, como mostra a tabela 3. A coluna indicada como *rinverse* denota valores que serão utilizados como elementos da diagonal da matriz de pesos ou ponderadores para análises que envolvem variância residual heterogênea.

Tabela 3. Exemplo de um arquivo de fenótipos utilizando-se informações de EBVs derregredidos e seus respectivos pesos (rinverse).

Animal	dEBV	rinverse
00000208FSB-M	-0.6499	1.8701
0000408FSB-F	0.4754	1.8698
0000608FSB-M	-0.490	1.8558
00000808FSB-F	-0.074	1.8698

Os resultados ou arquivos de saída do programa (*outputs*) são gerados na pasta de trabalho indicada e apresentam diferentes extensões como apresentado a seguir.

Passo-a-passo para as análises

Estimação de componentes de variâncias:

Para a obtenção dos valores genéticos dos animais, estimados e/ou derregredidos, é necessário que as diferenças entre o valor predito e verdadeiro de um animal sejam minimizadas e, para isso, os componentes de variância precisam ser estimados de forma acurada. Portanto, esse é primeiro passo da sequência de análises preliminares antes de rodar o Gensel.

É muito importante identificar um método estatístico adequado e que melhor reflita o comportamento biológico das características estudadas. Neste contexto, vários procedimentos de estimação dos componentes de variância já foram sugeridos para aplicação no melhoramento animal.

Alternativamente à aplicação dos métodos frequentistas nas análises genéticas, procedimentos bayesianos têm sido propostos na aplicação do modelo de touro (WANG et al., 1993), modelo animal (VAN TASSELL; VAN VLECK, 1996), modelo de limiar para dados categóricos (VAN TASSELL et al., 1998) e na obtenção da resposta à seleção (SORENSEN et al., 1994). Atualmente, a maioria dos programas de melhoramento utilizam equações de modelos mistos sob um modelo animal para estimação dos componentes de variância e valores genéticos, incluindo além dos efeitos aditivos de animal, os efeitos maternos aditivos, de ambiente permanente materno e/ou de ambiente permanente do animal. O modelo estatístico mais completo, em notação matricial, pode ser representado da seguinte forma:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wm} + \mathbf{Spe} + \mathbf{e}$$

Em que **y** é o vetor de observações, **X**, **Z**, **W** e **S** são as matrizes de incidência que relacionam as observações aos vetores **b** de efeitos fixos,

u de efeito aleatório genético direto, **m** de efeito aleatório genético materno e **pe** de efeito de ambiente permanente, e **e** é o vetor dos efeitos aleatórios residuais.

As pressuposições assumidas nesses modelos estão descritas abaixo:

$$\mathbf{E(y) = Xb,}\ \mathbf{E}\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{pe}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix},\ \mathbf{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{m}\\ \mathbf{pe}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\sigma}^2_{\mathbf{u}}\mathbf{A} & \boldsymbol{\sigma}_{\mathbf{um}}\mathbf{A} & \mathbf{0} & \mathbf{0}\\ \boldsymbol{\sigma}_{\mathbf{mu}}\mathbf{A} & \boldsymbol{\sigma}^2_{\mathbf{m}}\mathbf{A} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\boldsymbol{\sigma}^2_{\mathbf{pe}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\boldsymbol{\sigma}^2_{\mathbf{e}}\end{bmatrix}$$

Em que $\boldsymbol{\sigma}^2_{\mathbf{u}}$ é a variância genética aditiva, $\boldsymbol{\sigma}_{\mathbf{u}m} = \boldsymbol{\sigma}_{mu}$ é a covariância genética aditiva entre os efeitos diretos e maternos, $\boldsymbol{\sigma}^2_{m}$ é a variância genética aditiva para os efeitos maternos, $\boldsymbol{\sigma}^2_{pe}$ é a variância devido aos efeitos ambientais permanentes, $\boldsymbol{\sigma}^2_{e}$ é a variância residual, $\mathbf{A}$ é a matriz de parentesco de Wright e $\mathbf{I}$ é uma matriz identidade.

No Brasil, o primeiro estudo que utilizou a inferência bayesiana na estimação de componentes de variância, para características produtivas em bovinos, foi realizado por Magnabosco et al. (1997). A aplicação dos métodos de Cadeia de Markov de Monte Carlo (MCMC), dentre os quais se destaca a Amostragem de Gibbs (GS), é o que propicia a inferência bayesiana. O algoritmo de Gibbs é aplicado para gerar um valor para cada parâmetro desconhecido e apresenta fácil implementação, principalmente quando comparado a algoritmos baseados em processos não derivativos, uma vez que os resultados permitem uma inferência que gera distribuições posteriores marginais completas, a partir das quais são obtidas as estimativas dos componentes de variância e parâmetros genéticos.

Um conjunto de alternativas para a estimação dos componentes de variância são os programas da família Blupf90, utilizando-se o software Gibbs1f90 ou Gibbs2f90 (MISZTAL et al., 2002) para análise de

características com distribuição normal e o Thrgibbsf90 (MISZTAL et al., 2002) para dados categóricos. O tamanho das cadeias deve ser considerado de acordo com os dados analisados (p. ex.: 800.000 ciclos, aquecimento de 200.000 e tomada de amostras a cada 20 rodadas; ou 41.000 ciclos, com 1.000 de aquecimento e amostragens a cada 40 iterações etc.). Para estimação *a posteriori* utiliza-se o software Postgibbsf90 (MISZTAL et al., 2002). A convergência pode ser verificada através de inspeção gráfica, valores amostrados x iterações, e com o critério proposto por (GEWEKE, 1991) por meio do pacote BOA do programa R, versão 2.10.1 (SMITH, 2007).

Geralmente, antes da estimação dos componentes de (co)variâncias, são realizadas consistências no conjunto de dados, como, por exemplo, para as características contínuas são excluídos grupos contemporâneos com menos de cinco animais e medidas com 3.5 desvios-padrão acima ou abaixo da média do seu grupo e para as características categóricas, são eliminados grupos sem variabilidade.

Estas análises prévias são efetuadas geralmente usando métodos tradicionais, em que são utilizados somente dados de pedigree e fenótipo. Embora nada impeça o uso de genótipos desde esta etapa, em geral, estes são incorporados mais adiante somente nos estudos de associação genômica ampla (GWAS).

Correção dos fenótipos

Sabe-se que as avaliações de seleção genômica ampla (GWS) requerem três populações definidas: treinamento, validação e seleção. As equações de predição de valores genéticos genômicos, obtidos pela população de treinamento, são testadas para verificar suas acurácias nas amostras independentes, referentes à população de validação. Para computar essa acurácia, os valores genéticos genômicos são preditos e submetidos a uma análise de correlação com os valores fenotípicos observados. Esses fenótipos, por sua vez, devem ser corrigidos para os efeitos ambientais e dos genitores, trabalhando-se basicamente com o

efeito da segregação mendeliana derregredida, já que o dado ideal para a população de treinamento deve ser o mérito genético verdadeiro de indivíduos não aparentados. Sem essa correção dos valores fenotípicos, a acurácia da validação em uma amostra independente da população e, também, em indivíduos de outras gerações poderá ser subestimada (RESENDE et al., 2012). Desta forma, os EBVs também utilizados para análises de associação, podem ser derregredidos (DEBV) conforme Garrick et al. (2009), sendo então considerados equivalentes às informações fornecidas pelos próprios registros dos indivíduos e dos seus descendentes.

