ROSIANA RODRIGUES ALVES

SELEÇÃO POR TORNEIOS NAS ESTIMATIVAS DE ASSOCIAÇÃO ENTRE MARCADORES SNP'S E FENÓTIPOS

LAVRAS - MG
2014

ROSIANA RODRIGUES ALVES

SELEÇÃO POR TORNEIOS NAS ESTIMATIVAS DE ASSOCIAÇÃO ENTRE MARCADORES SNP'S E FENÓTIPOS

Tese apresentada à Universidade Federal de La-vras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador

Dr. Prof. Júlio Sílvio de Sousa Bueno Filho

LAVRAS - MG
2014

Ficha Catalográfica Elaborada pela Coordenadoria de Produtos e Serviços da Biblioteca Universitária da UFLA

Alves, Rosiana Rodrigues.
Seleção por torneios nas estimativas de associação entre marcadores SNP's e fenótipos / Rosiana Rodrigues Alves. – Lavras : UFLA, 2014.
70 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2014.
Orientador: Julio Silvio Sousa Bueno Filho.
Bibliografia.

1. SNP. 2. Seleção por torneio. 3. Lasso bayesiano. 4. Regressão linear múltipla. I. Universidade Federal de Lavras. II. Título.

CDD – 519.542

ROSIANA RODRIGUES ALVES

SELEÇÃO POR TORNEIOS NAS ESTIMATIVAS DE ASSOCIAÇÃO ENTRE MARCADORES SNP'S E FENÓTIPOS

Tese apresentada à Universidade Federal de La-vras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 27 de fevereiro de 2014.

Dr. Marcio Balestre	UFLA
Dr. Daniel Furtado Ferreira	UFLA
Dra. Maria Imaculada de Sousa Silva	UFU
Dra. Roseli Aparecida Leandro	ESALQ-USP

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS - MG
2014

A Deus,
à minha família,
aos meus irmãos,
aos meus pais,
aos meus avós,
aos meus amigos,
ofereço.

Aos meus Pais, Reginaldo e Aparecida; aos meus irmãos, e a minha afilhada Melissa.
dedico.

AGRADECIMENTOS

A Deus, por sempre iluminar e direcionar meus caminhos

Ao Júlio, pela sua amizade, atenção e dedicação durante todos os anos de orientação.

À Universidade Federal de Lavras, por toda minha formação acadêmica.

A Capes, pela concessão da bolsa.

Aos professores do Departamento de Ciências Exatas, pelos ensinamentos. Aos funcionários do Departamento de Ciências Exatas pela simpatia e boa vontade no atendimento.

A minha mãe e meu pai pelo amor e principalmente por me darem a base para as minhas conquistas. Aos meus avós, meus irmãos, Ronaldo, Rosana e Rosália, e toda minha família pela eterna amizade carinho e presença em todos momentos da minha vida, minha eterna gratidão.

Às meninas da Republica Bananinha, por dividirem comigo todos os momentos da minha vida universitária e por fazerem parte dessa grande família que nós construímos. Aos amigos do DEX, em especial Deyse, Flávia, Fran, por terem sido verdadeiras irmãs, e ao Walmes, pelos ensinamentos constantes e principalmente pela amizade. Aos irmãos (de orientação) Fábio, Manoel, Danilo, Camila pelo apoio e amizade. E Diógenes e Andrezza não existem palavras para agradecer tudo que vocês fizeram para mim e para conclusão do meu doutorado.

À Embrapa Pesca e Aquicultura por permitir que o doutorado fosse concluído.

A todos meus amigos minha eterna gratidão pela oportunidade de aprender, dividir conhecimentos, experiências e farras. À banca avaliadora, pela disponibilidade e auxílio.

A todos vocês meus sinceros agradecimentos pela contribuição para realização de um sonho.

RESUMO

Uma grande dificuldade em analisar dados de seleção genômica é que o número de preditores (marcadores SNPs) é muito maior que o número de animais avaliados. O número de correlações espúrias que surgem por mero acaso entre segregações de marcadores e fenótipos cresce exponencialmente. Uma nova série de abordagens têm sido propostas para solucionar o problema ($n << p$). O lasso bayesiano é uma opção estabelecida na literatura e métodos de seleção por torneios são sugestões recentes. Estes procedimentos consistem em dividir os SNP's em grupos aleatórios e fazer um "torneio" entre os efeitos estimados. Cada grupo é analisado em separado com algum modelo de regressão. Em nosso caso, eliminava-se o SNP como menor efeito. Os marcadores selecionados são então reunidos e entram na próxima fase em que são divididas em grupos por sorteio. Este processo é continuado até o numero de variáveis seja reduzido ao desejado. Neste trabalho, em um estudo de simulação ajustou-se torneios usando a regressão múltipla e o lasso bayesiano. As análises foram comparadas ao lasso bayesiano sem a utilização de torneios (com todas as marcas). As metodologias propostas foram aplicadas em um conjunto de dados de 384 bovinos da raça Canchim genotipados usando o BovineHD BeadChip com 708.641 SNP's identificados. Note-se que no estudo simulação foi utilizada a matriz de genótipos real e simulando os efeitos genéticos do SNP's e o vetor de fenótipos considerando três herdabilidades (25%, 50% e 100%). Para cada herdabilidade analisou-se torneios com três tamanhos de grupos (25, 50 e 100). A validação cruzada foi feita usando 1/8 das observações. Os resultados encontrados no torneios e no lasso bayesiano mostram que esses métodos não são muito diferentes em simulação. Para os dados reais a validação cruzada também foi equivalente. É preciso notar que os torneios permitem a paralelização direta da análise. Com o equipamento usado, torneios com regressão múltipla foram 10 vezes mais rápidos que o laso bayesiano. Os torneios são metodologias simples e rápidas com eficiência equiparável ao lasso bayesiano.

Palavras-chave: SNP. Regresão linear múltipla. Lasso Bayesiano. Seleção por torneios.

ABSTRACT

A major difficulty in analyzing genomic selection data is that the number of predictor (SNP markers) is much larger than the number of evaluated animals. The spurious correlations that arises by chance in joint segregation of SNPs and phenotypes grows exponentialy. A new series of methods have been proposed to tackle this question ($n << p$). Bayesian Lasso and its variations are established in the literature and tournament screening is among the new suggestions. The idea is to divide SNPs in randomly assembled groups and make a "tournament" of estimated effects. Each group is analysed separately with a regression model. In our case, SNPs with saller effects were out. Remaining markers are pooled and new phase of random groups were generated. The process goes on until the remaining SNPs are as few as desired. In this work, in a simulation study, multiple regression and Bayesian lasso models were adjusted within groups. Analyses were then compared to Bayesian lasso with all markers (no tournaments). Proposed methods were applied to 384 bovine (Canchim breed animals) genotyped using the BovineHD Bead-Chip with $708{,}641$ SNPs identified. Was also performed a simulation study using the real genotypes matrix and, simulating the genetical effects of SNPs and the vector of phenotypes, these, considering the following heritability estimates: 25%, 50% and 100%. For each heritability were analysed tournaments containing three sizes of groups: 25, 50 and 100. An 8-fold cross-validation was carried out. According to results from simulation, the tournaments and Bayesian lasso do not differ much. For real data, cross validation results are also equivalent. Note that tournaments allow for direct parallelization of analyses. With the used hardware tournaments with multiple regression were 10-fold faster than Bayesian lasso. Tournaments are simple and fast methods that yield equivalent results to Bayesian lasso.

Keywords: SNPs. Bayesian Lasso. Tournment Screening. Multiple Linear Regression.

SUMÁRIO

1	**INTRODUÇÃO**	8
2	**REFERENCIAL TÉORICO**	11
2.1	**SNP's e seleção assistida por marcadores moleculares**	11
2.2	**Lasso**	14
2.3	**Seleção por torneios**	16
3	**MATERIAL E MÉTODOS**	18
3.1	**Obtenção dos genotípos e fenótipos**	18
3.2	**Seleção de marcadores utilizando torneios**	20
3.2.1	**Situações consideradas no estudo de simulação**	22
3.2.2	**Avaliação da predição de valores genéticos**	22
3.3	**Análises com os fenótipos reais (AOL)**	24
4	**RESULTADOS E DISCUSSÃO**	25
4.1	**Predição de valores genéticos**	25
4.2	**Seleção de marcadores**	33
4.3	**Resultados das análises com fenótipos reais**	35
4.4	**Discussão**	37
5	**CONCLUSÕES**	40
	REFERÊNCIAS	41
	APÊNDICE	43

1 INTRODUÇÃO

O desenvolvimento de tecnologias para o sequenciamento do genoma possibilitou a identificação de polimorfismos nucleotídeo único (SNP – Single Nucleotide Polimorphism) distribuídos ao longo do genoma. Em seguida, a tecnologia evoluiu para a confecção de chips que permitem a genotipagem simultânea para milhares de SNP, disponibilizando um volume nunca visto de informações para um único animal. Nesse contexto, a seleção assistida por marcadores em escala ampla passa a ser denominada seleção genômica, na qual o valor genético dos candidatos à seleção é estimado usando apenas a informação molecular.

A seleção genômica ampla (GWS), proposta por Meuwissen, Hayes e Goddard (2001) consiste na utilização simultânea de centenas ou milhares de marcadores, os quais cobrem o genoma de maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Por atuar em todo o genoma é considerada ampla, capturando todos os genes que afetam um caráter quantitativo sem a necessidade de identificar previamente os marcadores com efeitos significativos (sem ouso de testes de significância para marcas individuais) .

Nesses conjuntos de dados com grande número de marcadores, é de se supor que apenas uma pequena parte deles contribuam para a variação genética e influencie a resposta fenotípica observada, havendo, potencialmente centenas de milhares de marcadores irrelevantes e redundantes em estudos de alta densidade de marcadores. Nesses tipos de conjunto de dados o número de marcadores, p geralmente é muito que o tamanho da amostra n. Em razão do preço da genotipagem por animais ou plantas ser maior que a obtenção de novos marcadores.

Um modelo de regressão linear múltipla pode ser usado para descrever a relação entre os fenótipos e marcadores. Os efeitos de todos os marcadores são estimados a partir do modelo e, em seguida, com base nessas estimativas, podem-se fazer testes de hipóteses e predições. Testes de hipóteses são utilizados para identificar SNPs com efeito significativo em nível populacional, ou seja, SNPs em desequilíbrio de ligação com QTLs. A partir de estimativas de efeitos de SNPs, podem-se predizer valores fenotípicos de novos indivíduos (que não foram utiliza-

dos na obtenção do modelo) candidatos à seleção em programas de melhoramento.

Uma abordagem, para seleção de variáveis, quando o número de variáveis independentes (p) é muito maior que número de observações (n), é utilizar um método de Encolhimento Bayesiano (Bayesian Shrinkage) que inclui todas variáveis no modelo e usa uma distribuição à priori informativa para reduzir os efeitos triviais a zero. O Lasso (Least Absolute Shrinkage and Selection Operator) é um método de encolhimento amplamente utilizado em análise de regressão para modelos de grandes dimensões (TIBSHIRANI, 1996). Park e Casella (2008) propuseram uma formulação bayesiana para o método LASSO de Tibshrinani (1996) denominado Lasso Bayesiano, em que o estimador Lasso é interpretado como uma estimativa da moda a posteriori em um contexto bayesiano, através da distribuição à priori Laplace (Exponencial Dupla).

Assim, as abordagens convencionais que têm sido abordadas na literatura têm se mostrado menos eficientes do que a mera estimação do parentesco entre os animais com o uso das marcas, para uso em modelos mistos posteriores. Por este motivo, uma nova série de abordagens têm sido propostas para solucionar o problema das seleção de variáveis quando a dimensão das colunas da matriz de genótipos é alta.

O que estas abordagens têm em comum é procurar primeiro reduzir a dimensão das colunas das matrizes de genótipos. Em seguida, um ajuste mais simples ou mais sofisticado é utilizado para encontrar os marcadores causais. Nesse contexto que Chen e Chen (2009) propuseram a seleção por torneios. Esse procedimento consiste em dividir os SNP's em grupos aleatórios. Cada grupo é analisado com um modelo de verossimilhança penalizada e um determinado número de variáveis são selecionadas. Os marcadores selecionados são então reunidas e entram na próxima fase. As variáveis retidas na fase anterior são divididas em grupos que não se sobrepõem. Este processo é continuado até o numero de variáveis seja reduzido a um nível desejável.

A proposta deste trabalho é apresentar e testar outros procedimentos de seleção de marcadores por torneios baseado nas ideias Chen e Chen (2009) para facilitar a seleção de marcadores. Durante cada etapa do torneio os grupos serão submetidos à regressão linear múltipla ou Lasso Bayesiano. E, ao final do torneio

os SNP's selecionados serão submetidos Lasso Bayesiano para estimar seus efeitos genéticos aditivos.

2 REFERENCIAL TÉORICO

2.1 SNP's e seleção assistida por marcadores moleculares

Marcadores moleculares são polimorfismos de DNA (VIGNAL et al., 2002). Eles podem ter diversas aplicações como, por exemplo, estudos de genética populacional,na realização de testes de paternidade, aplicações de cunho forense, estudos de associação com um dado fenótipo e estudos de mapeamento .

Avanços tecnológicos recentes trouxeram metodologias de alto desempenho e acurácia, e baixo custo e mão de obra para prospecção, caracterização e genotipagem de marcadores SNP (Single Nucleotide Polymorphism). Os marcadores SNP têm como base as alterações mais elementares da molécula de DNA, ou seja, mutações em bases únicas da cadeia de bases nitrogenadas (CAETANO, 2009).

Quando essas mutações ocorrem dentro de um gene, podem alterar a formação da respectiva proteína e, por consequência, gerar variabilidade fenotípica em duas características qualitativas e/ou quantitativas (complexas) interesse para a produção animal. Uma das grandes vantagens do marcador do tipo SNP é que possui baixa taxa de mutação, estimada em 10^{-9} mudanças de nucleotídeos por geração, entretanto, pode haver um confundimento na distinção entre mutação e SNP (VIGNAL et al., 2002). Todos os SNPs surgem a partir de uma mutação, sendo assim toda e qualquer conversão de um nucleotídeo a outro é considerado um evento mutacional. Porém, caso essa mutação não seja prejudicial para o organismo e seja passada ao longo das gerações para os organismos daquela população, a mutação passa a ser vista como polimorfismo de um único nucleotídeo (SNP) .

Normalmente, os marcadores SNP são bialélicos, portanto com baixo conteúdo de polimorfismo por loco, os SNPs oferecem algumas vantagens como: baixa taxa de mutação,codominância e abundância. Essas tecnologias trouxeram novas soluções para aplicações já solidificadas e, também, estão permitindo o desenvolvimento de novas aplicações (CAETANO, 2009).

Marcadores SNPs em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci – QTL*), podem ser utilizados para predizer

o valor genético genômico (*Genomic Breeding Value – GBV*) de cada indivíduo, e servir para a seleção, aumentando a acurácia na avaliação genética. A aplicação prática destas informações é um desafio, pois, geralmente não é possível a utilização adequada de métodos tradicionais, baseados em quadrados mínimos (Least Squares – LS), para estimar o efeito de cada SNP no fenótipo, uma vez que, geralmente, o número de marcadores é muito maior que o número de animais genotipados (SILVA et al., 2013).

O domínio da técnica de sequenciamento total do genoma e o desenvolvimento da tecnologia dos SNP chips promoveram um aumento exponencial do número de marcadores disponíveis, bem como o interesse pela MAS aplicada em ampla escala, denominada seleção genômica.

O desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular criou-se a expectativa de que informações genotípicas quando correlacionadas com características fenotípicas possam ser utilizadas na identificação e seleção de indivíduos com maiores valores genéticos. Com essa finalidade metodologias foram desenvolvidas para utilização de marcadores moleculares na identificação de locos que controlam características de interesse ao melhoramento, por exemplo, a seleção assistida por marcadores (Marker Assisted Selection - MAS) a genética de associação (Genome Wide Association Studies - GWAS) e a seleção genômica ampla (Genome Wide Selection - GWS). A idéia básica da MAS é explorar a dependência estatística (desequilíbrio de ligação) existente na distribuição conjunta dos marcadores e das regiões cromossômicas associadas às características quantitativas permitindo aprimorar as predições do mérito genético dos candidatos a seleção

A primeira metodologia proposta para estimar os efeitos dos marcadores foi pela aplicação do método dos mínimos quadrados. Fernando e Grossman (1989) estimaram o efeito de um locus pela teoria de modelos mistos. Whittaker, Thompson e Denham (2000) sugeriram que a metodologia de regressão de cumeeira fosse utilizada na estimação dos efeitos dos marcadores para superar o problema de dimensionalidade dos dados e de colinearidade entre efeitos (REZENDE, 2013). Meuwissen, Hayes e Goddard (2001) apresentaram a GWS que realiza a predição simultânea dos efeitos dos marcadores, sem o uso de testes de significân-

cia para marcas individuais, em situações que o número de marcadores pode ser muito maior que o número de indivíduos. Com base nos trabalhos pioneiros, nos últimos anos foram apresentados diversos métodos bayesianos de encolhimento e s métodos não paramétricos para seleção utilizando marcadores, por exemplo, De Los Campos et al. (2009), Gianola et al. (2009), Gianola, Fernando e Stella (2006) e Yi e Xu (2008).

A inovação de Meuwissen, Hayes e Goddard (2001) não foi em termos de metodologia estatística mas, em termos conceituais enfatizando o uso do conceito de desequilíbrio de ligação em nível populacional e não apenas dentro de família e o não uso de testes de significância para marcas. O maior mérito do trabalho foi a demonstração, via simulação, do fato de que a GWS pode realmente funcionar na prática.

A GWS enfatiza a predição simultânea (sem o uso de testes de significância para marcas individuais) dos efeitos genéticos de milhares de marcadores genéticos de DNA (SNP, DArT, microssatélites) dispersos em todo o genoma de um organismo, de forma a capturar os efeitos de todos os locos (tanto de pequenos quanto de grandes efeitos) e explicar toda a variação genética de um caráter quantitativo. A condição fundamental para isso é que haja desequilíbrio de ligação, em nível populacional, entre alelos dos marcadores e alelos dos genes que controlam o caráter. A predição dos efeitos genéticos é realizada com base em dados genotípicos e fenotípicos de indivíduos pertencentes a uma amostra da população de seleção.

O não uso de significância estatística para a seleção de marcas pela GWS a distingue da GWAS, a qual procura associação entre locos e caráter fenotípico em nível populacional, por meio de testes de hipóteses visando detectar efeitos com significância estatística. A GWAS sofre com a alta taxa de falsos negativos em razão do uso de pontos de corte muito rigorosos visando evitar a ocorrência de falsos positivos. A GWS equivale à GWAS aplicada sobre todos os locos simultaneamente e baseando-se em estimação e predição em vez de teste de hipótese. Dessa forma consegue explicar parte muito maior da variabilidade genética e evitar a chamada herdabilidade faltante ou perdida (missing heritability), típica dos estudos de análise de ligação e de associação.

2.2 Lasso

O Lasso (Least Absolute Shrinkage and Selection Operator) é um método de encolhimento, usado em análise de regressão, para modelos de grandes dimensões (TIBSHIRANI, 1996). O Lassso é, normalmente, usado para estimar os parâmetros de regressão $\beta = (\beta_1,\beta_2, ..., \beta_p)^T$ no modelo

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i = \mu + \mathbf{X_i}\beta + e_i \tag{1}$$

em que y_i é o valor fenótipico do i-ésimo indivíduo ($i = 1,2, ..., n$); μ é parâmetro comum a todas as observações x_{ij} denota o genótipo do marcador $j(j = 1, 2, ..., p)$ do indivíduo i; o coeficiente β_j representa o efeito do marcador j; e_i é o erro residual. Assume-se que $e_i \sim N(0, \sigma^2)$, $\mathbf{X_i} = (x_{i1},x_{i2}, \cdots ,x_{ip})$ e $\beta = (\beta_1, \beta_2, \cdots , \beta_p)^T$. O Lasso minimiza a soma de quadrados residuais, restringindo a soma dos valores absolutos dos coeficientes de regressão, $t \geq \sum |\beta_j|$ para $t \geq 0$. Essa restrição permite que algumas estimativas dos coeficientes de regressão sejam exatamente zero, realizando simultaneamente um procedimento *shrinkage* e seleção de modelos. A estimativa do Lasso é obtida por :

$$\min_{\beta,\lambda} \left[\left(y - \sum_{j=1}^{p} x_i\beta_j \right)^T \left(y - \sum_{j=1}^{p} x_i\beta_j \right) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2}$$

em que $\lambda \geq 0$ é multiplicador de Lagrange, que se relaciona implicitamente ao limite t e controla o grau de encolhimento.

A estimativa do Lasso pode ser interpretada como uma estimativa da moda a posteriori quando os parâmetros de regressão parâmetros de regressão têm prioris Laplace independentes (Exponencial Dupla) (PARK; CASELLA, 2008). Baseado nessa conexão alguns autores propuseram a utilização da exponencial dupla como priori. Park e Casella (2008) utilizam a priori Exponencial Dupla em um modelo hierárquico. O modelo hierárquico, quando comparado com a forma original da exponencial dupla, é mais facilmente tratável tanto do ponto de vista analí-

tico quanto computacional. Para obter as estimativas dos parâmetros de regressão $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ do modelo 1, é utilizado o amostador de Gibbs . As distribuições a priori do Lasso Bayesiano serão apresentadas como em Sun, Ibrahim e Zou (2010)

$$\begin{aligned}
\pi(\mu) &\propto 1, \\
\pi(\sigma^2) &\propto \frac{1}{\sigma^2}, \\
\pi(\beta_j|\tau_j^2) &\sim N(0,\tau_j^2) = \frac{1}{\sqrt{2\pi\tau_j^2}} \exp(-\frac{\beta_j^2}{2\tau_j^2}), \\
\pi(\tau_j^2|\frac{\lambda^2}{2} &\sim Exp\frac{\lambda^2}{2} = \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right), \\
\pi\left(\frac{\lambda^2}{2}\right) &\sim Gama(s,r) = \frac{r^2}{\Gamma(s)} \left(\frac{\lambda^2}{2}\right)^{s^{-1}} \exp\left(-r\frac{\lambda^2}{2}\right),
\end{aligned}$$

em que $j = 1, \cdots, p$ indicam os p marcadores. Também, é modelada a distribuição de $\frac{\lambda^2}{2}$ em vez de λ, e na priori de $\frac{\lambda^2}{2}$, s e r são parâmetros de forma e escala, respectivamente. Podem ser consideradas, ainda, duas opções de prioris para $\pi(\beta|\tau_j^2)$, a saber $\pi(\beta_j|\tau_j^2) \sim N(0,\tau_j^2)$ (YI; XU, 2008) e $\pi(\beta_j|\tau_j^2) \sim N(0,\sigma^2\tau_j^2)$ (PARK; CASELLA, 2008)

A distribuição a posteriori conjunta de todos os parâmetros $(\mu, \beta, \sigma^2, \tau^2, \lambda|y)$ pode ser expressa por:

$$\pi(\mu, \beta, \sigma^2, \tau^2, \lambda|y) \propto \prod_{i=1}^{n} \pi(y_i|\mu, \beta, \sigma^2)\pi(\mu)pi(\sigma^2) \prod_{j=1}^{p} \pi(\beta_j|\tau_j^2)\pi(\tau_j^2|\lambda)\pi(\frac{\lambda^2}{2}) \tag{3}$$

em que $\pi|\mu, X, \beta, \sigma^2) \sim N(\mu + X_i\beta, \sigma^2)$.

Sun, Ibrahim e Zou (2010) apresentam as distribuições condicionais completas a posteriori, utilizadas para obtenção de uma amostra da distribuição conjunta a posteriori pelo amostrador de Gibbs. Com base nessa amostra, são obtidas estimativas dos parâmetros de regressão $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$. Geralmente são utilizadas médias a posteriori para estimar β, cujas estimativas assumem valores

próximos de zero mas não exatamente zero.

2.3 Seleção por torneios

A seleção genômica é um típico problema ($p >> n$) o número de marcadores é muito maior que o número de indivíduos da amostra. Tipicamente há até centenas de milhares de marcadores (p) e apenas algumas centenas de indivíduos fenotipados e genotipados (n). Desta forma, os métodos de seleção de modelos convencionais são inaplicáveis (CHEN; CHEN, 2009).

Uma alternativa a esse problema foi apresentada por Chen e Chen (2009), chamada seleção por torneios em que conjunto de marcadores é dividido em grupos aleatórios e em cada grupo é realizada uma análise com um modelo de verossimilhança penalizada e um determinado número de marcadores são selecionados. O processo é repetido com os marcadores selecionados até que o número de marcadores seja reduzido a um nível desejado.

Se denotarmos por S^1 um conjunto de inteiros de 1 a p. Seja S^0 o subconjunto de S^1 que corresponde aos componentes não nulos de β (coeficientes das variáveis independentes), isto é, $\beta \neq 0$ se e somente se $j \in S^0$. $\nu(S)$ é o número de elementos dos subconjuntos S de S^1. Em particular, $\nu_0 = \nu(S_0)$. Desta forma, $\boldsymbol{X}(S)$ é a matriz correspondente às colunas do subconjunto $\boldsymbol{X}$, com índice de coluna em S, e associados ao efeito descrito por $\boldsymbol{\beta}(S)$. A função de log-verossimilhança penalizada negativa é dada por

$$l_p(\boldsymbol{\beta}(S), \sigma^2|\lambda) = -2\ln f(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta}(S) + n\sum_{j\in S} p_\lambda(|\beta_j|)),$$

em que $p_\lambda(.)$ é a função de penalidade regulada pelo parâmetro de ajuste λ. A função de penalidade pode ser tomada como $p_\lambda(\boldsymbol{\beta_j}|) = \lambda|\boldsymbol{\beta_j}|$, a penalidade L_1 usada no Lasso por Tibshirani (1996) ou outra penalidade.

A propriedade crucial necessária da função penalidade é que deve ser singular em zero. Em função desta singularidade, quando a verossimilhança penalizada negativa é minimizada com λ, fixado num determinado valor, um número de

componentes de β serão estimados como zero.

Chen e Chen (2009) demonstraram que a seleção por torneios proposta pelos mesmos satisfaz uma propriedade denominada *sure screening property*, que é a capacidade de selecionar as variáveis causais e eliminar as variáveis não causais (FAN; LV, 2008). Por analogia, isto seria o equivalente à seleção de grupos de variáveis em direção *backward* em modelos de regressão com $n > p$.

Seja n_g o número de elementos de cada grupo tal que $n_g < n$ e, seja k um número pré determinado de variáveis a serem selecionadas em cada grupo. A princípio k deve ser grande o suficiente para reter todas as variáveis causais dentro do grupo e ao mesmo tempo pequeno o suficiente para reduzir o número de variáveis eficientemente. Caso se tenha alguma ideia do tamanho de ν_0 (número total de variáveis causais), k pode ser escolhido com $2\nu_0$ ou $3\nu_0$.A seguir será apresentado detalhadamente a seleção por torneio:

Estágio 1 Divide-se S^1 em grupos aleatórios de tamanho de modo que

$$S^1 \quad = \quad S_{11_1} \cup \cdots \cup S_{J1_1}$$

em que J_1 é o maior inteiro tal que $[n_g J_1] \leqslant p$. Para cada grupo de S_{1J}, minimiza-se $l_p(\beta(S), \sigma^2|\lambda)$ ajustando λ para que sejam estimados apenas k componentes $\beta(S_{1j})$ não nulos. As variáveis correspondentes aos k coeficientes não nulos deste grupo são selecionadas, formando o grupo S^*_{1j} das k variáveis selecionadas. Por fim, são reunidos todos os grupos $S^*_{1j}, j = 1, 2, ..., J_1$ para formar o grupo de todas as variáveis selecionadas no primeiro estágio S^2. O número de variáveis é reduzido para k_{J1} neste estágio.

Estágio 2 O processo do Estagio 1 é repetido com S^1 substituído por S^2.

Demais estágios O processo acima descrito continua até que a dimensão do espaço de características seja reduzido a k.

Depois de reduzir o número original de variáveis $p(\gg n)$ para $k(< n)$, então qualquer método convencional de seleção de variáveis pode ser aplicado na seleção do modelo final. Chen e Chen (2009) utilizam para a seleção do modelo final um procedimento que combina a metodologia de verossimilhança penalizada com um critério de informação bayesiano extendido (*Extended Bayes Information Criterion - EBIC*).

3 MATERIAL E MÉTODOS

3.1 Obtenção dos genotípos e fenótipos

Nesse estudo são utilizados dados reais de bovinos provenientes de animais registrados na Associação Brasileira de Criadores de Canchim de sete rebanhos localizados em dois estados brasileiros (São Paulo e Goiás), cedidos pela EMBRAPA. Foram genotipados 400 animais com chips BovineHD BeadChip (Illumina Inc., San Diego, CA).

Após o processamento de controle de qualidade foram selecionados 384 animais e 708.641 SNP's. Foram eliminados os SNP's que apresentavam informação faltante em algum animal, ficando assim 561831 SNP's. Para diminuir a multicolinearidade foram eliminados também o SNP's que era exatamente igual ao SNP vizinho restando 526493 SNP's.

Os genótipos utilizados no estudo são então dados pela matriz de SNPs limpa como descrita acima. Os fenótipos foram obtidos por dois mecanismos de simulação, e além disso, foram analisados fenótipos reais. É muito importante notar que esta metodologia (simular efeitos a serem aplicados à matriz real de genótipos) permite uma generalização posterior dos resultados, pois em certa medida estarão preservados elementos importantes da populaçào original, em especial, o desequilíbrio de ligaçào entre as marcas deve ser apreciavelmente rpeservado.

A característica fenotípica real utilizada para as análises foi a área de olho do lombo (AOL) que é um indicador da composição da carcaça, relacionada à musculosidade do animal e ao rendimento dos cortes de alto valor comercial.

O estudo de simulação foi realizado com objetivo de verificar se a metodologia de Torneios proposta é capaz de selecionar SNPs corretamente ou SNPs próximos destes. Para o estudo de simulação, foram gerados vetores de efeitos genéticos aditivos de SNPs (β) contendo 48 SNPs com efeitos não nulos e o restante com efeitos nulos. Foram simulados dois vetores de efeitos de SNPs considerando as situações:

a) SNPs com efeitos não nulos, próximos entre si no cromossomo, distribuídos em 4 cromossomos (Figura 1);

b) SNPs com efeitos não nulos, um pouco mais dispersos no cromossomo, distri-buídos em 8 cromossomos (Figura 2).

Figura 1 Efeitos simulados dos SNPs, em valores absolutos, próximos entre si no cromossomo, distribuídos em 4 cromossomos

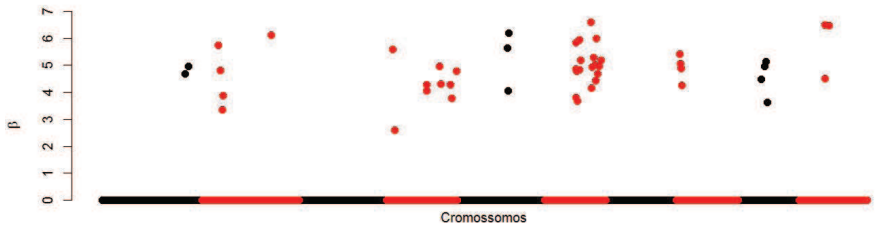

Figura 2 Efeitos simulados dos SNPs, em valores absolutos, um pouco mais dispersos no cromossomo, distribuídos em 8 cromossomos

Ao longo do cromossomo alguns SNP's próximos com efeito diferente de zero foram simulados para caracterizar regiões de QTL. Essa configuração foi

feita para verificar se os métodos selecionavam SNP's próximos a esses supostos QTL's.

