DIÓGENES FERREIRA FILHO

TORNEIOS ENTRE MARCADORES COMO FORMA DE ENRIQUECER PREDIÇÕES GENÉTICAS

LAVRAS – MG

2014

DIÓGENES FERREIRA FILHO

TORNEIOS ENTRE MARCADORES COMO FORMA DE ENRIQUECER PREDIÇÕES GENÉTICAS

Tese apresentada à Universidade Federal de Lavras, 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. Júlio Sílvio de Sousa Bueno Filho

LAVRAS – MG

2014

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

Ferreira Filho, Diógenes.
Torneios entre marcadores como forma de enriquecer predições genéticas / Diógenes Ferreira Filho. – Lavras : UFLA, 2014.
121 p. : il.

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

1. Torneios. 2. Lasso Bayesiano. 3. GWAS. 4. GWS. 5. SNPs. I. Universidade Federal de Lavras. II. Título.

CDD – 519.537

DIÓGENES FERREIRA FILHO

TORNEIOS ENTRE MARCADORES COMO FORMA DE ENRIQUECER PREDIÇÕES GENÉTICAS

Tese apresentada à Universidade Federal de Lavras, 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 14 de agosto de 2014.

Dr. Daniel Furtado Ferreira	UFLA
Dr. Denismar Alves Nogueira	UNIFAL
Dr. Joel Augusto Muniz	UFLA
Dr. Roberto Hiroshi Higa	EMBRAPA

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS – MG

2014

À minha família, o alicerce da minha vida.
Em especial, aos meus pais, Diógenes Ferreira e Aparecida Antônia Vendramel Ferreira, e à minha esposa, Larissa Carvalho Vilas Boas, por todo amor, apoio, respeito, confiança e paciência. Jamais teria conseguido sem vocês!!!

DEDICO

AGRADECIMENTOS

A Deus, por tudo que sou e conquistei.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX), pela oportunidade concedida para realização do doutorado.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), pela concessão da bolsa de estudos.

Aos membros da banca, pela disponibilidade em ajudar a finalizar este trabalho.

Ao meu orientador Júlio, pela sua grande ajuda, atenção e confiança depositada em mim.

À professora Sarah, por toda ajuda, atenção, paciência e compreensão.

À Embrapa Pecuária Sudeste por ceder os dados utilizados na tese.

Aos meus pais, Diógenes e Aparecida, sem os quais eu nunca teria chegado até aqui, por todo seu amor, cuidado, compreensão e apoio incondicional.

À minha esposa Larissa, amor da minha vida, por estar ao meu lado em todos os momentos, por me dar forças para seguir em frente, por todo seu amor, cuidado, compreensão e apoio incondicional. Jamais teria conseguido sem você.

Aos meus irmãos, Guilherme e Daniela por todo apoio, carinho e incentivo.

À minha sogra Maria de Fátima, por todo apoio, carinho, incentivo e compreensão.

A todos os meus amigos, que de alguma forma contribuíram para essa conquista.

Muito obrigado!!!