Esses DEBVs, por sua vez, serão utilizados nos estudos de GWAS com base no seguinte modelo de substituição alélica implementado no software Gensel versão 4.0 (FERNANDO; GARRICK, 2009):

$$y_i = \mu + \sum_{j=1}^{k} M_{ij}\, g_j + e_i$$

Onde

y_i = DEBV do indivíduo *i*

μ = Média da população

k = Número de marcadores SNPs no painel

M_{ij} = Número de alelos (0, 1 ou 2, ou seja, números de alelo B no A/B *calling system*) do marcador *j* no indivíduo *i*

g_j = Efeito aleatório para o marcador *j*

e_i = Resíduo, com variância heterogênea (dependendo da confiabilidade da informação do DEBV do indivíduo *i*)

Variáveis categóricas

Frequentemente, características utilizadas como critério de seleção são avaliadas fenotipicamente através da atribuição de escores visuais ou categorias, como no caso da pigmentação ocular, pelame, caracterização racial, facilidade de parto etc. Entretanto, por não apresentarem distribuição normal, certos cuidados devem ser tomados quanto à metodologia empregada para a avaliação genética destas características. As características categóricas, desta forma, violam várias pressuposições existentes na metodologia dos modelos mistos e, por isso, não devem ser analisadas através de um modelo linear, e sim por um modelo de limiar (*threshold*) (GIANOLA, 1982).

O modelo animal de limiar assume que, apesar das categorias terem uma distribuição observável descontínua limitada pelos limiares (Figura 1), estas possuem uma variação subjacente, que é de origem genética e ambiental, contínua e com distribuição normal (FALCONER; MACKAY, 1996). Com o uso do software Thrgibbs1f90 é possível realizar a predição de valores genéticos de forma mais fidedigna para variáveis categóricas através do modelo de limiar.

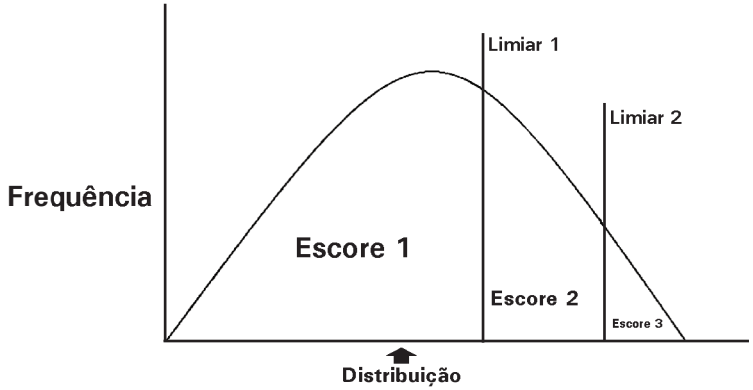


Figura1. Distribuição de uma população em uma característica de limiar contendo três escores e dois limiares

Escolha do método bayesiano para estimação dos efeitos dos marcadores

Dois fatores principais discriminam alguns modelos Bayesianos comumente utilizados:

1) o modelo pode ajustar efeitos aleatórios para todos os marcadores simultaneamente (Bayes A), ou assumir um modelo de mistura que classifica fatores aleatórios como pertencentes a uma de duas distribuições utilizadas (Bayes B e C). Em um modelo de mistura, a uma proporção dos marcadores (π) atribui-se uma distribuição com efeito nulo, e somente o restante dos marcadores (1-π) que são amostrados como tendo efeitos diferentes de zero e incluídos no modelo simultaneamente.

2) o parâmetro de variância assumido para os efeitos aleatórios pode ser especificado para cada marcador incluído (ou seja, BayesA, BayesB), ou comum a todos os marcadores (por exemplo, RRBLUP- *ridge-regression best linear unbiased prediction*).

No GenSel, especifica-se o valor de π no cartão de parâmetros através da palavra chave *probFixed* usando, por exemplo, o valor 0.99. O parâmetro π, que indica a proporção de marcadores a serem considerados com efeitos nulos na análise (99%), e pode, se for do interesse do usuário ser estimado utilizando os métodos Bayesianos numa extensão do BayesC, referida como BayesCπ. Todos estes métodos podem ser especificados em cada análise na opção *analysisType Bayes* e *bayesType BayesA, BayesB, BayesC ou BayesC* do cartão de parâmetros (*input*).

O cartão de parâmetros referente aos inputs das analises neste programa é definido como segue o exemplo:

```
markerFileName      ./input/genotypes.txt   #indicar o local e nome do arquivo de genótipos
phenotypeFileName   ./input/phenpredict_rd.txt  # indicar o local e nome do arquivo de fenótipos
```

mapOrderFileName ./input/map.txt *# indicar o local e nome do arquivo de mapa dos marcadores*

linkageMap BTA *# depende do arquivo de mapa: genoma bovino por exemplo, UMD ou BTA*

addMapInfoToMarkers yes *#Resulta na inclusão dos cromossomos e posições em pares de bases ao output*

analysisType Bayes *#Indicar o tipo de análise (ou Predict, StepWise, generateData e LD)*

bayesType BayesB *#Indicar o método de análise Bayesiana:* BayesA, BayesB, BayesC ou BayesCPi

chainLength 41000 *#número total de iterações*

burnIn 1000 *#número de ciclos a serem ignorados*

outputFreq 40 *#número de iterações para mostrar resultados (gerar amostragens). Default = 100*

probFixed 0.99 *#indicar proporção de marcadores com efeitos nulos (π); neste caso, 99%*

varGenotypic 0.0182 *#indicar variância genética da característica estimada a priori*

varResidual 0.0715 *#indicar variância residual da característica estimada a priori*

numMHIter 10 *# número de iterações Metropolis-Hastings usadas na determinação da variância para cada covariável pelos métodos BayesA ou BayesB. O valor 1 é considerado para BayesA e 10 para BayesB (default).*

mcmcSamples yes *#indicar número de iterações Metropolis-Hastings utilizadas para determinar a variância de cada covariável nos métodos BayesA e BayesB*

plotPosteriors yes *#indicar se distribuições a posteriori serão plotadas no output*

windowBV yes *#formar janelas de 1 Mb com base no arquivo de mapas (computa variâncias das janelas). Caso não utilize mapa, o número de marcadores em cada janela pode ser definido (*windowWidth *5, por exemplo)*

FindScale no *#indicar se o fator de escala será estimado a partir dos dados (yes) ou se π será determinado a priori (no)*

modelSequence yes *#gerar um arquivo extra de output listando as covariáveis incluídas no modelo a cada amostragem post burnin*

isCategorical no *#indicar se serão utilizados dados categóricos (modelo de limiar)*

De acordo com os métodos BayesA e BayesB, cada SNP é considerado como tendo um efeito de variância específico, que por sua vez apresenta uma distribuição Qui-quadrado invertida $X^{-2}(v, S)$ com graus de liberdade v e escala S. Entretanto, a especificação *a priori* para os efeitos dos SNPs em BayesB permite que uma proporção de marcadores tenham efeitos nulos, com probabilidade fixa de π, enquanto os outros marcadores possuem efeitos com distribuição normal e variância locus-específica, com probabilidade igual a $1-\pi$. Por outro lado, com o método BayesA todos os SNPs são considerados no modelo, ou seja $\pi = 0$, para cada ciclo de Markov Chain Monte Carlo (MCMC).