Os SNPs com efeito diferente de zero serão chamados aqui como SNPs causais, o que é razoável no estudo de simulação. A simulação foi feita com um conjunto de dados reduzido, foram utilizados 10% dos SNPs do conjunto original de dados, os quais foram provenientes dos cromossomos 20 ao 29, selecionados a cada dez SNPs igualmente espaçados, de forma que o conjunto de dados reduzido foi composto por 11.812 SNPs. A redução foi feita devido ter sido realizado 100 repetições de cada análise, reduzindo assim o tempo e custo computacional das análises.

Com a matriz de genótipos reais reduzida ($\mathbf{X}$) e efeitos simulados dos SNPs (β) foi possível calcular os valores genéticos genômicos dos animais, GBV = $\mathbf{X}\beta$. De posse dos GBVs foram simulados vetores de fenótipos dos animais, considerando diferentes herdabilidades (0,25, 0,5 e 1,0), de acordo com o modelo 4. Assim, os fenótipos foram simulados como:

$$y = 1\mu + X\beta + \varepsilon \tag{4}$$

em que $\mu = 100$ e ε é o vetor de efeitos residuais, gerados segundo uma distribuição normal com média zero e variância σ^2 compatível com a herdabilidade desejada, $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$. A herdabilidade $h2 = 1.0$ foi considerada para avaliar a metodologia na ausência de erro.

3.2 Seleção de marcadores utilizando torneios

A seleção de marcadores foi realizada utilizando uma estratégia semelhante a seleção por torneios apresentados por Chen e Chen (2009). Entretanto neste trabalho será eliminado apenas um marcador por grupo em cada etapa do torneio. E para selecionar esse marcador a ser eliminado ao invés de utilizar metodologia de verossimilhança penalizada foi utilizado a regressão linear múltipla clássica ou o LASSO bayesiano.

O primeiro passo para a realização do torneio é a divisão do número total de marcadores S^1 em grupos aleatórios menores $S_{11} \cup S_{12} \cup \cdots S_{1J_1}$, cada

grupo com tamanho $n_g < p$. A seguir são apresentadas as etapas que compõem a metodologia de seleção de marcadores por torneios proposta neste trabalho.

Etapa 1 É feita a divisão do número total de marcadores S^1 em grupos menores $S_{11} \cup S_{12} \cup \cdots \cup S_{1J_1}$, cada grupo com tamanho $n_g < p$. Em cada grupo S_{1j} foi ajustado um modelo utilizando uma das seguintes metodologias:

- Regressão linear múltipla clássica;
- LASSO Bayesiano

Obtidas as estimativas dos efeitos dos SNPs, $\beta(S_{1j})$, foi eliminado um marcador de acordo com os seguintes critérios:

- Utilizando regressão linear múltipla é eliminado o marcador cuja estimativa tem o menor p-valor ou, no caso de singularidade perfeita, é eliminado (aleatoriamente) um dos marcadores cujo efeito não pode ser estimado;
- Utilizando o LASSO Bayesiano é eliminado o marcador com menor valor, em módulo, da estimativa do efeito dividida pelo respectivo desvio padrão;

Os marcadores não eliminados de todos os grupos são agrupados novamente em um único grupo S^2.

Etapa 2 Repita o procedimento da Etapa 1 substituindo S^1 por S^2

Demais etapas Repita o procedimento acima até que o número total de marcadores seja reduzido para o nível desejado.

Ao final do torneio é ajustado um modelo, utilizando LASSO Bayesiano, aos SNPs selecionados pelo torneio. Considerando que foram simulados 48 SNPs causais e, que apenas um SNP de cada grupo é eliminado em cada etapa do torneio, serão realizados torneios considerando diferentes tamanhos de grupos proporcionais ao número de SNPs causais simulados. Assim, foram avaliados torneios com os seguintes tamanhos de grupo (n_g): 25, 50 e 100.

A sequência de eliminação dos marcadores durante as etapas do torneio é armazenada gerando um *ranking* dos marcadores, de forma que os marcadores mais importantes são os que restaram ao final do torneio seguidos pelos marcadores eliminados, na seguinte ordem: dos últimos marcadores a serem eliminados no torneio até os primeiros.

Para reduzir o tempo de análise pela metodologia de torneios foi implementada uma programação paralela que permite que as análises em cada grupo sejam executadas simultaneamente, uma em cada core do computador, utilizando a biblioteca multicore do software R (R CORE TEAM, 2014).

Para avaliar a metodologia de torneios quanto à capacidade de selecionar marcadores para predição de valores genéticos genômicos foi realizada uma comparação com a capacidade preditiva do LASSO Bayesiano sem torneios. Para ajustar o LASSO Bayesiano foi utilizada a biblioteca BLR do software R (R CORE TEAM, 2014).

3.2.1 Situações consideradas no estudo de simulação

Resumindo as situações consideradas no estudo de simulação e avaliação da metodologia de torneios:

- SNPs causais agrupados;
- SNPs causais dispersos;
- Diferentes herdabilidades: 0,25; 0,5 e 1,0;
- Diferentes tamanho dos grupos: 25, 50 e 100;
- Diferentes metodologia utilizadas na etapa de seleção dos marcadores do torneio: Regressão linear múltipla (LM) e LASSO Bayesiano (BL);

3.2.2 Avaliação da predição de valores genéticos

Para avaliação da capacidade de predição dos valores genéticos genômicos dos indivíduos foram calculadas as correlações entre os GBVs estimados e os fenótipos simulados ($r_{X\beta,y}$) bem como as correlações entre os GBVs estimados e simulados ($r_{X\hat{\beta},X\beta}$) nas populações de estimação e de validação considerando um esquema de validação cruzada. Para os fenótipos reais a validação cruzada foi realizada dividindo-se o conjunto de dados dos 384 animais em 8 grupos de 48 animais. Em cada etapa da validação cruzada um grupo foi retirado da análise para

ser utilizado na validação e o grupo formado pelos 7 grupos restantes foi utilizado para estimação do modelo.

O modelo obtido na população de estimação foi então aplicado na população de validação para predizer os GBVs dos indivíduos desta população. Foram então calculadas as correlações $r_{X\beta,y}$ e $r_{X\hat{\beta},X\beta}$ na população de validação. O processo foi repetido para cada um dos 8 grupos e os resultados finais foram as correlações médias das 8 análises.

Para comparar as capacidades de predição das diferentes metodologias foram utilizadas as correlações $r_{X\beta,X\beta}$ e $r_{X\beta,y}$ avaliadas em diferentes números de marcadores utilizados para o ajuste do modelo. Isto foi feito da seguinte forma:

Torneios Para as metodologias de torneios inicialmente foi realizado um torneio sem validação cruzada para selecionar 100 SNPs ao final do torneio e, consequentemente obter o *ranking* dos demais SNPs conforme a ordem de eliminação no torneio. De posse do *ranking* foram ajustados modelos com o LASSO Bayesiano, em esquema de validação cruzada, para diferentes números de SNPs selecionados (100, 250, 500, 1000, 2000, 4000, 8000, 11812).

LASSO Bayesiano sem torneios Inicialmente foi ajustado um modelo pelo LASSO Bayesiano a todos os SNPs (modelo completo) sem validação cruzada e foi obtido o *ranking* dos SNPs de acordo com os módulos das estimativas de seus efeitos obtidas neste modelo. De posse do *ranking* foram ajustados modelos com o LASSO Bayesiano, em esquema de validação cruzada, para diferentes números de SNPs selecionados (100, 250, 500, 1000, 2000, 4000, 8000, 11812).

Dessa forma, tanto as análises com torneios quanto o LASSO Bayesiano sem torneios foram utilizados para fornecer o *ranking* dos SNPs mais importantes e em ambos os casos foram ajustados modelos utilizando LASSO Bayesiano para diferentes números de SNPs (obedecendo o ranking) em esquema de validação cruzada. Portanto o que diferenciou as análises com torneios das análises utilizando o LASSO Bayesiano sem torneios foi a forma de obter o *ranking* dos SNPs, ou equivalentemente, a forma de selecionar os SNPs para compor o modelo.

3.3 Análises com os fenótipos reais (AOL)

Foram realizadas análises com os fenótipos reais de área de olho de lombo corrigidos para efeitos de grupo de contemporâneos utilizando as metodologias de torneios com regressão linear múltipla com grupos aleatórios. Nos torneios com dados de AOL foi utilizado o tamanho de grupo que obteve melhor desempenho nos estudos de simulação. Estas análises foram comparadas à do LASSO Bayesiano aplicado diretamente (sem torneios). Para comparação do desempenho das metodologias utilizadas na análise dos dados de AOL foram consideradas as correlações entre os GBVs estimados e os fenótipos, ou seja, $r_{X\beta,y}$. Também foram feitos gráficos dos módulos das estimativas dos efeitos dos 100 SNPs de maior importância, de acordo com cada metodologia, em suas posições nos cromossomos, para visualização dos SNPs selecionados.

4 Resultados e Discussão

Nas primeiras duas subseções serão apresentados resultados da predição e da escolha de marcadores no exemplo simulado. Na terceira subseção serão apresentados resultados para o exemplo com o fenótipo real (AOL). Na última subseção serão discutidos em conjunto os resultados de simulação e do exemplo real.

4.1 Predição de valores genéticos

Nas figuras 3 a 6 são apresentadas as correlações genéticas aentre os GBV estimados e os valores fenotípicos simulados, na média das simulações. Tais figuras facilitam a avaliação da influência do tamanho de grupos nas diferentes metodologias de torneio, considerando os dois cenários simulados.

Nas figuras 3 e 4 são apresentados os gráficos das correlações médias entre os GBVs estimados e fenótipos simulados em populações de validação para diferentes tamanhos de grupos. Foram consideradas siuações com números de marcadores selecionados variando de 100 a 11812, para o torneio realizado utilizando a regressão linear múltipla (LM). Quando os SNPs causais estão agrupados, o tamanho de grupo 25 apresenta as maiores correlações para a herdabilidade é baixa e número de marcadores selecionados no modelo final é menor. À medida que o número de marcadores aumenta no modelo final a correlação tende a diminuir. No entanto, quando os SNPs causais estão dispersos o padrão de correlação muda, como podemos ver na figura 4. Aqui, para a herdabilidade $0{,}25$ o aumento da correlação ocorre até o modelo final apresentar mais de 2000 marcadores e só então a correlação diminui. Em geral as correlações nas situações em que os SNPs causais estão agrupados são maiores do que quando estão espalhados.

Nas figuras 5 e 6 são apresentadas as correlações médias entre os GBVs estimados e fenótipos simulados em populações de validação nas situações em que torneios são realizados utilizando o Lasso bayesiano (BL) para seleção de marcadores estão apresentados. Da mesma forma que aconteceu com os modelos de regressão, os torneios com BL apresentam correlações menores quando os efeitos

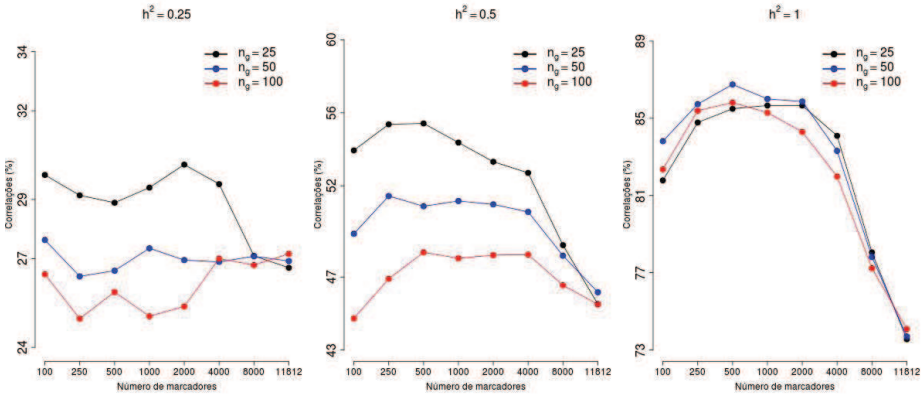

Figura 3 Correlações (%) entre GBVs estimados e fenótipos ($r_{X\hat{\beta},y}$) no grupo de validação, para diferentes números de SNPs selecionados em torneios utilizando regressão linear múltipla, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais agrupados

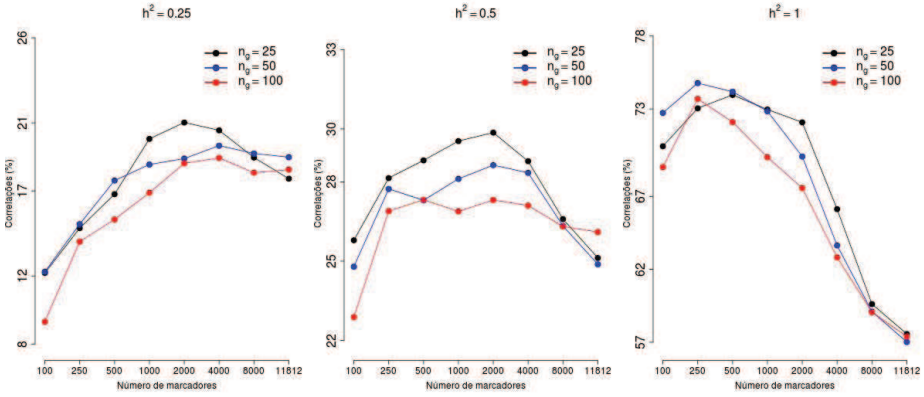

Figura 4 Correlações (%) entre GBVs estimados e fenótipos ($r_{X\hat{\beta},y}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando regressão linear múltipla, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais dispersos

paramétricos dos SNPs estão dispersos do que agrupados.

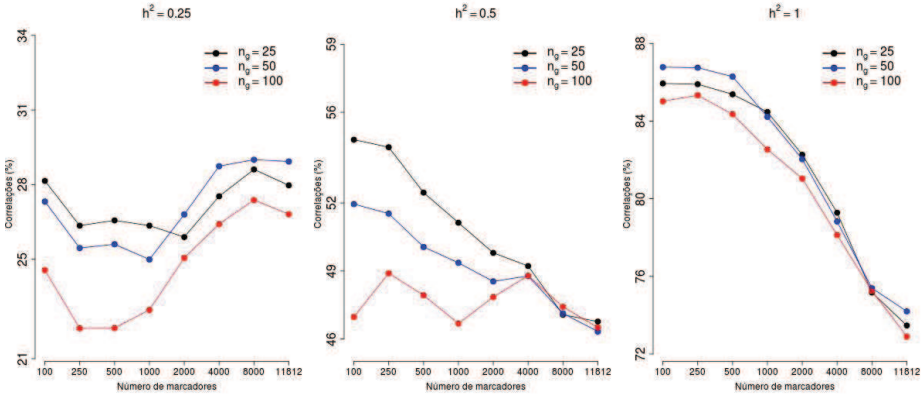

Figura 5 Correlações (%)entre GBVs estimados e fenótipos ($r_{X\hat{\beta},y}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando BL na etapa de seleção de marcadores, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais agrupados

Nas figuras 7 à 10 são apresentados gráficos das correlações médias entre os GBVs estimados e simulados em populações de validação para diferentes tamanhos de grupos, em situações de diferentes números de marcadores selecionados e para as diferentes metodologias de torneios testadas.

A capacidade de predição das metodologias de seleção de marcadores utilizando torneio com tamanho de grupo de 25 foram comparadas a seleção utilizando apenas Lasso Bayesiano. Apenas esse tamanho de grupo foi escolhido devido ao seu melhor desempenho.