RESUMO

Em estudos de associação genômica ampla (GWAS) e seleção genômica ampla (GWS) há dois problemas metodológicos que limitam a análise estatística: alta dimensionalidade ($n \ll p$) e multicolinearidade. Neste trabalho, foi revisitada uma estratégia de organização de torneios entre amostras aleatórias de marcadores, em que cada amostra tem boas propriedades estatísticas para estimação ($n > p$). Tais torneios são elaborados de modo a eliminar marcadores mais lentamente, usando regressão linear múltipla, adaptando sugestões anteriores encontradas na literatura. Isto não apenas contorna o problema $n \ll p$, mas também minimiza associações espúrias. Outra possível melhoria foi investigada, e se baseia em formar os grupos com marcadores tomados de diferentes cromossomos para minimizar a colinearidade dentro de grupos. Foram comparadas as estratégias em ambos os estudos com dados simulados e reais. A simulação foi realizada com genótipos reais, os quais foram, posteriormente, analisados com fenótipos reais. Os dados são provenientes de um estudo de SNPs em gado de corte (384 animais da raça Canchin genotipados para 526.493 SNPs e fenotipados para área de olho de lombo). Foram utilizadas, como critério de comparação, a capacidade de selecionar SNPs próximos do efeito simulado, as capacidades de predição genotípica e fenotípica, e também uma validação cruzada para os dados reais. O Lasso Bayesiano (BL) foi utilizado como referência (estimando os efeitos de todos os marcadores para selecioná-los) e também para obter estimativas dos efeitos dos SNPs selecionados no final dos torneios. Na maioria das situações simuladas os torneios foram igualmente precisos e ligeiramente mais acurados que o BL. No entanto, quando se usou dados reais, os torneios (ambas as estratégias) superaram muito a acurácia de predição obtida pelo BL. Para fins de GWAS, ambas as estratégias de torneios tendem a selecionar os mesmos SNPs, de forma mais consistente que o BL, que tende a selecionar qualquer uma das segregações que representam o mesmo efeito. Reduzir a colinearidade mostrou-se uma boa estratégia, mesmo que posteriormente a análise seja feita com o BL. Entre as estratégias de torneios, a mais simples (grupos formados aleatoriamente) foi a melhor, produzindo o mesmo resultado e, em um tempo que foi uma fração das outras metodologias. Para os dados reais, os resultados são promissores. Ao selecionar 104 SNPs, a correlação entre GBVs preditos e fenótipos alcançou 90,32% no conjunto de validação, mostrando a eficiência dos torneios na identificação de SNPs relevantes (ou segregações) para GWS. O código R para melhores benefícios da estratégia de torneios por meio de programação paralela simples é disponibilizado.

Palavras-chave: Torneios. Lasso Bayesiano. GWAS. GWS. SNPs.

ABSTRACT

In genome-wide association studies (GWAS) and genome-wide selection (GWS) there are two methodological issues that restrict statistical analysis: high dimensionality ($n \ll p$) and multicollinearity. In this work, we revisit an organization strategy of tournaments between random marker samples, in which each sample presents good statistical properties for estimation ($n > p$). Such tournaments are elaborated in such a way to eliminate markers more slowly, using multiple linear regression, adapting previous suggestions found in literature. This not only circumvents the $n \ll p$ problem but also minimizes spurious associations. Another possible improvement was investigated, and is based on forming groups with markers taken from different chromosomes to minimize within group collinearity. The strategies were compared in both studies using simulated and real data. The simulation was performed with real genotypes, which were, subsequently, analyzed with real phenotypes. The data are derived from a study with SNPs in beef cattle (384 animals of the Canchim breed, genotyped for 526,493 SNPs and phenotyped for the loin eye area). As comparison criteria, we used the capacity of selecting SNPs near the simulated effect, the genotype and phenotype prediction capabilities, and also an cross validation for the real data. The Bayesian Lasso (BL) was used as reference (estimating the effects of all markers to select them) and also to obtain estimates of the effects of the SNPs selected at the end of the tournaments. In most simulated situations, the tournaments were equally precise and a slightly more accurate than the BL. However, when real data was used, the tournaments (both strategies) far overcomes the prediction accuracy obtained by the BL. For GWAS purposes, both tournament strategies tend to select the same SNPs, and clearly overcomes the BL, which tends to select any of the segregations that represent the same effect. Reducing collinearity showed to be a good strategy, even if later the analysis be performed with the BL. Among the tournament strategies, the simpler (groups randomly formed) was the best overall, producing the same result and, in time that was a fraction of the other methodologies. For real data, the results are promising. When selecting 104 SNPs, the correlation between predicted GBVs and phenotypes reached 90.32% in the validation set, showing the efficiency of the tournaments in identifying relevant SNPs (or segregations) for GWS. The R code for better benefits tournaments strategy by simple parallel programming is available.

Keywords: Tournaments. Bayesian Lasso. GWAS. GWS. SNPs.