O modelo estatístico utilizado para as análises Bayesianas que estimar os efeitos dos marcadores pode ser demonstrado por:

$$y = \sum_{i=1}^{k} \delta_i \mathbf{m}_i g_i + e,$$

Onde y é o vetor de fenótipos (DEBV);

k é o número total de SNPs analisados;

δ_i é o indicador de quando o SNP i serpa incluído $(\delta_i = 1)$ ou excluído

$(\delta_i = 0)$ do modelo para dada interação da MCMC;

$\mathbf{m}_i$ é o vetor de genótipos de um marcador i, no caso do Gensel codificado como -10/0/10 em vez do tradicional 0/1/2;

g_i é o efeito de substituição alélica do marcador i com sua própria variância σ^2_{gi} e um efeito nulo a priori com probabilidade π ou um efeito diferente de zero com probabilidade $1-\pi$, e é o vetor de resíduos aleatórios normalmente distribuídos.

No modelo BayesCπ, a probabilidade de um SNP ter efeito nulo é tratada como desconhecida e uma variância de efeito comum é assumida para todos os SNPs, enquanto para BayesA, δ_i é sempre igual a 1. Geralmente, testa-se primeiramente a estimação dos efeitos dos marcadores para todos os indivíduos com o método BayesC e estimasse o valor de π na população e característica de interesse, assumindo a priori que π=0.5, como proposto por Sun et al. (2011) e outros (OLIVEIRA et al., 2014). Posteriormente, utiliza-se o valor estimado (por.ex. média a posteriorior) em análises com o método BayesB nas quais pode ser testada a estimação dos efeitos dos marcadores fixando-se π no valor estimado através do BayesCπ.

O método BayesC também tem sido utilizado em estudos de associação via GenSel e.g. (SCHURINK et al., 2012). Este método assume variância comum para todos os SNPs no modelo e parece ser menos sensível às variâncias genéticas determinadas *a priori*, comparado com o método BayesB como descrito por Meuwissen et al. (2001). Entretanto, diversos autores têm preferido a aplicação do método Bayes B e.g (CARDOSO et al., 2015; ONTERU et al., 2013; SUN et al., 2011; WOLC et al., 2012) devido às melhores acurácias de predição obtidas, bem como devido ao fato de considerarem diferentes variâncias genéticas para cada marcador, o que se aproxima das situações biológicas para características quantitativas, permitindo maior variação na magnitude dos efeitos de cada marcador (alguns podem ter efeitos muito pequenos e outros efeitos maiores).

Todas as análises no GenSel irão gerar um resumo dos resultados de acordo com o nome dado à análise, por exemplo, "run_BBpi99.out1" (ver quadro a abaixo). O sufixo "1" indica o número de vezes que genótipos e fenótipos foram lidos e que passaram pelo processo de filtro para subsequente análise indicada (no caso, BayesB com o valor π igual a 99% ou 0,99). Análises repetidas dos mesmos dados apresentarão sufixos na sequência (2,3,4,5 etc.).

```
        read 3455 lines in text genotype file ./input/genotypes.txt
        Binary genotype file ./input/genotypes.txt.newbin created
        Merging animals with Phenotypes in ./input/phenotypes.txt with
Genotypes in ./input/genotypes.txt.newbin


       mean of 2pq = 0.359515
       Number of records read in genotype file:  3455
       Number of records in phenotype file:      3455
       Number of matching records:               3455
       Number of marker loci in model:           41045
       Number of fixed levels in model:          0
        Warning :  no fixed effects will inflate SSE and reduce heritability
if mean is not zero


        Model Info:


           Model Term       Number of Levels
               Markers               41045
```

O exemplo acima apresenta parte do resumo de output gerado a partir de 3.455 animais com dados de fenótipos e 41.045 genótipos registrados para cada indivíduo. Não foram identificados efeitos fixos uma vez já considerados quando calculados os valores de EBV (ver `Warning :`).

Regiões cromossômicas ou janelas e efeitos de marcadores

Os resultados dos estudos de associação ampla obtidos por meio do GenSel são apresentados de duas formas principais: a) Através da porcentagem da variância genética explicada por cada segmento (janela) cromossômico, composto por um conjunto de marcadores adjacentes (Tabela 1) ou b) através dos efeitos referentes a cada marcador, em que parâmetros da estatística bayesiana (*Model Frequency* e *t-like*) e as estimativas de seus efeitos são apresentados individualmente por marcador (Tabela 2). Estudos de associação envolvem a localização em escala genômica por fragmentos cromossômicos que podem ser caracterizados regredindo os dados de fenótipos por genótipos acessados. Para encontrar associações significativas, a estratégia bayesiana utiliza amostragens das distribuições a posteriori dos efeitos de genótipos.

Utilizando-se milhares de SNPs (marcadores) a fim de inferir associações, é esperado que marcadores próximos uns aos outros estivessem altamente correlacionados, e que nem todo marcador apresenta forte associação com a característica. Considerando um grupo de marcadores que flanqueiam determinada região genômica, ou janela cromossômica, espera-se que a maior parte da variabilidade de determinado loco de característica seja explicada conjuntamente. Assim, assumindo modelos de regressão múltipla, as inferências de associações podem ser baseadas em janelas cromossômicas, e não em marcadores pontualmente.

Relativamente ao total da variância genética $\widetilde{\sigma_a^2}$ existente, aquela fração explicada pelos marcadores $\widetilde{\sigma_{aw}^2}$ pertencentes a uma janela (w) pode ser definido por:

Em que,

$$q_w = \widetilde{\sigma^2_{aw}} \Big/ \widetilde{\sigma^2_a}$$

$$\widetilde{\sigma^2_{aw}} = \sum_{j=1}^{n} a^2_{wj} \Big/ n - \left[\left(\sum_{j=1}^{n} a_{wj} \Big/ n\right)\right]^2 \text{ e}$$

$$\widetilde{\sigma^2_a} = \sum_{j=1}^{n} a^2_j \Big/ n - \left[\left(\sum_{j=1}^{n} a_j \Big/ n\right)\right]^2$$

Sendo que o componente do valor genotípico atribuído à janela cromossômica w é definido por $a_w = M_w g_w$ onde M_w é a matriz de genótipos dos marcadores (covariáveis) da janela w, e g_w é o vetor de efeitos destes marcadores.

Variâncias genéticas das janelas são computadas multiplicando-se o número de alelos representados pelas covariávies (SNPs) consecutivas de uma janela, pelos seus respectivos efeitos médios de substituição *a posteriori*. Após computar os valores genéticos destes SNPs (dentro de cada janela) para todos os animais na população, a variância destes valores genéticos é então calculada. Janelas que contribuírem com maiores variâncias genéticas são consideradas aquelas mais associadas à característica (PETERS et al., 2012).