Nas figuras 11 a 12 são apresentados gráficos das correlações entre GBVs estimados e os fenótipos ($r_{X\hat{\beta},y}$) simulados em populações de validação em torneios com LM, torneios utilizando o BL e o BL aplicado a todas as marcas, sem utilização de torneios. Foram consideradas também diferentes herdabilidades e diferentes números de SNPs selecionados com SNPs causais agrupados. Na figura 11 as correlações $r_{X\hat{\beta},y}$ se mantém constantes a medida que varia o número de marcadores selecionados independentemente do método de seleção. Comparando

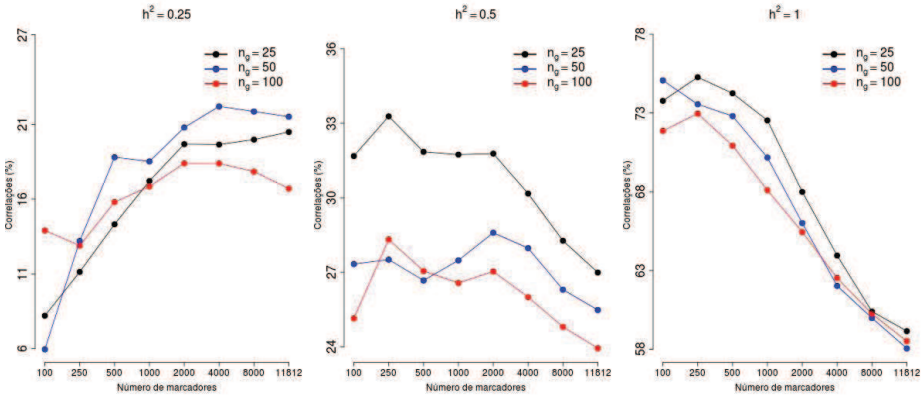

Figura 6 Correlações (%)entre GBVs estimados e fenótipos ($r_{X\hat{\beta},y}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando LASSO Bayesiano na etapa de seleção de marcadores, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais dispersos

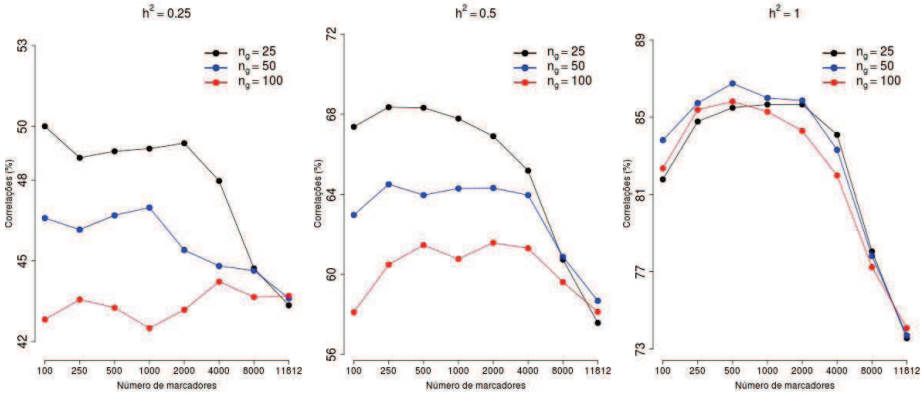

Figura 7 Correlações (em porcentagem) entre GBVs estimados e fenótipos ($r_{X\hat{\beta},X\beta}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando regressão linear múltipla, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais agrupados

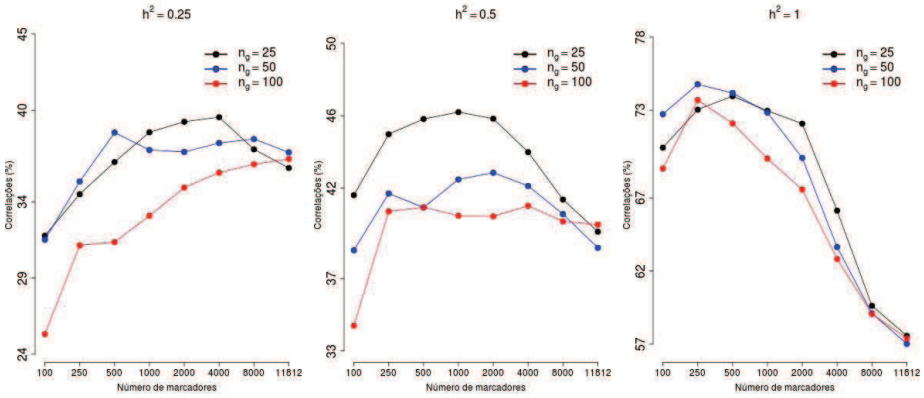

Figura 8 Correlações (%)entre GBVs estimados e simulados ($r_{X\hat{\beta},X\beta}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando regressão linear múltipla, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais dispersos

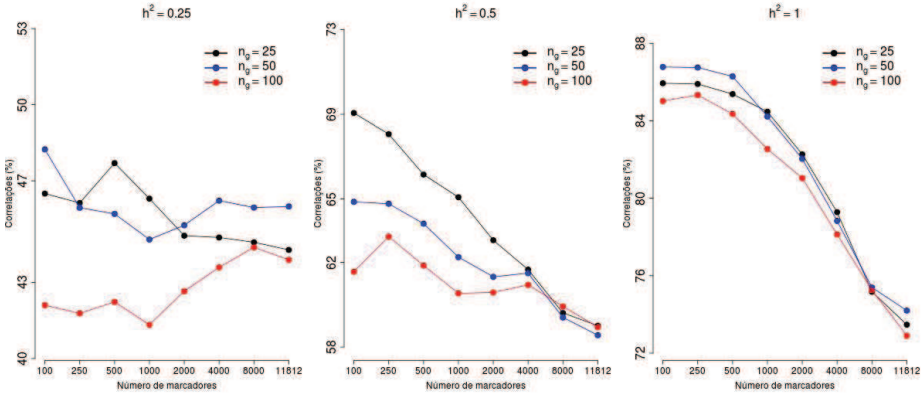

Figura 9 Correlações (%)entre GBVs estimados e simulados ($r_{X\hat{\beta},X\beta}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando LASSO Bayesiano na etapa de seleção de marcadores, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais agrupados

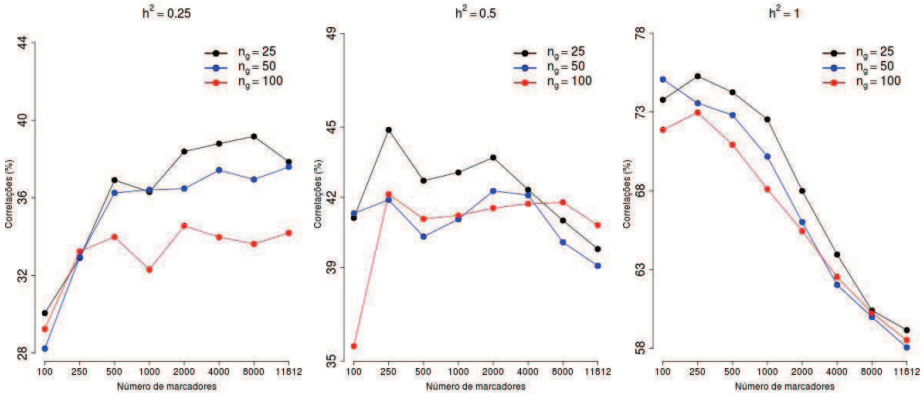

Figura 10 Correlações (%)entre GBVs estimados e simulados ($r_{X\hat{\beta},X\beta}$) no grupo de validação, para diferentes números de SNPs selecionados em Torneios utilizando LASSO Bayesiano na etapa de seleção de marcadores, considerando diferentes tamanhos de grupos e herdabilidades, no cenário de SNPs causais dispersos

os métodos eles apresentam correlações bem próximas. E uma outra informação interessante que pode ser observada nas tabelas 2 a 4 do Apêndice A é que os desvios padrão dos coeficientes de correlação na validação cruzada para os torneios com LM são bem menores que os dos BL, com ou sem torneios. Quando os SNPs causais simulados estão dispersos (figura 12) em herdabilidades baixas os métodos apresentam as correlações muito próximas, mas com um comportamento um pouco diferente ao apresentado na figura 11. Os SNPs causais simulados dispersos fez com que as correlações fossem mais baixas em geral, e ainda é necessário um grupo maior de marcadores no modelo final para que se tenha a melhor predição.

Nas figuras 13 e 14 são apresentados gráficos das correlações entre GBVs estimados e simulados em populações de validação em torneios com LM, torneios utilizando o BL e o BL sem utilização de torneios. Podemos observar que os métodos de seleção de marcadores apresentados na figura 13 são muitos semelhantes quando comparados através da correlação $r_{X\hat{\beta},X\beta}$. Essas correlações tendem a diminuir quando aumentam o número de marcadores no modelo final. Quando os SNPs causais estão dispersos (figura 14) as correlações entre os GBVs estima-

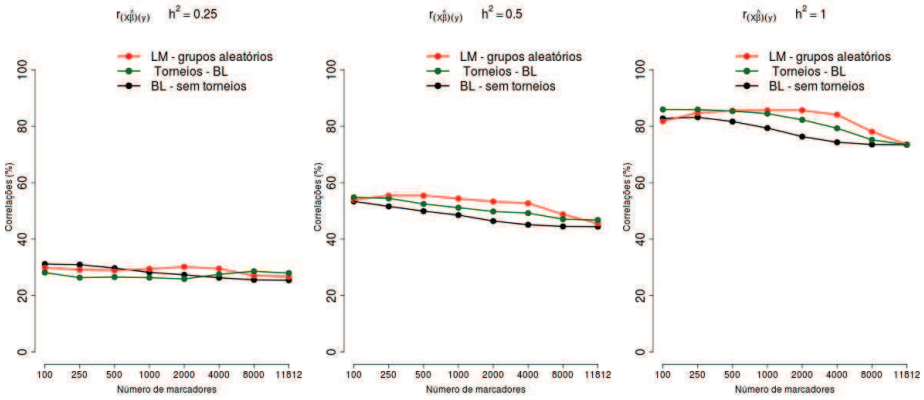

Figura 11 Correlações $r_{X\hat{\beta},y}$ na população de validação, para diferentes números de SNPs selecionados por meio de (i) Torneios utilizando LM, (ii) Torneios utilizando Lasso Bayesiano na etapa de seleção de marcadores e, (iii) Lasso Bayesiano sem torneios. Herdabilidades 0,25; 0,5 e 1, no cenário de SNPs causais agrupados

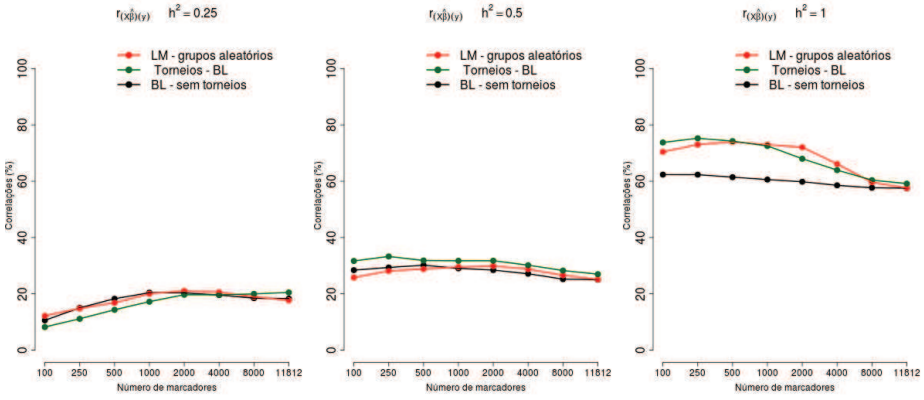

Figura 12 Correlações $r_{X\hat{\beta},y}$ na população de validação, para diferentes números de SNPs selecionados por meio de (i) Torneios LM, (ii) Torneios BL na etapa de seleção de marcadores e, (iii) Lasso Bayesiano sem torneios (maiores estimativas). Herdabilidades 0,25; 0,5 e 1, no cenário de SNPs causais dispersos

dos e simulados tem um pequeno aumento até numero de marcadores no modelo final esteja entre 1000 e 2000 nas herdabilidades 0,25 e 0,50 e depois elas diminuem. Nessas herdabilidades as correlações dos três métodos apresentados são muitos próximas. Contudo na herdabilidade 1,0 os torneios apresentam correlações ($r_{X\hat{\beta},X\beta}$) superiores ao Lasso Bayesiano até que se tenha 4000 marcadores no modelo final, a partir desse número o métodos são iguais.

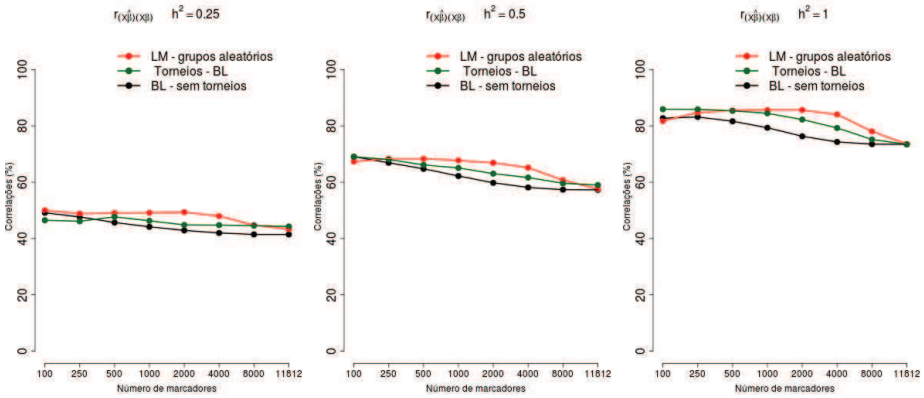

Figura 13 Correlações $r_{X\hat{\beta},X\beta}$ na população de validação, para diferentes números de SNPs selecionados por meio de (i) Torneios LM com grupos, (ii) Torneios BL na etapa de seleção de marcadores e, (iii) Lasso Bayesiano sem torneios (maiores estimativas). Herdabilidades 0,25; 0,5 e 1, no cenário de SNPs causais agrupados

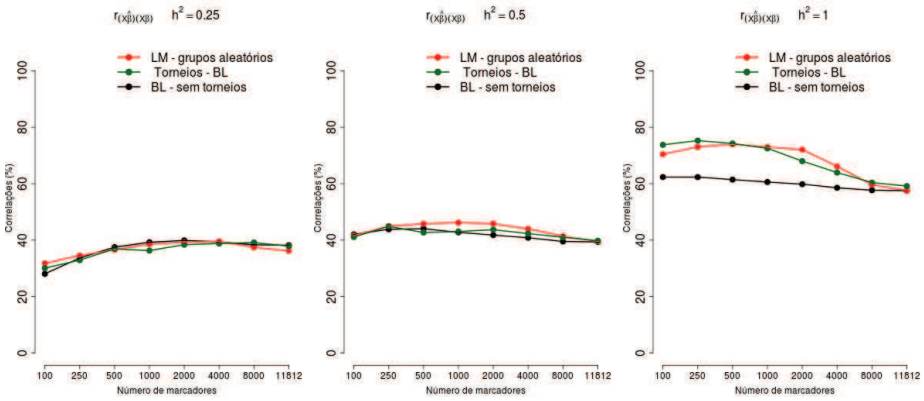

Figura 14 Correlações $r_{X\hat{\beta},X\beta}$ na população de validação, para diferentes números de SNPs selecionados por meio de (i) Torneios LM, (ii) Torneios BL e, (iii) Lasso Bayesiano sem torneios. Herdabilidades 0,25; 0,5 e 1, no cenário de SNPs causais dispersos

Pode-se observar que as metodologias utilizando torneios apresentaram correlações próximas ao BL. Na situação de SNPs causais estão dispersos é necessário mais marcadores no modelo de estimação para que se tenha melhor predição. E ainda podemos observar que quando os SNPs causais simulados são espalhados a predição em todos os métodos de seleção são piores do que na situações em que os SNPs causais estão próximos.

4.2 Seleção de marcadores

Para avaliar a metodologia quanto à capacidade de selecionar SNPs corretamente, ou seja, selecionar os SNPs causais simulados (ou próximos destes), serão utilizados gráficos que apresentam os módulos das estimativas dos efeitos dos SNPs selecionados "$\times$", juntamente com os módulos dos efeitos simulados "$\bullet$", em suas respectivas localizações no genoma.

Os gráficos apresentam os módulos das estimativas dos efeitos dos 100 SNPs que permaneceram ao final do torneios utilizando o Lasso Bayesiano. Para critério de comparação também foram feitos gráficos dos 100 SNPs com maiores

módulos das estimativas obtidas pelo Lasso Bayesiano, sem utilização de torneios.