No GenSel, as janelas cromossômicas são construídas baseando-se nos cromossomos e posições dos marcadores, em pares de bases, de acordo com o arquivo de mapa (Tabela 2). Cada janela consiste de marcadores localizados no mesmo cromossomo dentro de aproximadamente 1 Mb. A palavra *yes* após o parâmetro *windowBV* (na especificação do arquivo de

parâmetros) determina a formação destas janelas e o cômputo de valores de amostragens para as variâncias dos valores genômicos (*genomic breeding values*, GBV) para cada janela. A frequência com que estas amostragens são computadas é determinada na especificação do parâmetro *outputFreq*. As amostragens individuais de variâncias das janelas, expressas como a proporção de variância genômica, são gravadas em um arquivo de saída (*output*) para propósitos de inferências (Tabela 1).

Um exemplo do resumo da estatística para resultados de variâncias das janelas (*output*) está na Tabela 4, considerando uma análises BayesB, onde $\pi = 0,99$, utilizando marcadores numericamente codificados (*start* – primeiro SNP representativo da janela e *end*- último SNP representativo da janela). Neste exemplo, as quatro primeiras janelas, sequencialmente enumeradas ao longo do genoma, são apresentadas já em ordem decrescente em termos de contribuição da variância explicada por cada uma das janelas de 1 Mb. A janela mais representativa explica 1,67% da variância genética (*%Var*) e consiste de 17 marcadores (*#SNPs*). A coluna *p>0* reflete a proporção de modelos em que esta janela foi incluída e, portanto, explicando mais de 0% da variância genética. A coluna *p>Average* reflete a proporção de modelos em que esta janela explicou mais do que o montante da variância que poderia ser explicada se todas as janelas apresentassem o mesmo efeito. Note que as três primeiras janelas, individualmente, explicam mais de 1% da variância genética relacionada à característica avaliada, mas conjuntamente, as quatro explicam 4,83%.

O GenSel também plota gráficos das distribuições a posteriori das 10 maiores janelas em termos de porcentagens de variâncias genéticas explicadas, basta especificar yes ao parâmetro plotPosteriors.

Janelas consideradas significativas ou mais informativas para determinada característica avaliada podem ser consideradas QTLs (*Quantitative Trait Loci*). Segundo (ONTERU et al., 2013), QTLs candidatos podem ser considerados se a porcentagem esperada de variância genética contida em uma janela for pelo menos 5 vezes maior do que 100%/n (n= número de janelas propostas considerando todo o

genoma. Para o genoma bovino UMD 3.1 são consideradas 2.519 janelas, ou seja, próximo a 0,04%). Portanto, uma janela de 1 Mb que explica pelo menos 0,2% da variância genética pode ser considerada um QTL. A Figura2 exemplifica uma forma de representação das janelas identificadas após a análise. Neste gráfico, a porcentagem de variação explicada por cada janela é representada em um Manharttan plot.

.

Tabela 4. Exemplo do arquivo de saída (*output*) do GenSel referente a parâmetros de quatro janelas cromossômicas identificadas

Window	start	end	#SNPs	%Var	Cum%Var	p > 0	p > Average	map_pos0	map_posn	chr_Mb
1621	27751	115263	17	1.67	1.67	0.841842	0.82983	15_37063494	15_37958705	15_37
1319	112243	12954	17	1.19	2.86	0.714715	0.698699	11_101022540	11_101963555	11_101
1164	1854	24556	27	3.99	3.99	0.68969	0.665666	10_51001410	10_51980270	10_51
1553	115527	1112	15	4.83	4.83	0.653654	0.630631	14_54008792	14_54947391	14_54

A Tabela 5 ilustra uma apresentação de parâmetros referentes a cada marcador do arquivo de mapa/genótipos. Suponhamos que a análise de 50.000 marcadores SNPs considerou um valor de π igual a 0,99, o que resultaria em média 500 SNPs com efeitos amostrados em cada ciclo da cadeia de MCMC. Desta forma, uma estatística importante de considerar, é a proporção de amostragens que determinado locus/marcador obteve efeito não nulo ($1-\pi$). Esta estatística é denominada de model frequency (proporção de cadeias que incluíram determinado marcador no modelo). Um marcador escolhido aleatoriamente para $\pi = 0,95$, seria esperado a priori ter 0,95 amostragens com efeito nulo e 0,05 amostragens com efeito diferente de zero. Após as análises, marcadores mais informativos apresentariam maiores valores de model frequency.

A segunda coluna da Tabela 5 (*effect*) identifica a média *a posteriori* dos efeitos de substituição alélica (trocar um alelo B pelo alelo A). A coluna *EffectVar* representa a média *a posteriori* das variâncias dos efeitos do locus. *Window frequency* (*WindowFreq*) indica a proporção de modelos que incluíram algum locus na janela, enquanto *gene frequency* (*GenFreq*) é a frequência do alelo B. *GenVar* é a contribuição marginal deste locus para a variância genética, computado a partir da frequência alélica e média *a posteriori* do efeito de substituição. *EffectDelta 1* é a média *a posteriori* do efeito do locus apenas daqueles modelos que o incluíram, diferente da coluna *Effect* que considera todos os modelos, inclusive aqueles que o marcador não foi incluído e, portanto, teve efeito nulo. Este valor, multiplicado pelo *model frequency* resulta no efeito da coluna 2. *SDDelta 1* é o desvio padrão dos efeitos amostrado naquelas iterações onde este locus foi incluído. O *t-like* é média *a posteriori* do efeito amostrado dividido pelo seu desvio padrão quando incluído no modelo. Este parâmetro não segue a distribuição *t* mas pode ser também um indicador de marcadores com altos efeitos consistentemente amostrados. A última coluna, *shrink*, informa o encurtamento (ou redução em direção a zero) da estimativa de modelo misto para aquele locus relativo à sua estimativa de quadrados mínimos.

Desequilíbrio de ligação

Considerando a utilização de marcadores genéticos em larga escala para estudos genômicos, podemos definir desequilíbrio de ligação (LD) como a propriedade na qual um alelo de um SNP é herdado em conjunto ou está correlacionado com o alelo de outro SNP dentro de uma população (BUSH; MOORE, 2012). Para geneticistas de populações, o LD reflete uma tentativa matemática de descrever mudanças na variação genética dentro de uma população pelo tempo. Isto porque, eventos de recombinações acontecem a cada geração quebrando os blocos de segmentos cromossômicos. Em uma população de tamanho fixo sob acasalamentos ao acaso, quebras de partes justapostas dos cromossomos (com alelos ligados) aumentam, devido aos eventos de recombinações aleatórias consecutivos, até que todos os alelos de toda a população estejam em equilíbrio de ligação ou independentes.

Todas as medidas de LD propostas baseiam-se na diferença existente entre a frequência observada de co-ocorrências para dois alelos (ou seja, um haplótipo de dois marcadores) e a frequência esperada se os dois marcadores fossem independentes. As duas medidas mais utilizadas são a D' e r^2 (ALTSHULER et al., 2005; DEVLIN; RISCH, 1995). Para propósitos de análises genéticas, o LD é geralmente reportado em termos de r^2, uma medida estatística de correlação. Assim, maiores valores de r^2 (0-1) indicam que dois SNPs transmitem informações semelhantes, uma vez que um alelo de um deles é sempre observado com um alelo de um segundo SNP; então somente um dos dois precisa ser genotipado para capturar a variação alélica. Essa informação pode ser usada para reduzir o número de marcadores e o custo da aplicação da informação genômica no melhoramento animal.

Existem diferentes programas computacionais, pacotes/funções do R que permitem o cálculo do desequilíbrio de ligação entre marcadores a partir de uma matriz de genótipos, bem como representações gráficas elegantes dos blocos de ligação (CLAYTON, 2012; PURCELL et al., 2007; SHIN et al., 2006). A figura 3 exemplifica uma forma de representação de análise de LD de uma janela cromossômica pelo pacote do R LDheatmap (SHIN et al., 2006).

Vale ressaltar que é importante entender como são mensuradas as frequências de haplótipos de determinada amostra, pois cada SNP é genotipado independentemente e a fase de ligação ou cromossomo de origem para cada alelo é desconhecida. Existem métodos bem documentados para inferir fases de haplótipos, quando utilizados ao invés de marcas únicas (BROWNING; BROWNING, 2011).

Tabela 5. Exemplo do arquivo de saída (*output*) do GenSel referente a parâmetros de três loci analisados.

ID.bt	Effect	EffectVar	ModelFreq	Windowfreq	Genefreq	GenVar
Hapmap43437-BTA-101873	7.3e-06	5.9e-07	0.0062	0.0444	0.355	2.4e-11
ARS-BFGL-NGS-66449	-4.7e-05	1.1e-06	0.0093	0.1006	0.046	2.0e-10
BTB-01560502	5.0e-05	9.4e-07	0.0099	0.0868	0.887	5.2e-10
ID.bt	**EffectDelta1**	**SDDelta1**	**t-like**	**shrink**	**map-posn**	**Window**
Hapmap43437-BTA-101873	1.1e-03	4.9e-03	0.239	0.778	1_135098	2
ARS-BFGL-NGS-66449	-5.1e-03	9.8e-03	0.525	0.331	1_571340	5
BTB-01560502	5.1e-03	8.1e-03	0.630	0.523	1_5880498	48

Seleção de marcadores mais informativos

Marcadores SNPs selecionados especificamente para capturar a variação em diferentes regiões do genoma são chamados de *tag SNPs,* pois alelos destes SNPs "marcam" trechos ou pedaços em LD. Não podemos esquecer que padrões de LD existem entre populações específicas, entre raças, assim como *tag SNPs* selecionados a partir de uma população podem não ser os mesmos em diferentes populações.

As estratégias de identificação destas marcas mais informativas (*tag SNPs*) tanto dentro de determinadas populações, como para característica específicas, além das diferentes estratégias de validação destes, são cruciais em estudos de associação.

De acordo com os resultados de parâmetros bayesianos gerados pelo programa GenSel, e descritos anteriormente neste documento, a seleção de marcas pode basear-se nos resultados de Model Frequency (MF, proporção de cadeias que incluíram determinado marcador no modelo), t-like (TL, a razão do efeito médio a posteriori somente das cadeias que incluíram determinado marcador no modelo, dividido pelo desvio padrão das distribuições destes efeitos), desequilíbrio de ligação (DL) e frequência do alelo menor (MAF - minor allele frequency). A seguir, descrevemos as etapas de uma estratégia de seleção de marcadores desenvolvida por Sollero et al. (2017) para propor painéis de baixa densidade para uso na seleção genômica e os comandos utilizados no Pacote R (R FOUNDATION FOR STATISTICAL COMPUTING, 2013):

1) Selecionar janelas mais informativas (top Windows): janelas que explicaram pelo menos 5 x 100%/n (n = número de janelas de 1Mb total no genoma) da variância genética da característica (ONTERU et al., 2013);

2) Determinar o valor mínimo de MF (minMF) encontrado entre todos os SNPs que apresentaram os valores máximos de MF dentro de cada top window;

3) Selecionar somente SNPs incluídos nas top Windows com valores acima do minMF;

4) Determinar o valor mínimo para TL (minTL) observado entre os SNPs selecionados no passo anterior;

5) Selecionar, dentre todos os SNPs das top Windows, além daqueles selecionados de acordo com minMF (etapa 4), aqueles acima no minTL (minMF + minTL);

6) Avaliar o DL (previamente calculado) entre pares de marcas da lista final (minMF + minTL) definindo um limiar para exclusão de SNP; se acima deste valor (ex. DL 0,4);

7) Utilizar MAF como critério de desempate entre SNPs; permanecendo aqueles com maior valor.

Obs: um exemplo completo de Script R para seleção de marcadores é apresentado no Anexo 1.

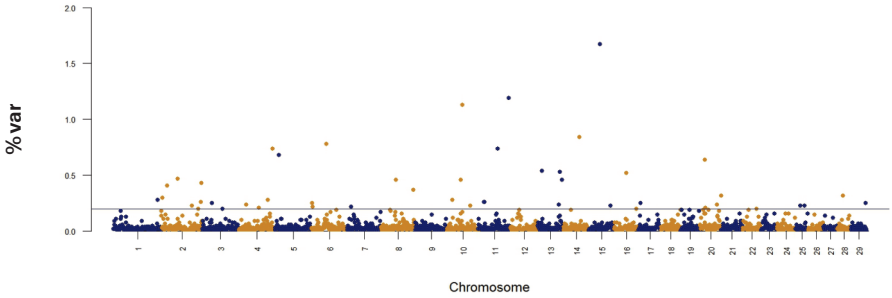


Figura 2. Exemplo de um Manhattan Plot das porcentagens de variâncias genéticas (%Var) explicada por cada janela (região cromossômica) para determinada característica no genoma bovino, como resultado do programa GenSel, apresentadas em ordem cromossômica (BTA1:BTA29). A linha define o limiar (0,2%), acima do qual, janelas são consideradas potenciais QTL. Eixo x apresenta os cromossomos autossomais de bovinos.

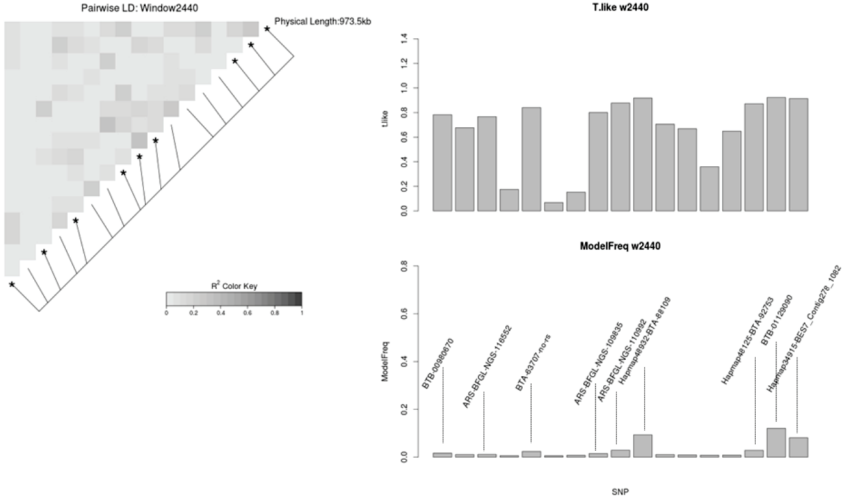


Figura 3. Apresentação gráfica de uma janela cromossômica hipotética identificando parâmetros t-like e Model Frequency para cada marcador, além do heatmap representado pelas correlações entre marcadores (desequilíbrio de ligação). (B) indica marcador mais informativos segundo os critérios de seleção de marcas estabelecidos; (*) indica marcador excluído devido ao critério de desequilíbrio de ligação.