Na Figura 15 é apresentado o gráfico das estimativas de 100 SNPs selecionados pela metodologia de Torneios com Regressão linear, com grupos formados por 25 SNPs, considerando uma herdabilidade de 0.25 no cenário de SNPs causais agrupados. Pode-se observar que foram selecionados SNPs nas proximidades dos SNPs causais simulados e distantes também. Inclusive em cromossomos onde não haviam SNPs causais. Este fato se repetiu em todas as metodologias e, principalmente para herdabilidade de 0.25. Nas figuras 15 e 16 são apresentados o mesmo tipo de gráfico porém para SNPs selecionados por torneios com Regressão linear e para SNPs de maiores módulos das estimativas no Lasso Bayesiano sem torneios, respectivamente.

Os demais gráficos considerando outras herdabilidades e situações de SNPs causais agrupados e dispersos são apresentados nas figuras 20 a 30 (APÊNDICES B e C). Aparentemente todas as metodologias conseguem selecionar SNPs próximos aos SNPs causais, no entanto todas as metodologias estão selecionando muitos SNPs que estão distantes dos SNPs causais. Um possível motivo para as metodologias estarem selecionando muitos SNPs incorretamente é que ao final dos torneios foram selecionados mais SNPs do que a quantidade de SNPs causais simulados.

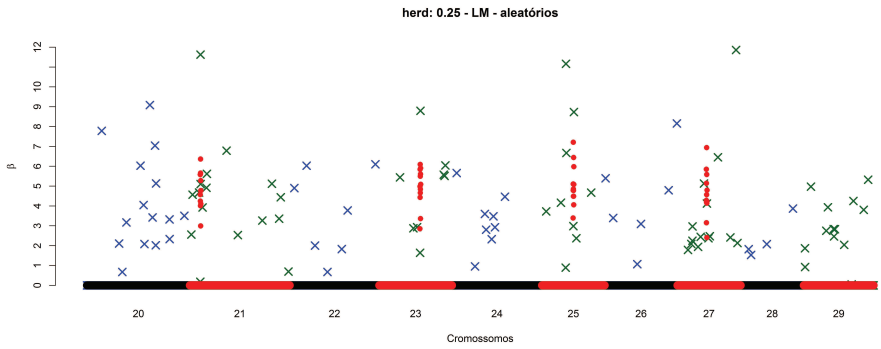

Figura 15 Estimativas dos efeitos dos SNPs obtidas pelo Torneio com regressão linear múltipla para uma herdabilidade de 0.25, com os SNPs causais agrupados

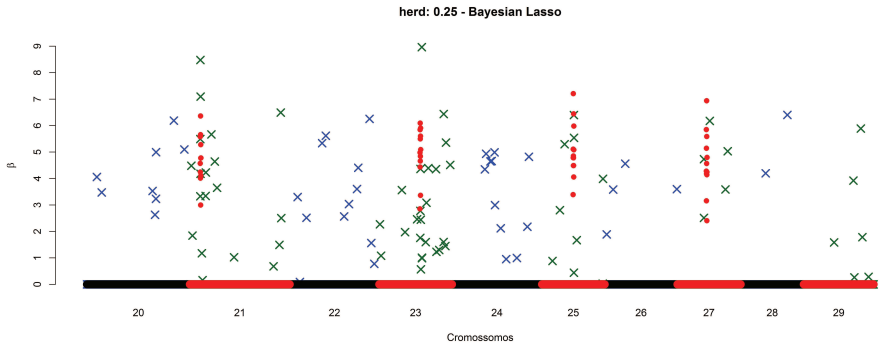

Figura 16 Estimativas dos efeitos dos SNPs obtidas pelo Lasso Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas (em módulo) no modelo completo, para uma herdabilidade de 0.25, no cenário de SNPs causais agrupados

4.3 Resultados das análises com fenótipos reais

Na figura 17 são apresentados os resultados das análises com os dados de área de olho de lombo (AOL). Na análise com os fenótipos reais de AOL utilizou-se o banco de dados completo dos genótipos dos SNPs, ao contrário das análises com fenótipos simulados em que foi utilizado um conjunto de dados de genótipos reduzido. As metodologias LM e o Lasso Bayesiano sem torneios apresentaram correlações semelhantes, sendo que LM apresentou maior correlação do que o Lasso Bayesiano quando se utilizou 100 marcadores. O tempo de análise de cada metodologia para se obter o ranking dos marcadores pode ser observado na tabela 1. Quanto ao tempo, fica claro na análise dos dados reais que os torneios com LM são muito mais rápidos que o Lasso Bayesiano sem torneios.

Quanto à seleção de SNPs pode-se observar nas figura 18 e 19 as estimativas e posições nos cromossomos dos 100 SNPs selecionados por Torneios LM e os 100 SNPs de maiores módulos das estimativas no Lasso Bayesiano sem torneios, respectivamente.

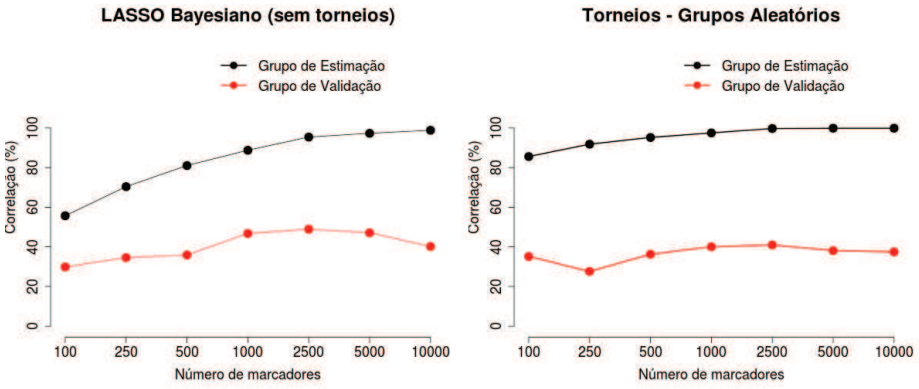

Figura 17 Correlações em populações de estimação e de validação nas análises com (i) Lasso Bayesiano sem torneios, (ii) Torneios LM para diferentes números de marcadores selecionados

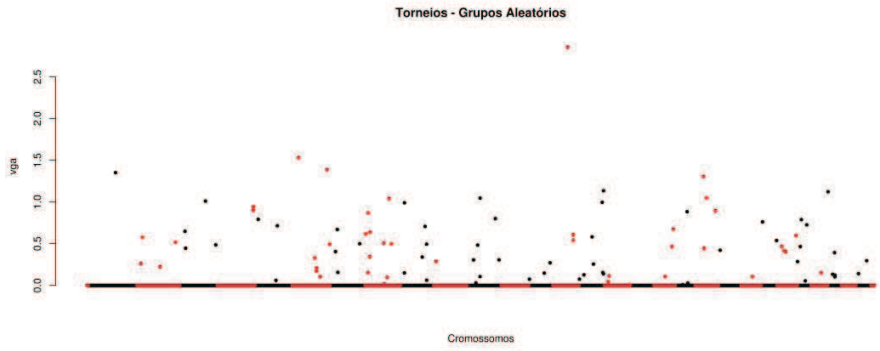

Figura 18 Estimativas dos efeitos de 100 SNPs selecionados por Torneios utilizando regressão linear múltipla para os fenótipos reais de área de olho de lombo

Tabela 1 Tempo de execução das análises para obtenção do ranqueamento dos marcadores para os dados de AOL

Metodologia	Tempo de Análise
Torneios utilizando regressão linear Lasso	19 minutos
Bayesiano sem torneios	3 horas e 8 minutos

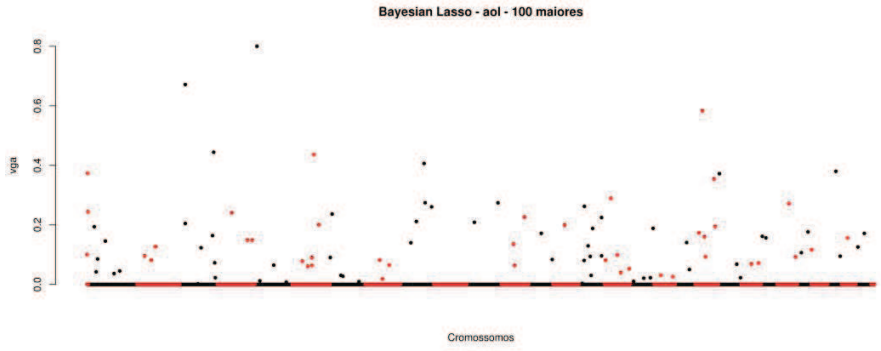

Figura 19 Estimativas dos efeitos de 100 SNPs selecionados pelos maiores módulos das estimativas no modelo completo ajustado com o Lasso Bayesiano sem torneios para os fenótipos reais de área de olho de lombo

4.4 Discussão

O fato de os SNPs estarem agrupados em regiões de QTL facilita a descoberta destas regiões e traz correspondentes melhorias nas estimativas de correlações genéticas. Pode-se dizer que a seleção genômica seria boa alternativa mesmo em casos de genes principais controlando o caracter de interesse. Na simulação houve uma tendência a que menores grupos resultassem em melhor seleção. O tamanho de grupo deveria afetar menos aos torneios BL, mas estes também foram afetados. Parte disto pode ser devida ao pequeno número de SNPs causais simulados, pois $n = 25$ era o único tamanho de grupo menor que o número de SNPs causais simulados. Os torneios BL foram melhores que os LM apenas com herdabilidade 1. Deve-se assumir que do ponto de vista prático os torneios com LM

sejam preferíveis, dado o menor tempo de análise.

As correlações entre GBV estimados e simulados trazem praticamente o mesmo padrão, mas neste caso no cenário de SNPs causais agrupados e herdabilidade $0{,}5$ os torneios com LM seriam preferíveis, enquanto que para a mesma situação e efeitos dispersos o BL supera o LM. Em ambos os casos os menores tamanhos de grupo permitiram chegar próximo aos limites teóricos destas correlações (rais quadradada da herdabilidade). Para a herdabilidade 1 nenhum método e tamanho de grupo permitiu se aproximar do limite teórico. Isto provavelmente se deve ao fato de a montagem dos grupos ter terminado com 100 marcadores selecionados. Desta forma, os demais marcadores (ou segregações incluidas) representam erro.

As análises apresentadas nas figuras 11 a 14, envolvendo populações de validação simuladas, não revelam grandes diferenças entre os métodos. Destaca-se no entanto a consistente melhor performance dos torneios para situações de herdabilidade 1. É claro que isto não indica que os métodos são bons, mas parece ser indício de que são menos afetados por segregações espúrias. Analisando as figuras 15 e 16 referentes a situações plausíveis de em estudos de associação, constatamos que o que importa na seleção genômica é acertar segregações equivalentes às dos SNPs causais e não estes últimos *per se*. Assim, as 100 maiores estimativas de efeitos de SNPs não correspondem aos efeitos simulados, embora a seleção seja razoavelmente bem correlacionada aos valores genéticos. É certo que mais estudos são necessários antes de descartar a identificação de gens, mas a tendência a métodos de seleção genômica serem mais factíveis que o mapeamento por associação são consistentes com a literatura.

Um resultado bastante animador para os torneios LM foi o desempenho equivalente em termos de correlação e validação cruzada quando comparado ao Lasso bayesiano sem torneios. Isto feito com um tempo de processamo=ento aproximadamente dez vezes menor. Note que, no caso do LM a possibilidade de paralelização é ilimitada, indicando que o processamento poderia ser ainda muito mais rápido em um cluster. O mesmo não se pode dizer do Lasso bayesiano, porque a obtenção de cadeias convergentes não escala com a paralelização.

O resultado em termos de estimativas de efeitos não deve ser relevante

para a comparação dos métodos, conforme verificado no estudo de simulação, mas a velocidade dos torneios LM aliada à sua precisão equivalente a uma das metodologias padrão que demanda extremo esforço computacional faz com que seja uma possível recomendação para o uso prático.

5 CONCLUSÕES

Com base nos estudos de simulação não se pode tirar conclusões definitivas a respeito da validade dos métodos na seleção de marcadores associados à característica de interesse.

Quanto à seleção genômica em estudos de simulação com diferentes tamanhos de grupo e herdabilidades, a metodologia de torneios entre marcadores utilizando regressão linear múltipla clássica (LM) foi análoga à de torneios utilizando lasso bayesiano, inclusive superando o LASSO Bayesiano sem torneios, para algumas das situações simuladas.

Uma vantagem da metodologia de torneios LM em relação ao LASSO Bayesiano (com ou sem torneios) é que ela apresenta desempenho computacional muito melhor.

A utilização de torneios entre marcadores LM potencialmente consegue amenizar os efeitos da multicolinearidade aumentando consideravelmente a capacidade de predição de valores genéticos dos indivíduos tanto em simulação quanto na análise de dados reais e deve ser considerada para estudos posteriores.

REFERÊNCIAS

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, n. 8, p. 64–71, 2009.

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

DE LOS CAMPOS, G. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, Bethesda, v. 182, n. 1, p. 375–385, 2009.

FAN, J.; LV, J. Sure independence screening for ultra-high dimensional feature space. **Journal of the Royal Statistical Society: Series B**, Princeton, v. 70, n. 5, p. 849–911, Nov. 2008.

FERNANDO, R. L.; GROSSMAN, M. Marker assisted selection using best linear unbiased prediction. **Genetics Selection Evolution**, London, v. 21, p. 467–477, 1989.

GIANOLA, D. et al. Additive genetic variability and the bayesian alphabet. **Genetics**, Bethesda, v. 183, n. 1, p. 347–363, 2009.

GIANOLA, D.; FERNANDO, R. L.; STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, Bethesda, v. 173, n. 3, p. 1761–1776, 2006.

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

PARK, T.; CASELLA, G. The bayesian lasso. **Journal of the American Statistical Association**, New York, v. 103, n. 482, p. 681–686, 2008.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 10 mar. 2014.

REZENDE, F. M. de. **Incorporação de informações de marcadores genéticos em programas de melhoramento genético de bovinos de corte**. 2013. 88 p. Tese (Doutorado em Zootecnia) - Universidade de São Paulo, São Paulo, 2013.

SILVA, F. et al. Genome wide selection for growth curves. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 65, n. 5, p. 1519–1526, 2013.

SUN, W.; IBRAHIM, J. G.; ZOU, F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. **Genetics**, Austin, v. 185, n. 1, p. 349–359, Jan. 2010.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 58, p. 267–288, Jan. 1996.

VIGNAL, A. et al. A review on snp and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, Paris, v. 34, n. 3, p. 275–306, 2002.

WHITTAKER, J.; THOMPSON, R.; DENHAM, M. Marker-assisted selection using ridge regression. **Genetics Research**, London, v. 75, p. 249–252, 2000.

YI, N.; XU, S. Bayesian LASSO for quantitative trait loci mapping. **Genetics**, Austin, v. 179, n. 2, p. 1045–1055, Jan. 2008.