Considerações finais

Estudos de associação genômica ampla para características zootécnicas têm trazido oportunidades especialmente para a descoberta de variantes causais. Além disso, não se pode esquecer que o conceito fundamental de *genome breeding values* (GBV) ou valores genômicos é que estes resultam da soma dos efeitos aditivos dos QTLs de todos os QTLs que influenciam determinada característica, tornando ainda mais importante a escolha dos métodos estatísticos na estimação dos efeitos dos marcadores antes das predições dos valores genômicos.

As estratégias de seleção de marcadores (tag SNP) também vêm favorecendo a implantação e aplicação de testes genômicos, os quais abrem novas oportunidades quanto à praticidade e à redução dos custos de avaliações genômicas de animais em larga escala, propondo chips de baixas densidades de marcadores.

Neste documento foram apresentados conceitos e comandos para facilitar o emprego do Software Gensel nos estudos de associação genômica ampla (GWAS) e programação em R para formar painéis de baixa densidade usando os resultados do Gensel. Como isso, espera-se que os leitores possam usar esta publicação como guia para identificar regiões genômicas relacionadas ao desempenho de animais de produção, bem como identificar os marcadores para compor painéis de baixa densidade a serem usados na aplicação prática da seleção genômica no melhoramento animal.

Referências

ALTSHULER, D.; BROOKS, L. D.; CHAKRAVARTI, A.; COLLINS, F. S.; DALY, M. J.; DONNELLY, P.; GIBBS, R. A.; BELMONT, J. W.; BOUDREAU, A.; LEAL, S. M.; HARDENBOL, P.; PASTERNAK, S.; WHEELER, D. A.; WILLIS, T. D.; YU, F. L.; YANG, H. M.; ZENG, C. Q.; GAO, Y.; HU, H. R.; HU, W. T.; LI, C. H.; LIN, W.; LIU, S. Q.; PAN, H.; TANG, X. L.; WANG, J.; WANG, W.; YU, J.; ZHANG, B.; ZHANG, Q. R.; ZHAO, H. B.; ZHAO, H.; ZHOU, J.; GABRIEL, S. B.; BARRY, R.; BLUMENSTIEL, B.; CAMARGO, A.; DEFELICE, M.; FAGGART, M.; GOYETTE, M.; GUPTA, S.; MOORE, J.; NGUYEN, H.; ONOFRIO, R. C.; PARKIN, M.; ROY, J.; STAHL, E.; WINCHESTER, E.; ZIAUGRA, L.; SHEN, Y.; YAO, Z. J.; HUANG, W.; CHU, X.; HE, Y. G.; JIN, L.; LIU, Y. F.; SHEN, Y. Y.; SUN, W. W.; WANG, H. F.; WANG, Y.; WANG, Y.; WANG, Y.; XIONG, X. Y.; XU, L.; WAYE, M. M. Y.; TSUI, S. K. W.; XUE, H.; WONG, J. T. F.; GALVER, I. L. M.; FAN, J. B.; MURRAY, S. S.; OLIPHANT, A. R.; CHEE, M. S.; MONTPETIT, A.; CHAGNON, F.; FERRETTI, V.; LEBOEUF, M.; OLIVIER, J. F.; PHILLIPS, M. S.; ROUMY, S.; SALLEE, C.; VERNER, A.; HUDSON, T. J.; FRAZER, K. A.; BALLINGER, D. G.; COX, D. R.; HINDS, D. A.; STUVE, L. L.; KWOK, P. Y.; CAI, D. M.; KOBOLDT, D. C.; MILLER, R. D.; PAWLIKOWSKA, L.; TAILLON-MILLER, P.; XIAO, M.; TSUI, L. C.; MAK, W.; SHAM, P. C.; SONG, Y. Q.; TAM, P. K. H.; NAKAMURA, Y.; KAWAGUCHI, T.; KITAMOTO, T.; MORIZONO, T.; NAGASHIMA, A.; OHNISHI, Y.; SEKINE, A.; TANAKA, T.; TSUNODA, T.; DELOUKAS, P.; BIRD, C. P.; DELGADO, M.; DERMITZAKIS, E. T.; GWILLIAM, R.; HUNT, S.; MORRISON, J.; POWELL, D.; STRANGER, B. E.; WHITTAKER, P.; BENTLEY, D. R.; DALY, M. J.; BAKKER, P. I. W. de; BARRETT, J.; FRY, B.; MALLER, J.; MCCARROLL, S.; PATTERSON, N.; PE'ER, I.; PURCELL, S.; RICHTER, D. J.; SABETI, P.; SAXENA, R.; SCHAFFNER, S. F.; VARILLY, P.; STEIN, L. D.; KRISHNAN, L.; SMITH, A. V.; THORISSON, G. A.; CHEN, P. E.; CUTLER, D. J.; KASHUK, C. S.; LIN, S.; ABECASIS, G. R.; GUAN, W. H.; MUNRO, H. M.; QIN, Z. H. S.; THOMAS, D. J.; MCVEAN, G.; BOTTOLO, L.; EYHERAMENDY, S.; FREEMAN, C.; MARCHINI, J.; MYERS, S.; SPENCER, C.; STEPHENS, M.; CARDON, L. R.; CLARKE, G.; EVANS, D. M.; MORRIS, A. P.; WEIR, B. S.; TSUNODA, T.; MULLIKIN, J. C.; SHERRY, S. T.; FEOLO, M.; ZHANG, H. C.; ZENG, C. Q.; ZHAO, H.; MATSUDA, I.; FUKUSHIMA, Y.; MACER, D. R.; SUDA, E.; ROTIMI, C. N.; ADEBAMOWO, C. A.; AJAYI, I.; ANIAGWU, T.; MARSHALL, P. A.; NKWODIMMAH, C.; ROYAL, C. D. M.; LEPPERT, M. F.; DIXON, M.; PEIFFER, A.; QIU, R. Z.; KENT, A.; KATO, K.; NIIKAWA, N.; ADEWOLE, I. F.; KNOPPERS, B. M.; FOSTER, M. W.; CLAYTON,

E. W.; MUZNY, D.; NAZARETH, L.; SODERGREN, E.; WEINSTOCK, G. M.; WHEELER, D. A.; YAKUB, I.; GABRIEL, S. B.; RICHTER, D. J.; ZIAUGRA, L.; BIRREN, B. W.; WILSON, R. K.; FULTON, L. L.; ROGERS, J.; BURTON, J.; CARTER, N. P.; CLEE, C. M.; GRIFFITHS, M.; JONES, M. C.; MCLAY, K.; PLUMB, R. W.; ROSS, M. T.; SIMS, S. K.; WILLEY, D. L.; CHEN, Z.; HAN, H.; KANG, L.; GODBOUT, M.; WALLENBURG, J. C.; ARCHEVEQUE, P. L.; BELLEMARE, G.; SAEKI, K.; WANG, H. G.; AN, D. C.; FU, H. B.; LI, Q.; WANG, Z.; WANG, R. W.; HOLDEN, A. L.; BROOKS, L. D.; MCEWEN, J. E.; BIRD, C. R.; GUYER, M. S.; NAILER, P. J.; WANG, V. O.; PETERSON, J. L.; SHI, M.; SPIEGEL, J.; SUNG, L. M.; WITONSKY, J.; ZACHARIA, L. F.; KENNEDY, K.; JAMIESON, R.; STEWART, J. A haplotype map of the human genome. **Nature**, London, v. 437, n. 7063, p. 1299–1320, Oct. 2005.