APÊNDICE

APÊNDICE A - Estatísticas descritivas das correlações das 100 repetições das análises apresentadas

Tabela 2 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com o LASSO Bayesiano (sem torneios) considerando diferentes herdabilidades (h^2), diferentes números de SNPs de maiores módulos dos efeitos estimados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados

Tipo de População	Nº de SNPs utilizados	h^2		
		0,25	0,50	1
Estimação	100	78,81±3,24	82,28±3,07	89,85±1,90
	250	88,02±1,64	89,81±1,38	93,82±0,77
	500	93,02±0,88	93,94±0,78	93,21±0,41
	1000	96,30±0,41	96,77±0,38	98,00±0,21
	2000	98,16±0,24	98,51±0,21	99,13±0,08
	4000	92,12±0,22	99,36±0,16	99,70±0,05
	8000	99,48±0,23	99,66±0,17	99,90±0,06
	11812	99,51±0,23	99,68±0,17	99,91±0,06
Validação	100	31,18±13,53	53,38±11,36	82,75±5,36
	250	30,96±10,60	51,59±10,75	83,18±3,82
	500	29,71±9,89	49,91±10,65	81,66±3,62
	1000	28,23±10,21	48,52±10,77	79,35±4,02
	2000	27,36±10,34	46,40±10,37	76,32±4,42
	4000	26,32±10,88	45,11±10,48	74,31±5,15
	8000	25,56±10,73	44,50±10,42	73,53±5,57
	11812	25,43±10,77	44,42±10,41	73,47±5,62

Tabela 3 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com Torneios utilizando regressão linear múltipla clássica e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	86,82±0,31	89,09±0,25	93,93±0,21
		250	91,79±0,28	93,11±0,22	96,80±0,15
		500	95,13±0,15	95,76±0,15	98,60±0,09
		1000	97,34±0,12	97,70±0,08	99,83±0,03
		2000	98,42±0,14	98,91±0,09	99,97±0,01
		4000	98,78±0,14	99,28±0,12	99,96±0,01
		8000	99,30±0,12	99,59±0,04	99,92±0,01
		11812	99,53±0,07	99,70±0,07	99,91±0,03
	50	100	90,96±0,28	92,29±0,23	95,76±0,22
		250	94,95±0,13	95,42±0,14	97,86±0,09
		500	97,24±0,14	97,37±0,11	99,19±0,06
		1000	98,72±0,19	98,74±0,11	99,94±0,02
		2000	99,35±0,08	99,39±0,11	99,98±0,01
		4000	99,43±0,10	99,55±0,09	99,95±0,04
		8000	99,43±0,08	99,59±0,10	99,92±0,03
		11812	99,56±0,07	99,66±0,06	99,92±0,02
	100	100	91,80±0,28	93,08±0,14	96,16±0,19
		250	96,68±0,12	97,05±0,13	98,43±0,11
		500	98,55±0,05	98,54±0,14	99,47±0,06
		1000	99,53±0,07	99,54±0,06	99,97±0,01
		2000	99,75±0,13	99,85±0,02	99,99±0,00
		4000	99,72±0,03	99,81±0,04	99,96±0,03
		8000	99,60±0,08	99,73±0,05	99,94±0,02
		11812	99,58±0,08	99,70±0,07	99,91±0,02

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	29,83±2,19	53,94±3,48	81,78±1,24
		250	29,14±1,44	55,37±2,85	84,78±1,16
		500	28,89±1,51	55,42±2,03	85,49±0,72
		1000	29,40±1,21	54,37±2,31	85,66±0,51
		2000	30,18±1,22	53,32±1,86	85,66±0,67
		4000	29,52±1,03	52,71±1,68	84,09±0,89
		8000	27,09±1,25	48,75±1,88	78,05±1,30
		11812	26,69±1,17	45,52±1,92	73,55±1,78
	50	100	27,63±1,61	49,37±3,42	83,81±0,88
		250	26,40±2,49	51,44±3,70	85,73±0,92
		500	26,59±2,45	50,88±3,24	86,75±1,12
		1000	27,35±1,54	51,17±3,83	86,00±0,91
		2000	26,95±1,97	50,98±2,79	85,87±0,64
		4000	26,89±2,89	50,57±1,92	83,31±0,95
		8000	27,08±2,45	48,17±1,90	77,81±1,05
		11812	26,92±2,37	46,17±1,70	73,70±1,50
	100	100	26,47±3,19	44,73±4,60	82,35±2,33
		250	24,97±2,59	46,90±3,98	85,39±1,62
		500	25,87±2,91	48,35±3,50	85,81±1,17
		1000	25,05±4,04	48,03±2,64	85,29±0,86
		2000	25,38±3,73	48,20±1,83	84,29±1,07
		4000	27,00±4,23	48,23±1,79	81,98±0,89
		8000	26,78±3,38	46,55±1,80	77,23±1,06
		11812	27,16±3,57	45,50±2,17	74,08±0,96

Tabela 4 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com Torneios utilizando LASSO Bayesiano e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	87,64±0,69	90,39±0,64	95,83±0,42
		250	91,99±0,66	94,06±0,70	97,83±0,27
		500	94,95±0,68	96,20±0,56	99,10±0,19
		1000	97,42±0,64	98,02±0,48	99,87±0,08
		2000	98,79±0,52	99,25±0,29	99,98±0,02
		4000	99,27±0,64	99,60±0,16	99,96±0,03
		8000	99,45±0,24	99,66±0,17	99,92±0,08
		11812	99,58±0,20	99,67±0,17	99,91±0,11
	50	100	91,44±0,75	93,10±0,54	96,90±0,28
		250	95,04±0,72	95,89±0,49	98,39±0,24
		500	97,26±0,48	97,66±0,44	99,42±0,15
		1000	98,68±0,42	98,85±0,37	99,92±0,12
		2000	99,42±0,38	99,60±0,23	99,98±0,02
		4000	99,47±0,36	99,54±0,57	99,97±0,02
		8000	99,49±0,30	99,65±0,21	99,93±0,08
		11812	99,57±0,22	99,70±0,10	99,92±0,04
	100	100	92,78±0,57	94,08±0,42	97,26±0,24
		250	97,37±0,32	97,75±0,22	98,98±0,15
		500	98,75±0,25	98,85±0,25	99,71±0,08
		1000	99,56±0,20	99,62±0,20	99,98±0,02
		2000	99,69±0,22	99,81±0,14	99,99±0,01
		4000	99,64±0,21	99,75±0,15	99,94±0,18
		8000	99,48±0,25	99,72±0,12	99,93±0,04
		11812	99,55±0,19	99,74±0,10	99,90±0,12

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	28,15±12,10	54,79±8,61	85,94±4,55
		250	26,35±11,97	54,46±8,60	85,90±4,19
		500	26,56±12,39	52,46±9,33	85,38±3,50
		1000	26,35±11,47	51,13±9,29	84,47±3,50
		2000	25,89±12,13	49,80±9,05	82,27±4,40
		4000	27,53±11,55	49,22±7,69	79,28±5,06
		8000	28,61±12,11	47,07±9,60	75,17±5,60
		11812	27,97±12,36	46,77±9,15	73,47±6,02
	50	100	27,32±10,52	51,95±10,37	86,78±3,02
		250	25,45±11,64	51,53±10,53	86,75±3,38
		500	25,60±12,24	50,06±10,83	86,29±3,00
		1000	24,99±12,57	49,36±11,12	84,22±4,00
		2000	26,80±13,28	48,54±12,73	82,04±4,57
		4000	28,74±14,96	48,78±12,46	78,83±5,81
		8000	29,00±13,51	47,13±12,41	75,39±7,26
		11812	28,93±13,79	46,33±11,78	74,20±7,32
	100	100	24,56±10,61	46,97±9,25	85,02±3,63
		250	22,22±13,12	48,90±8,60	85,33±3,33
		500	22,23±12,55	47,93±10,05	84,35±3,80
		1000	22,96±12,87	46,68±10,93	82,54±4,11
		2000	25,05±11,39	47,85±10,25	81,04±4,12
		4000	26,41±10,80	48,79±10,20	78,13±5,19
		8000	27,38±11,15	47,42±10,08	75,25±5,58
		11812	26,81±10,32	46,51±10,24	72,90±6,44

Tabela 5 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com o LASSO Bayesiano (sem torneios) considerando diferentes herdabilidades (h^2), diferentes números de SNPs de maiores módulos dos efeitos estimados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Nº de SNPs utilizados	h^2		
		0,25	0,5	1
Estimação	100	77,21±3,67	78,34±3,13	83,97±2,19
	250	87,22±1,77	88,09±1,48	90,88±1,07
	500	92,69±0,88	93,02±0,82	94,70±0,56
	1000	96,09±0,48	96,29±0,36	97,21±0,35
	2000	98,07±0,23	98,11±0,27	98,70±0,18
	4000	99,17±0,17	99,09±0,23	99,45±0,14
	8000	99,56±0,19	99,45±0,23	99,72±0,14
	11812	99,59±0,18	99,48±0,23	99,75±0,15
Validação	100	10,58±13,01	28,41±13,91	62,36±10,25
	250	15,00±13,44	29,36±12,63	62,34±9,37
	500	18,25±13,07	30,18±12,44	61,45±9,23
	1000	20,42±12,90	29,05±12,30	60,59±9,10
	2000	20,34±12,33	28,44±11,71	59,82±9,62
	4000	19,57±12,55	27,13±11,60	58,55±10,06
	8000	18,47±12,79	25,21±11,14	57,69±10,26
	11812	18,33±12,64	25,03±11,15	57,54±10,28

Tabela 6 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com Torneios utilizando regressão linear múltipla clássica e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	85,33±0,33	84,43±0,30	90,32±0,36
		250	91,68±0,36	90,23±0,43	94,64±0,17
		500	95,24±0,20	93,38±0,28	97,36±0,13
		1000	97,49±0,27	95,67±0,35	99,27±0,09
		2000	98,26±0,22	97,06±0,14	99,77±0,05
		4000	98,54±0,20	97,98±0,07	99,71±0,03
		8000	99,16±0,10	98,97±0,08	99,70±0,05
		11812	99,60±0,07	99,45±0,09	99,79±0,04
	50	100	89,90±0,18	89,79±0,24	93,73±0,29
		250	94,88±0,24	93,94±0,20	96,69±0,16
		500	97,44±0,15	96,42±0,18	98,45±0,12
		1000	98,85±0,10	97,71±0,09	99,60±0,05
		2000	99,37±0,12	98,31±0,16	99,84±0,04
		4000	99,39±0,15	98,71±0,19	99,70±0,14
		8000	99,41±0,10	99,17±0,10	99,75±0,04
		11812	99,62±0,06	99,49±0,08	99,78±0,05
	100	100	90,45±0,16	90,74±0,22	94,15±0,20
		250	96,65±0,15	96,26±0,13	97,59±0,08
		500	98,62±0,11	98,08±0,12	98,89±0,05
		1000	99,69±0,02	99,17±0,12	99,80±0,02
		2000	99,86±0,03	99,53±0,11	99,92±0,02
		4000	99,80±0,04	99,37±0,13	99,86±0,02
		8000	99,56±0,08	99,37±0,08	99,79±0,04
		11812	99,60±0,05	99,42±0,05	99,77±0,03

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	12,18±3,45	25,79±1,93	70,44±1,90
		250	14,81±3,69	28,14±1,50	73,04±1,57
		500	16,81±3,30	28,81±2,04	73,96±1,79
		1000	20,06±3,49	29,54±3,13	72,96±1,33
		2000	21,02±3,42	29,86±2,31	72,08±0,42
		4000	20,56±2,99	28,78±2,42	66,13±0,90
		8000	18,96±2,39	26,59±2,41	59,62±0,96
		11812	17,72±1,66	25,12±2,14	57,57±1,04
	50	100	12,25±2,96	24,79±3,03	72,73±1,77
		250	15,05±2,94	27,73±3,34	74,77±1,00
		500	17,61±2,54	27,30±3,56	74,18±1,57
		1000	18,55±2,76	28,11±2,95	72,84±1,28
		2000	18,90±2,47	28,63±3,00	69,74±2,00
		4000	19,66±1,97	28,34±2,32	63,64±1,79
		8000	19,20±2,18	26,33±1,17	59,10±1,74
		11812	18,98±1,55	24,88±1,45	57,02±2,26
	100	100	9,31±4,94	22,88±2,73	69,01±3,06
		250	14,02±4,22	26,89±2,15	73,69±2,00
		500	15,32±3,11	27,32±1,48	72,10±1,51
		1000	16,89±2,84	26,88±1,81	69,69±2,36
		2000	18,62±3,00	27,31±1,87	67,57±2,43
		4000	18,94±3,35	27,10±2,07	62,82±1,93
		8000	18,07±3,47	26,31±2,19	59,06±1,68
		11812	18,25±3,01	26,10±2,00	57,39±1,56

Tabela 7 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e y obtidas nas análises com Torneios utilizando LASSO Bayesiano e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	86,26±0,31	87,12±0,04	93,37±0,10
		250	92,00±0,36	92,08±0,14	96,38±0,01
		500	95,17±0,26	94,85±0,05	98,32±0,15
		1000	98,02±0,24	97,13±0,08	99,50±0,03
		2000	99,24±0,09	98,42±0,12	99,82±0,04
		4000	99,52±0,07	99,03±0,15	99,76±0,10
		8000	99,51±0,02	99,40±0,05	99,81±0,03
		11812	99,63±0,06	99,42±0,09	99,77±0,02
	50	100	90,68±0,02	91,40±0,05	95,37±0,01
		250	95,23±0,27	95,28±0,12	97,52±0,06
		500	97,70±0,27	97,40±0,22	98,87±0,04
		1000	99,15±0,16	98,60±0,08	99,61±0,06
		2000	99,69±0,03	99,25±0,13	99,86±0,03
		4000	99,60±0,10	99,46±0,03	99,84±0,03
		8000	99,56±0,11	99,45±0,11	99,83±0,01
		11812	99,58±0,05	99,47±0,06	99,76±0,06
	100	100	91,93±0,00	92,50±0,10	95,82±0,21
		250	97,65±0,11	97,23±0,17	98,51±0,05
		500	99,25±0,01	98,72±0,04	99,45±0,00
		1000	99,87±0,01	99,54±0,04	99,90±0,02
		2000	99,90±0,02	99,77±0,02	99,94±0,01
		4000	99,78±0,04	99,54±0,11	99,87±0,02
		8000	99,59±0,08	99,49±0,06	99,81±0,04
		11812	99,54±0,09	99,46±0,10	99,78±0,08

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	8,20±2,45	31,68±2,27	73,77±0,32
		250	11,13±1,49	33,28±0,73	75,27±1,08
		500	14,32±1,63	31,85±1,97	74,25±0,10
		1000	17,21±1,15	31,74±0,48	72,53±1,39
		2000	19,68±1,95	31,78±0,67	67,99±1,96
		4000	19,65±0,30	30,17±0,38	63,95±2,13
		8000	19,98±0,92	28,27±0,89	60,39±1,74
		11812	20,49±0,66	26,99±1,67	59,15±0,85
	50	100	5,95±3,54	27,33±3,09	75,07±0,39
		250	13,20±4,47	27,51±1,91	73,55±0,59
		500	18,80±3,44	26,67±1,83	72,80±0,82
		1000	18,52±3,97	27,48±0,07	70,17±2,39
		2000	20,79±3,36	28,59±1,14	66,01±2,21
		4000	22,20±2,51	27,97±1,55	62,02±0,99
		8000	21,86±1,63	26,30±1,72	59,98±1,49
		11812	21,51±1,69	25,48±1,92	58,05±1,84
	100	100	13,90±1,26	25,14±1,42	71,86±1,27
		250	12,89±0,23	28,32±1,42	72,96±0,61
		500	15,80±0,63	27,05±1,69	70,92±1,23
		1000	16,84±1,05	26,57±0,10	68,10±1,34
		2000	18,39±0,45	27,03±0,46	65,44±0,17
		4000	18,38±0,44	26,00±2,96	62,52±0,22
		8000	17,84±0,04	24,80±1,82	60,24±0,14
		11812	16,71±0,61	23,94±2,05	58,51±0,65

Tabela 8 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com o LASSO Bayesiano (sem torneios) considerando diferentes herdabilidades (h^2), diferentes números de SNPs de maiores módulos dos efeitos estimados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Nº de SNPs utilizados	h^2		
		0,25	0,5	1
Estimação	100	62,58±6,20	81,36±4,88	89,85±1,90
	250	63,24±3,44	82,86±2,54	93,82±0,77
	500	62,79±2,14	83,23±1,40	96,21±0,41
	1000	62,12±1,72	82,91±1,02	98,00±0,21
	2000	61,32±1,50	82,54±0,79	99,13±0,08
	4000	61,15±1,32	82,17±0,71	99,70±0,05
	8000	60,94±1,12	82,03±0,69	99,90±0,06
	11812	60,94±1,10	82,02±0,71	99,91±0,06
Validação	100	49,15±16,52	69,06±12,03	82,75±5,36
	250	47,67±13,86	66,91±10,63	83,18±3,82
	500	45,68±12,25	64,76±9,66	81,66±3,62
	1000	44,16±12,42	62,21±9,75	79,35±4,02
	2000	42,89±12,55	59,79±9,04	76,32±4,42
	4000	41,97±12,28	58,15±9,56	74,31±5,15
	8000	41,46±12,57	57,36±9,65	73,53±5,57
	11812	41,45±12,54	57,29±9,62	73,47±5,62

Tabela 9 Médias $\pm$ desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com Torneios utilizando regressão linear múltipla clássica e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	66,26±0,27	84,40±0,37	93,93±0,21
		250	66,12±0,36	85,36±0,19	96,80±0,15
		500	65,33±0,28	85,25±0,15	98,60±0,09
		1000	64,10±0,17	84,59±0,07	99,83±0,03
		2000	63,12±0,23	83,56±0,14	99,97±0,01
		4000	62,55±0,21	82,95±0,10	99,96±0,01
		8000	61,48±0,26	82,31±0,10	99,92±0,01
		11812	60,89±0,16	81,98±0,10	99,91±0,03
	50	100	64,08±0,31	83,73±0,31	95,76±0,22
		250	63,94±0,29	84,60±0,22	97,86±0,09
		500	63,37±0,23	84,31±0,15	99,19±0,06
		1000	62,21±0,29	83,59±0,16	99,94±0,02
		2000	61,32±0,17	82,79±0,21	99,98±0,01
		4000	61,16±0,21	82,45±0,11	99,95±0,04
		8000	61,12±0,15	82,26±0,14	99,92±0,03
		11812	60,82±0,19	82,05±0,12	99,92±0,02
	100	100	62,04±0,60	82,93±0,45	96,16±0,19
		250	62,48±0,34	83,53±0,14	98,43±0,11
		500	61,71±0,14	83,30±0,12	99,47±0,06
		1000	60,56±0,15	82,30±0,12	99,97±0,01
		2000	60,11±0,28	81,63±0,06	99,99±0,00
		4000	60,34±0,12	81,80±0,11	99,96±0,03
		8000	60,68±0,21	81,94±0,11	99,94±0,02
		11812	60,78±0,19	81,98±0,15	99,91±0,02

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	50,00±1,56	67,37±2,84	81,78±1,24
		250	48,83±2,37	68,35±2,15	84,78±1,16
		500	49,07±0,74	68,32±1,70	85,49±0,72
		1000	49,17±0,80	67,78±1,74	85,66±0,51
		2000	49,37±1,50	66,90±1,79	85,66±0,67
		4000	47,97±1,12	65,18±1,74	84,09±0,89
		8000	44,72±1,31	60,74±1,65	78,05±1,30
		11812	43,35±1,41	57,58±1,54	73,55±1,78
	50	100	46,59±2,61	62,97±2,95	83,81±0,88
		250	46,16±2,36	64,50±3,05	85,73±0,92
		500	46,69±2,75	63,96±2,69	86,75±1,12
		1000	46,98±3,26	64,29±2,59	86,00±0,91
		2000	45,40±3,17	64,31±2,09	85,87±0,64
		4000	44,81±2,74	63,96±1,53	83,31±0,95
		8000	44,64±2,38	60,88±1,87	77,81±1,05
		11812	43,61±1,98	58,68±1,47	73,70±1,50
	100	100	42,82±3,18	58,11±3,31	82,35±2,33
		250	43,56±2,46	60,49±3,19	85,39±1,62
		500	43,26±2,27	61,46±3,43	85,81±1,17
		1000	42,50±2,65	60,77±2,26	85,29±0,86
		2000	43,18±2,22	61,58±1,73	84,29±1,07
		4000	44,23±1,91	61,30±1,47	81,98±0,89
		8000	43,66±1,65	59,61±1,26	77,23±1,06
		11812	43,70±1,96	58,13±1,45	74,08±0,96

Tabela 10 Médias ± desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com Torneios utilizando regressão linear múltipla clássica e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	66,26±0,27	84,40±0,37	93,93±0,21
		250	66,12±0,36	85,36±0,19	96,80±0,15
		500	65,33±0,28	85,25±0,15	98,60±0,09
		1000	64,10±0,17	84,59±0,07	99,83±0,03
		2000	63,12±0,23	83,56±0,14	99,97±0,01
		4000	62,55±0,21	82,95±0,10	99,96±0,01
		8000	61,48±0,26	82,31±0,10	99,92±0,01
		11812	60,89±0,16	81,98±0,10	99,91±0,03
	50	100	64,08±0,31	83,73±0,31	95,76±0,22
		250	63,94±0,29	84,60±0,22	97,86±0,09
		500	63,37±0,23	84,31±0,15	99,19±0,06
		1000	62,21±0,29	83,59±0,16	99,94±0,02
		2000	61,32±0,17	82,79±0,21	99,98±0,01
		4000	61,16±0,21	82,45±0,11	99,95±0,04
		8000	61,12±0,15	82,26±0,14	99,92±0,03
		11812	60,82±0,19	82,05±0,12	99,92±0,02
	100	100	62,04±0,60	82,93±0,45	96,16±0,19
		250	62,48±0,34	83,53±0,14	98,43±0,11
		500	61,71±0,14	83,30±0,12	99,47±0,06
		1000	60,56±0,15	82,30±0,12	99,97±0,01
		2000	60,11±0,28	81,63±0,06	99,99±0,00
		4000	60,34±0,12	81,80±0,11	99,96±0,03
		8000	60,68±0,21	81,94±0,11	99,94±0,02
		11812	60,78±0,19	81,98±0,15	99,91±0,02

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	50,00±1,56	67,37±2,84	81,78±1,24
		250	48,83±2,37	68,35±2,15	84,78±1,16
		500	49,07±0,74	68,32±1,70	85,49±0,72
		1000	49,17±0,80	67,78±1,74	85,66±0,51
		2000	49,37±1,50	66,90±1,79	85,66±0,67
		4000	47,97±1,12	65,18±1,74	84,09±0,89
		8000	44,72±1,31	60,74±1,65	78,05±1,30
		11812	43,35±1,41	57,58±1,54	73,55±1,78
	50	100	46,59±2,61	62,97±2,95	83,81±0,88
		250	46,16±2,36	64,50±3,05	85,73±0,92
		500	46,69±2,75	63,96±2,69	86,75±1,12
		1000	46,98±3,26	64,29±2,59	86,00±0,91
		2000	45,40±3,17	64,31±2,09	85,87±0,64
		4000	44,81±2,74	63,96±1,53	83,31±0,95
		8000	44,64±2,38	60,88±1,87	77,81±1,05
		11812	43,61±1,98	58,68±1,47	73,70±1,50
	100	100	42,82±3,18	58,11±3,31	82,35±2,33
		250	43,56±2,46	60,49±3,19	85,39±1,62
		500	43,26±2,27	61,46±3,43	85,81±1,17
		1000	42,50±2,65	60,77±2,26	85,29±0,86
		2000	43,18±2,22	61,58±1,73	84,29±1,07
		4000	44,23±1,91	61,30±1,47	81,98±0,89
		8000	43,66±1,65	59,61±1,26	77,23±1,06
		11812	43,70±1,96	58,13±1,45	74,08±0,96

Tabela 11 Médias ± desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com Torneios utilizando LASSO Bayesiano e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais agrupados.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	66,48±1,70	84,99±0,94	95,83±0,42
		250	66,32±1,86	85,17±0,76	97,83±0,27
		500	65,59±1,75	84,94±0,80	99,10±0,19
		1000	63,93±1,53	84,17±0,81	99,87±0,08
		2000	62,33±1,68	82,98±0,84	99,98±0,02
		4000	61,44±1,64	82,32±0,78	99,96±0,03
		8000	61,12±1,49	82,08±0,81	99,92±0,08
		11812	60,78±1,51	82,04±0,79	99,91±0,11
	50	100	64,21±1,72	84,10±0,84	96,90±0,28
		250	64,05±1,54	84,39±0,74	98,39±0,24
		500	63,26±1,50	84,10±0,61	99,42±0,15
		1000	62,18±1,34	83,34±0,67	99,92±0,12
		2000	60,99±1,46	82,27±0,70	99,98±0,02
		4000	61,00±1,35	82,09±0,77	99,97±0,02
		8000	60,95±1,28	82,09±0,77	99,93±0,08
		11812	60,78±1,34	82,01±0,74	99,92±0,04
	100	100	62,29±1,93	82,93±1,02	97,26±0,24
		250	62,00±1,31	83,25±0,83	98,98±0,15
		500	61,33±1,37	83,04±0,84	99,71±0,08
		1000	60,36±1,43	82,12±0,92	99,98±0,02
		2000	60,24±1,55	81,68±0,94	99,99±0,01
		4000	60,53±1,35	81,91±0,92	99,94±0,18
		8000	61,06±1,31	81,95±0,80	99,93±0,04
		11812	60,85±1,35	81,91±0,76	99,90±0,12

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	46,50±9,59	69,06±6,12	85,94±4,55
		250	46,13±9,46	68,06±6,63	85,90±4,19
		500	47,70±11,47	66,15±7,57	85,38±3,50
		1000	46,30±10,39	65,08±7,74	84,47±3,50
		2000	44,84±11,35	63,05±8,93	82,27±4,40
		4000	44,77±10,52	61,66±8,55	79,28±5,06
		8000	44,58±10,36	59,61±10,54	75,17±5,60
		11812	44,28±10,56	59,01±10,35	73,47±6,02
	50	100	48,24±10,93	64,86±8,36	86,78±3,02
		250	45,94±10,99	64,77±7,02	86,75±3,38
		500	45,70±11,23	63,83±7,29	86,29±3,00
		1000	44,69±11,77	62,25±8,42	84,22±4,00
		2000	45,25±11,56	61,32±9,83	82,04±4,57
		4000	46,22±12,31	61,49±8,82	78,83±5,81
		8000	45,95±12,05	59,40±8,94	75,39±7,26
		11812	45,99±12,07	58,56±8,64	74,20±7,32
	100	100	42,10±10,88	61,56±8,28	85,02±3,63
		250	41,78±12,85	63,22±8,90	85,33±3,33
		500	42,23±11,68	61,85±9,54	84,35±3,80
		1000	41,33±11,50	60,54±10,27	82,54±4,11
		2000	42,65±10,89	60,58±9,55	81,04±4,12
		4000	43,59±10,02	59,92±8,83	78,13±5,19
		8000	44,38±10,02	59,92±8,83	75,25±5,58
		11812	43,89±9,92	58,94±9,20	72,90±6,44

Tabela 12 Médias ± desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com o LASSO Bayesiano (sem torneios) considerando diferentes herdabilidades (h^2), diferentes números de SNPs de maiores módulos dos efeitos estimados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Nº de SNPs utilizados	h^2		
		0,25	0,5	1
Estimação	100	46,35±4,36	63,24±4,80	83,97±2,19
	250	50,31±3,38	67,76±2,74	90,88±1,07
	500	52,03±2,48	69,48±1,76	94,70±0,56
	1000	52,74±2,15	70,54±1,26	97,21±0,35
	2000	52,83±1,98	70,97±1,03	98,70±0,18
	4000	52,75±1,96	71,05±1,05	99,45±0,14
	8000	52,41±1,86	70,92±0,99	99,72±0,14
	11812	52,34±1,87	70,90±0,99	99,75±0,15
Validação	100	28,05±12,08	42,02±13,03	62,36±10,25
	250	33,65±11,22	43,83±10,92	62,34±9,37
	500	37,51±10,39	44,05±11,10	61,45±9,23
	1000	39,25±10,07	42,76±11,10	60,59±9,10
	2000	39,91±9,60	41,79±11,19	59,82±9,62
	4000	39,38±9,91	40,85±11,38	58,55±10,06
	8000	38,33±9,86	39,52±11,47	57,69±10,26
	11812	38,21±9,90	39,38±11,47	57,54±10,28

Tabela 13 Médias ± desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com Torneios utilizando regressão linear múltipla clássica e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	54,90±0,69	70,96±0,44	90,32±0,36
		250	56,32±0,38	73,23±0,29	94,64±0,17
		500	56,08±0,32	73,88±0,24	97,36±0,13
		1000	55,25±0,31	73,64±0,21	99,27±0,09
		2000	54,66±0,22	73,15±0,14	99,77±0,05
		4000	54,38±0,24	72,61±0,07	99,71±0,03
		8000	53,35±0,18	71,61±0,09	99,70±0,05
		11812	52,33±0,18	70,96±0,13	99,79±0,04
	50	100	53,71±0,49	70,76±0,39	93,73±0,29
		250	54,60±0,39	72,47±0,22	96,69±0,16
		500	54,26±0,23	72,67±0,19	98,45±0,12
		1000	53,36±0,17	72,45±0,13	99,60±0,05
		2000	52,67±0,23	72,21±0,18	99,84±0,04
		4000	52,79±0,27	71,85±0,13	99,70±0,14
		8000	52,77±0,22	71,36±0,11	99,75±0,04
		11812	52,28±0,16	70,88±0,14	99,78±0,05
	100	100	51,20±0,66	68,65±0,52	94,15±0,20
		250	52,94±0,31	71,13±0,24	97,59±0,08
		500	52,67±0,14	71,32±0,22	98,89±0,05
		1000	51,62±0,08	70,96±0,18	99,80±0,02
		2000	51,29±0,12	70,71±0,22	99,92±0,02
		4000	51,69±0,13	70,97±0,10	99,86±0,02
		8000	52,39±0,16	71,05±0,12	99,79±0,04
		11812	52,35±0,10	71,00±0,09	99,77±0,03

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	31,78±4,33	41,61±3,02	70,44±1,90
		250	34,50±3,06	44,98±1,34	73,04±1,57
		500	36,61±2,69	45,82±1,38	73,96±1,79
		1000	38,56±1,74	46,20±1,81	72,96±1,33
		2000	39,25±2,40	45,84±1,62	72,08±0,42
		4000	39,55±1,97	43,99±2,37	66,13±0,90
		8000	37,44±1,70	41,37±2,19	59,62±0,96
		11812	36,22±1,54	39,59±2,13	57,57±1,04
	50	100	31,51±3,13	38,56±4,01	72,73±1,77
		250	35,33±2,27	41,70±3,35	74,77±1,00
		500	38,54±1,37	40,93±3,79	74,18±1,57
		1000	37,40±1,81	42,48±3,45	72,84±1,28
		2000	37,27±2,18	42,85±3,51	69,74±2,00
		4000	37,85±2,60	42,12±2,87	63,64±1,79
		8000	38,12±2,00	40,56±1,46	59,10±1,74
		11812	37,24±1,04	38,70±2,37	57,02±2,26
	100	100	25,32±5,09	34,39±3,14	69,01±3,06
		250	31,14±3,51	40,71±2,34	73,69±2,00
		500	31,35±3,74	40,94±2,16	72,10±1,51
		1000	33,08±3,39	40,47±2,19	69,69±2,36
		2000	34,93±2,54	40,44±1,66	67,57±2,43
		4000	35,91±2,24	41,02±1,65	62,82±1,93
		8000	36,46±2,15	40,16±1,75	59,06±1,68
		11812	36,81±2,16	39,98±1,86	57,39±1,56

Tabela 14 Médias ± desvios padrão (em %) das correlações entre $X\widehat{\beta}$ e $X\beta$ obtidas nas análises com Torneios utilizando LASSO Bayesiano e grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes herdabilidades (h^2), diferentes números de SNPs selecionados, para populações de estimação e validação num esquema de validação cruzada, com 100 repetições de cada análise, em um cenário de SNPs causais dispersos.