BOICHARD, D.; CHUNG, H.; DASSONNEVILLE, R.; DAVID, X.; EGGEN, A.; FRITZ, S.; GIETZEN, K. J.; HAYES, B. J.; LAWLEY, C. T.; SONSTEGARD, T. S.; VAN TASSELL, C. P.; VANRADEN, P. M.; VIAUD-MARTINEZ, K. A.; WIGGANS, G. R. Design of a bovine low-density SNP array optimized for imputation. **PLoS One**, San Francisco, v. 7, n. 3, e34130, Mar. 2012.

BROWNING, S. R.; BROWNING, B. L. Haplotype phasing: existing methods and new developments. **Nature Reviews Genetics**, London, v. 12, n. 10, p. 703–714, Oct. 2011.

BUSH, W. S.; MOORE, J. H. Genome-wide association studies. **PLoS Computational Biology**, San Francisco, v. 8, n. 12, p. e1002822, Dec. 2012.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, Viçosa, MG, v. 38, p. 64–71, jul. 2009. Número especial.

CAMPOS, G. de los; GIANOLA, D.; ALLISON, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. **Nature Reviews Genetics**, London, v. 11, n. 12, p. 880–886, Dec. 2010.

CARDOSO, F. F.; GOMES, C. C. G.; SOLLERO, B. P.; OLIVEIRA, M. M.; ROSO, V. M.; PICCOLI, M. L.; HIGA, R. H.; YOKOO, M. J.; CAETANO, A. R.; AGUILLAR, I. Genomic prediction for tick resistance in Braford and Hereford cattle. **Journal of Animal Science**, Savoy, v. 93, n. 6, p. 2693-2705, June 2015.

CLAYTON, D. **SnpStats**: SnpMatrix and XSnpMatrix classes and methods. Version 1.10.0. [S.l.]: Bioconductor, 2012. Disponível em: <https://bioconductor.org/packages/2.12/bioc/html/snpStats.html>. Acesso em: 30 out. 2013.

DEVLIN, B.; RISCH, N. A comparison of linkage disequilibrium measures for fine-scale mapping. **Genomics**, San Diego, v. 29, n. 2, p. 311–322, Sept. 1995.

ELSIK, C. G.; TELLAM, R. L.; WORLEY, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, Washington, DC, v. 324, n. 5926, p. 522–528, Apr. 2009.

FALCONER, D. S.; MACKAY, T. F.C.; FRANKHAM, R. Introduction to quantitative genetics. (4th ed. London: Longman, 1996. 464 p.

FARIA, C. U. de; MAGNABOSCO, C. de U.; REYES, A. de los; LÔBO, R. B.; BEZERRA, L. A. F. Inferência bayesiana e sua aplicação na avaliação genética de bovinos da raça nelore: revisão bibliográfica. **Ciência Animal Brasileira**, Goiânia, v. 8, n. 1, p. 75–86, jan./mar. 2007.

FERNANDO, R. L.; GARRICK, D. J. **GenSel**: user manual for a portfolio of genomic selection related analyses. Ames: Iowa State University, 2009. 24 p. Disponível em: <https://www.biomedcentral.com/content/supplementary/1471-2105-12-186-s1.pdf>. Acesso em: 30 nov. 2016.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, London, v. 41, 31 Dec. 2009. Article 55.

GEWEKE, J. **Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments**. Mineapolis: University of Minnesota, 1991. 31 p. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.7672&rep=rep1&type=pdf>. Acesso em: 29 nov. 2016.

GIANOLA, D. Theory and analysis of threshold characters. **Journal of Animal Science**, Savoy, v. 54, n. 5, p. 1079–1096, 1982.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, London, v. 10, n. 6, p. 381–391, June 2009.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, London, v. 12, 23 May 2011. Article 186.

MAGNABOSCO, C. U.; LÔBO, R. B.; REYES, A.; MARTINS, E. N.; FAMULA, T. Inferência bayesiana na estimação de parâmetros genéticos para peso aos 205 dias de idade em bovinos da raça Nelore. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 34., 1997, Juiz de Fora. **Anais**... Juiz de Fora: SBZ, 1997. v. 3. p. 165-167.

MATUKUMALLI, L. K.; LAWLEY, C. T.; SCHNABEL, R. D.; TAYLOR, J. F.; ALLAN, M. F.; HEATON, M. P.; O'CONNELL, J.; MOORE, S. S.; SMITH, T. P. L.; SONSTEGARD, T. S.; VAN TASSELL , C. P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, San Francisco, v. 4, n. 4, e5350, Apr. 2009.

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

MISZTAL, I.; TSURUTA, S.; STRABEL, T.; AUVRAY, B.; DRUET, T.; LEE, D. H. BLUPF90 and related programs (BGF90). In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 7., 2002, Montpellier. **Proceedings**... Montpellier: INRA, 2002. v. 28. p. 21-22. (Communication n. 28–27).

OLIVEIRA P. S. N.; CESAR, S. M.; NASCIMENTO, M. L.; CHAVES, A. M.; TIZIOTO, P. C.; TULLIO, R. R.; LANNA, D. P. D.; ROSA, A. N.; SONSTEGARD, T. S.; MOURAO, G. B.; REECY, J. M.; GARRICK, D. J.; MUDADU, M. A.; COUTINHO, L. L.; REGITANO, L. C. A. Identification of genomic regions associated with feed efficiency in Nelore cattle. **BMC Genetics**, London, v. 15, 10 Sept. 2014. Article 100.

ONTERU, S. K.; GORBACH, D. M.; YOUNG, J. M.; GARRICK, D. J.; DEKKERS, J. C. M. Whole genome association studies of residual feed intake and related traits in the pig. **PLoS One**, San Francisco, v. 8, n. 6, p. e61756, Jun. 2013.

PETERS, S. O.; KIZILKAYA, K.; GARRICK, D. J.; FERNANDO, R. L.; REECY, J. M.; WEABER, R. L.; SILVER, G. A.; THOMAS, M. G. Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. **Journal of Animal Science**, Savoy, v. 90, n. 10, p. 3398–3409, Oct. 2012.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; BAKKER, P. I. W. de; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **American Journal of Human Genetics**, Chicago, v. 81, n. 3, p. 559–575, Sept. 2007.