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Estimação	25	100	56,24±0,01	70,77±0,16	93,37±0,10
		250	56,70±0,03	72,50±0,16	96,38±0,01
		500	56,10±0,19	73,01±0,02	98,32±0,15
		1000	54,47±0,30	72,74±0,09	99,50±0,03
		2000	53,08±0,24	72,21±0,00	99,82±0,04
		4000	52,53±0,20	71,55±0,09	99,76±0,10
		8000	52,57±0,06	71,13±0,08	99,81±0,03
		11812	52,29±0,19	71,06±0,17	99,77±0,02
	50	100	54,45±0,46	70,72±0,28	95,37±0,01
		250	54,78±0,01	71,90±0,12	97,52±0,06
		500	54,07±0,21	72,00±0,04	98,87±0,04
		1000	52,93±0,20	71,61±0,16	99,61±0,06
		2000	51,93±0,04	71,17±0,11	99,86±0,03
		4000	52,36±0,33	70,98±0,02	99,84±0,03
		8000	52,41±0,20	70,94±0,19	99,83±0,01
		11812	52,44±0,18	70,93±0,08	99,76±0,06
	100	100	51,63±0,45	69,58±0,33	95,82±0,21
		250	52,75±0,06	71,18±0,15	98,51±0,05
		500	52,10±0,03	70,99±0,04	99,45±0,00
		1000	51,11±0,07	70,49±0,07	99,90±0,02
		2000	51,18±0,14	70,23±0,06	99,94±0,01
		4000	51,72±0,15	70,74±0,19	99,87±0,02
		8000	52,32±0,15	70,90±0,16	99,81±0,04
		11812	52,51±0,20	70,89±0,12	99,78±0,08

continua...

...conclusão

Tipo de população	Tamanho do grupo	Nº de SNPs utilizados	h^2		
			0,25	0,5	1
Validação	25	100	30,06±1,77	41,12±3,90	73,77±0,32
		250	32,90±4,16	44,88±1,62	75,27±1,08
		500	36,91±1,76	42,71±2,73	74,25±0,10
		1000	36,30±1,31	43,06±0,18	72,53±1,39
		2000	38,38±3,77	43,70±1,38	67,99±1,96
		4000	38,79±3,49	42,32±0,20	63,95±2,13
		8000	39,16±1,99	41,01±1,25	60,39±1,74
		11812	37,85±2,73	39,79±1,30	59,15±0,85
	50	100	28,23±5,04	41,31±1,52	75,07±0,39
		250	32,95±4,03	41,89±1,53	73,55±0,59
		500	36,25±1,57	40,32±2,39	72,80±0,82
		1000	36,41±0,85	41,06±0,59	70,17±2,39
		2000	36,47±0,02	42,27±1,02	66,01±2,21
		4000	37,42±0,48	42,09±0,16	62,02±0,99
		8000	36,94±1,05	40,08±0,51	59,98±1,49
		11812	37,59±0,52	39,07±0,28	58,05±1,84
	100	100	29,24±1,79	35,64±0,68	71,86±1,27
		250	33,24±1,33	42,13±2,65	72,96±0,61
		500	33,98±0,67	41,08±1,86	70,92±1,23
		1000	32,30±1,75	41,22±1,92	68,10±1,34
		2000	34,55±0,92	41,54±0,12	65,44±0,17
		4000	33,97±1,07	41,72±2,73	62,52±0,22
		8000	33,62±1,20	41,78±2,45	60,24±0,14
		11812	34,19±0,10	40,81±2,43	58,51±0,65

APÊNDICE B - Estimativas dos efeitos dos SNPs para o Lasso Bayesiano sem torneios

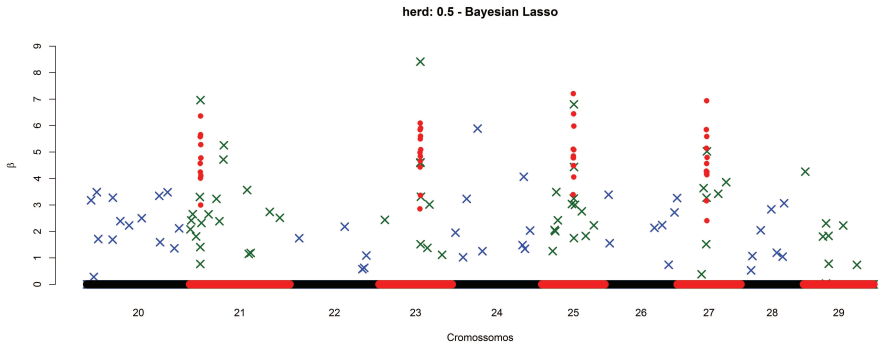

Figura 20 Estimativas dos efeitos dos SNPs obtidas pelo LASSO Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas do modelo completo, para uma herdabilidade de 0.5, com SNPs causais agrupados

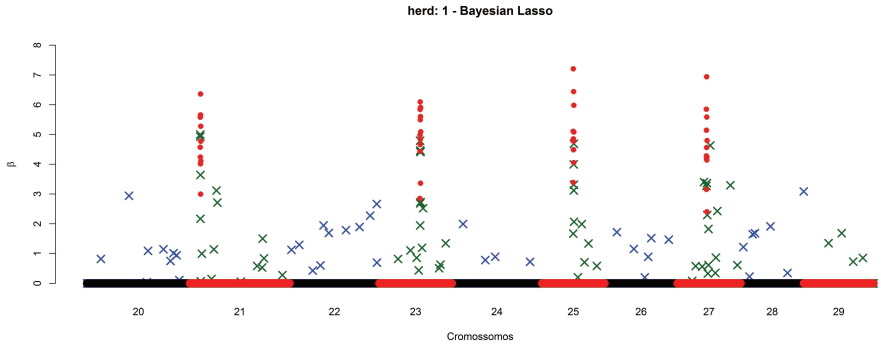

Figura 21 Estimativas dos efeitos dos SNPs obtidas pelo LASSO Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas (em módulo) do modelo completo, para uma herdabilidade de 1.0, em um cenário de SNPs causais agrupados

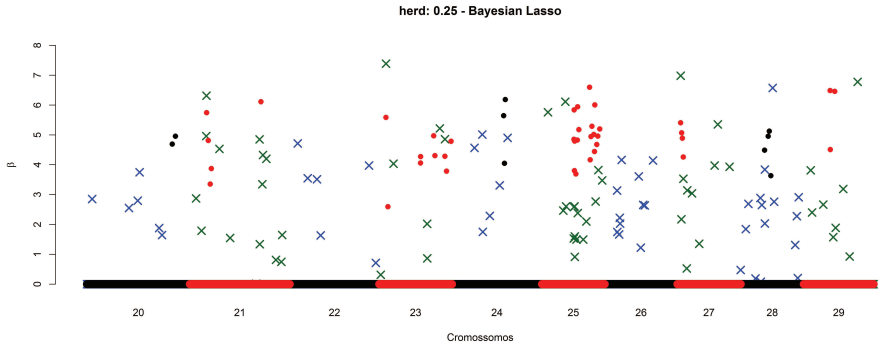

Figura 22 Estimativas dos efeitos dos SNPs obtidas pelo LASSO Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas (em módulo) do modelo completo, para uma herdabilidade de 0.25, em um cenário de SNPs causais dispersos

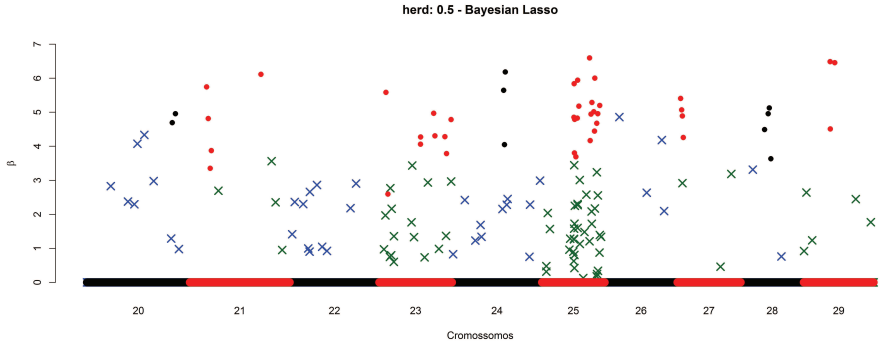

Figura 23 Estimativas dos efeitos dos SNPs obtidas pelo LASSO Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas (em módulo) do modelo completo, para uma herdabilidade de 0.5, em um cenário de SNPs causais dispersos

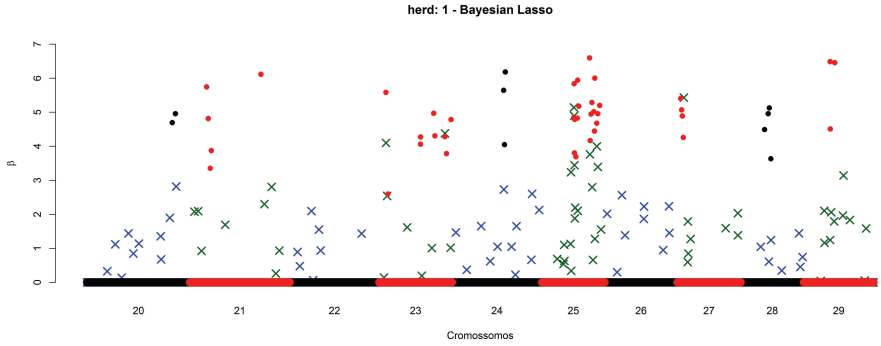

Figura 24 Estimativas dos efeitos dos SNPs obtidas pelo LASSO Bayesiano sem torneios, ajustado aos SNPs que obtiveram as 100 maiores estimativas (em módulo) do modelo completo, para uma herdabilidade de 1.0, em um cenário de SNPs causais dispersos

APÊNDICE C - Gráficos das estimativas dos efeitos dos SNPs para Torneios utilizando o LASSO Bayesiano na etapa de seleção de marcadores

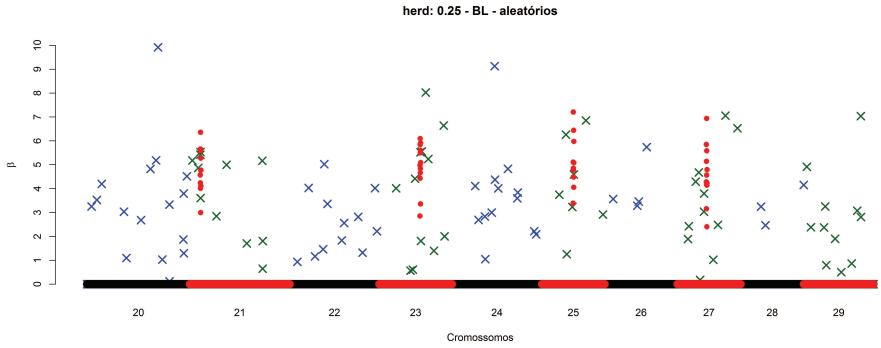

Figura 25 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 0.25, em um cenário de SNPs causais agrupados

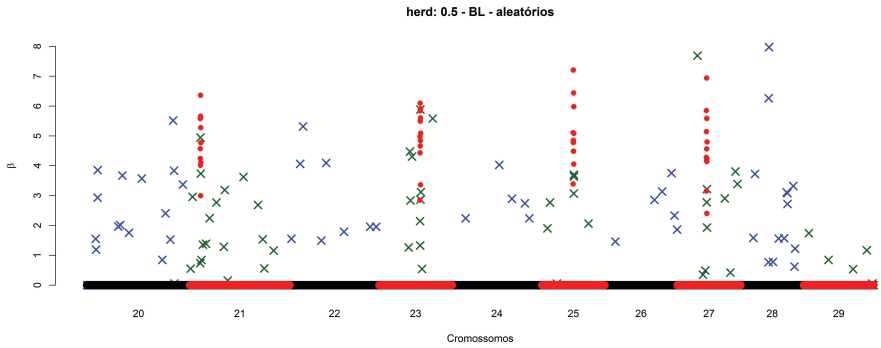

Figura 26 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 0.5, em um cenário de SNPs causais agrupados

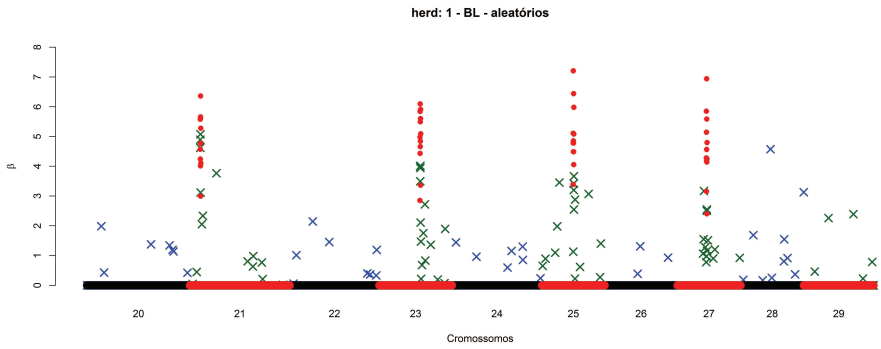

Figura 27 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 1.0, em um cenário de SNPs causais agrupados

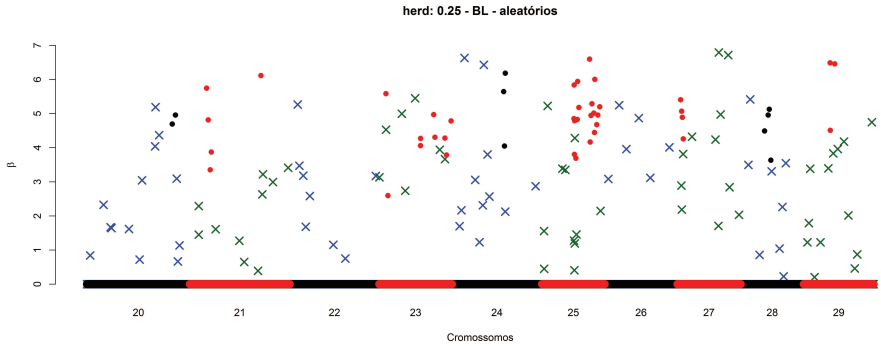

Figura 28 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 0.25, em um cenário de SNPs causais dispersos

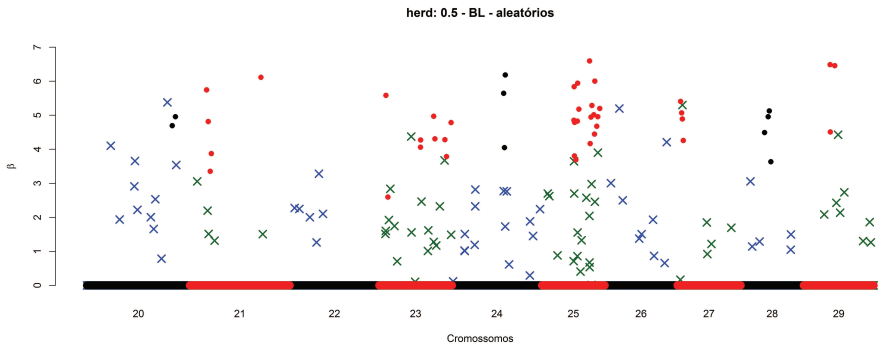

Figura 29 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 0.5, em um cenário de SNPs causais dispersos

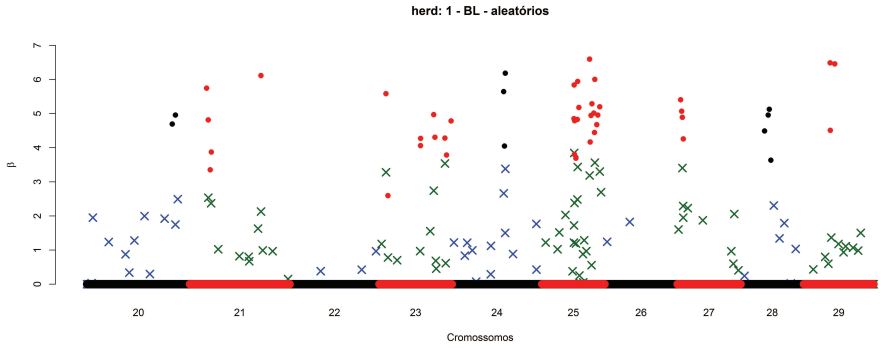

Figura 30 Estimativas dos efeitos dos SNPs obtidas pelo Torneio utilizando LASSO Bayesiano, para uma herdabilidade de 1.0, em um cenário de SNPs causais dispersos