R FOUNDATION FOR STATISTICAL COMPUTING. **R**: a language and environment for statistical computing. Vienna, Austria, [2015]. Disponível em: <http://www.R-project.org/>. Acesso em: 25 nov. 2013.

RESENDE, M. F. R.; MUNOZ, P.; ACOSTA, J. J.; PETER, G. F.; DAVIS, J. M.; GRATTAPAGLIA, D.; RESENDE, M. D. V.; KIRST, M. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytologist**, New Jersey, v.193, n. 3, p. 617–624, Feb. 2012.

SCHURINK, A.; WOLC, A.; DUCRO, B. J.; FRANKENA, K.; GARRICK, D. J.; DEKKERS, J. C. M.; VAN ARENDONK, J. A. M. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. **Genetics Selection Evolution**, London, v. 44, 30 Oct. 2012. Article 31.

SHIN, J.-H.; BLAY, S.; MCNENEY, B.; GRAHAM, J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. **Journal of Statistical Software**, Innsbruck, v. 16, p. 1–10, Oct. 2006.

SMITH, B. J. Boa: an R package for MCMC output convergence assessment and posterior inference. **Journal of Statistical Software**, Los Angeles, v. 21, n. 11, p. 1–37, Oct. 2007.

SOLLERO, P. S.; JUNQUEIRA, V.S.; GOMES, C. C. G.; CAETANO, A. R.; CARDOSO, F.F. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. **Genetics Selection Evolution**, v. 49, 15 June 2017. Article 49.

SORENSEN, D. A.; WANG, C. S.; JENSEN, J.; GIANOLA, D. Bayesian analysis of genetic change due to selection using Gibbs sampling. **Genetics Selection Evolution**, Paris, v. 26, n. 4, p. 333-360, 1994.

STEPHENS, M.; BALDING, D. J. Bayesian statistical methods for genetic association studies. **Nature Reviews Genetics**, London, v. 10, n. 10, p. 681–690, Oct. 2009.

SUN, X.; HABIER, D.; FERNANDO, R. L.; GARRICK, D. J.; DEKKERS, J. C. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods. **BMC Proceedings**, London, v. 5, S13–S13, May 2011. Supplement 3.

VAN TASSELL, C. P.; VAN VLECK, L. D. Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. **Journal of Animal Science**, Savoy, v. 74, n. 11, p. 2586–2597, Nov. 1996.

VAN TASSELL, C. P.; VAN VLECK, L. D.; GREGORY, K. E. Bayesian analysis of twinning and ovulation rates using a multiple-trait threshold model and Gibbs sampling. **Journal of Animal Science**, Savoy, v. 76, n. 8, p. 2048–2061, Aug. 1998.

WANG, C. S.; RUTLEDGE, J. J.; GIANOLA, D. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. **Genetics Selection Evolution**, Paris, v. 25, n. 1, p. 41-62, 1993.

WOLC, A.; ARANGO, J.; SETTAR, P.; FULTON, J. E.; O'SULLIVAN, N. P.; PREISINGER, R.; HABIER, D.; FERNANDO, R.; GARRICK, D. J.; HILL, W. G.; DEKKERS, J. C. M. Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. **Animal Genetics**, Hoboken, v. 43, p. 87–96, July 2012. Supplement 1.

Anexo

Linhas de comandos no programa R para seleção e marcadores a partir dos resultados do GenSel:

```
#Importando arquivos
mkrs <- read.csv("snp_content.csv", sep=";", header=T)
#informações dos marcadores (t.like, #model freq., window, posição,
#cromossomos etc.)
ld <- read.csv("LD_topwindows.csv", sep=";", header=T)
#matriz de LD entre os marcadores #identificados nas colnames e
#rownames.


# Valor máximo de Model Freq em cada janela (window)
TopMF_W<-tapply(mkrs$ModelFreq,mkrs$Window.x,max)
summary(TopMF_W)

# Limiares:
# Menor valor entre os valores TopMF_W
minTopMF_W<-min(TopMF_W)
MFt<-minTopMF_W
#Valor de LD acima do qual um dos SNPs é excluído
ldt<-.4
```

```
#########################################################
    #          FUNÇAO PARA SELEÇÃO DE MARCAS            #
#########################################################
subset_mkrs<-function(MFt,ldt) {

# Número total de marcadores acima do limiar de model frequency (MFt)
print(paste("N markers ModelFreq>=MFt:",sum(mkrs$ModelFreq>=MFt)))


# Distribuição t.like dos marcadores com valores de model freq> MFt
summary(mkrs$t.like[mkrs$ModelFreq>=MFt])

# Número total de marcadores acima do limiar de t.like além
daqueles > MFt
t.liket<-min(mkrs$t.like[mkrs$ModelFreq>=MFt])
print(paste("N markers t.like >= min:" , sum(mkrs$t.like>=t.liket)))

# Combinando os dois critérios de seleção
sum(mkrs$ModelFreq>=MFt | mkrs$t.like>=t.liket)
topmkrs1<-mkrs$idx[mkrs$ModelFreq>=MFt | mkrs$t.like>=t.liket]
print(paste("N markers before LD:",length(topmkrs1)))

# Avaliando LD
ld_topmkrs1<-ld[rownames(ld) %in% topmkrs1,colnames(ld) %in% topmkrs1]
idx <- which(ld_topmkrs1>ldt, arr.in=TRUE)
ld.idx<-ld_topmkrs1[idx[,1],idx[,2]]
ld.mkrs<-data.frame(rownames(ld.idx),colnames(ld.idx),
         diag(as.matrix(ld.idx)))
colnames(ld.mkrs)<-c("mkr.x","mkr.y","ld")
cols<-c("idx","Window.x","BTA_chr","BTA_chr",
        "MAF","ModelFreq","t.like")
ld.mkrs<-merge(ld.mkrs,mkrs[,cols],by.x="mkr.x",by.y="idx")
ld.mkrs<-merge(ld.mkrs,mkrs[,cols],by.x="mkr.y",by.y="idx")
ld.mkrs


#Condições para avaliar exclusão
fav.x<-(ld.mkrs$ModelFreq.x>ld.mkrs$ModelFreq.y)+
       (ld.mkrs$t.like.x>ld.mkrs$t.like.y)+
       (ld.mkrs$MAF.x>ld.mkrs$MAF.y)

#Remoção de SNPs
ld.mkrs$remove<-as.character(ld.mkrs$mkr.y)
idx<-!(fav.x>=2)
ld.mkrs$remove[idx]<-as.character(ld.mkrs$mkr.x[idx])
ld.mkrs
```

```
#Painel de SNPs selecionados
keepmkrs<-topmkrs1[!topmkrs1 %in% ld.mkrs$remove]
print(paste("Final N markers after LD:",length(keepmkrs)))
return(list(ld.mkrs,keepmkrs))
}

#Retornando o número de marcadores identificados/selecionados
set<-subset_mkrs(min(tapply(mkrs$ModelFreq,mkrs$Window.x,max)),.4)
#[1] "N markers ModelFreq>=MFt: …"
#[1] "N markers t.like >= min: …"
#[1] "N markers before LD: …"
#[1] "Final N markers after LD: …"
```

Embrapa

Pecuária Sul

CGPE 13287

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

BRASIL
GOVERNO FEDERAL