LISTA DE FIGURAS

Figura 1 Single Nucleotide Polimorphism (SNP)..... 26

Figura 2 Gráficos das funções densidade das distribuições Exponencial Dupla e Normal..... 34

Figura 3 Dados de área de olho de lombo corrigidos para efeitos de grupos de contemporâneos de 384 bovinos da raça Canchim..... 43

Figura 4 Normal QQ-Plot para dados de AOL corrigidos para efeitos de grupos de contemporâneos de 384 bovinos da raça Canchim..... 43

Figura 5 Efeitos simulados de 48 SNPs com efeitos não nulos, distribuídos entre 4 cromossomos, próximos entre si no cromossomo..... 45

Figura 6 Efeitos simulados de 48 SNPs com efeitos não nulos, distribuídos entre oito cromossomos, um pouco mais dispersos entre si no cromossomo..... 46

Figura 7 Efeitos simulados de 250 SNPs com efeitos não nulos, distribuídos por todos os cromossomos, dispersos entre si no cromossomo..... 46

Figura 8 GBVs simulados em uma situação de 48 SNPs não nulos agrupados..... 48

Figura 9 GBVs simulados em uma situação de 48 SNPs não nulos dispersos..... 48

Figura 10 GBVs simulados em uma situação de 250 SNPs não nulos dispersos..... 48

Figura 11 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 48 SNPs não nulos agrupados, considerando diferentes herdabilidades..... 49

Figura 12 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 48 SNPs não nulos dispersos, considerando diferentes herdabilidades 49
Figura 13 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 250 SNPs não nulos dispersos, considerando diferentes herdabilidades 49
Figura 14 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 60
Figura 15 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 61
Figura 16 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 63
Figura 17 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 64

Figura 18 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 65
Figura 19 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos..... 66
Figura 20 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos..... 66
Figura 21 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada,para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 67
Figura 22 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos..... 68

Figura 23 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos.....68
Figura 24 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 69
Figura 25 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos.....70
Figura 26 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos.....70
Figura 27 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 72
Figura 28 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes

números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 73
Figura 29 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 73
Figura 30 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 0,25 76
Figura 31 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 0,5 77
Figura 32 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 1,0 78

Figura 33 Correlações médias entre GBVs preditos e fenótipos para amostras de estimação (a) e validação (b), considerando diferentes metodologias e diferentes números de SNPs selecionados, para a característica área de olho de lombo 80
Figura 34 Frequências dos SNPs selecionados em 100 torneios com grupos aleatórios, com 100 SNPs selecionados em cada torneio, para a característica área de olho de lombo 83
Figura 35 Estimativas dos efeitos de 100 SNPs selecionados em um torneio com grupos aleatórios, com HPDs de 95% para os marcadores cujos efeitos diferem significativamente de zero, para a característica área de olho de lombo..... 84
Figura 36 Frequências, em 100 análises, dos SNPs selecionados por HPDs de 95%, calculados para os SNPs previamente selecionados em torneios com grupos aleatórios para a característica área de olho de lombo 85
Figura 37 Resultado de uma regressão linear múltipla clássica utilizando 100 SNPs dispostos sequencialmente no cromossomo 1 93
Figura 38 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 94
Figura 39 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos..... 94

Figura 40 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 95
Figura 41 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 95
Figura 42 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 96
Figura 43 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 96
Figura 44 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 97
Figura 45 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados

e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 97
Figura 46 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados 98
Figura 47 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 98
Figura 48 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 99
Figura 49 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 99
Figura 50 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos 100
Figura 51 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos 100

Figura 52 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 0,25 101
Figura 53 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 0,5 102
Figura 54 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 1,0 103
Figura 55 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 0,25 104
Figura 56 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um

cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 0,5 105

Figura 57 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 1,0 106

LISTA DE TABELAS

Tabela 1 Quantidade de SNPs por cromossomo após o controle de qualidade 41

Tabela 2 Estatísticas descritivas dos dados originais e corrigidos para efeitos de grupos de contemporâneos (GC) da característica área de olho de lombo (AOL), de 384 bovinos da raça Canchim 42

Tabela 3 Quantidade de SNPs por cromossomo no conjunto de dados reduzido utilizado no estudo de simulação 44

Tabela 4 Medidas descritivas dos coeficientes de regressão e de seus erros padrões obtidos em regressões lineares múltiplas, com grupos de SNPs próximos (localizados sequencialmente) no cromossomo 1, grupos formados por SNPs tomados aleatoriamente no cromossomo 1, e quantidade de coeficientes que não puderam ser estimados devido às colinearidades perfeitas 58

Tabela 5 Tempo de execução das análises para obtenção da classificação dos marcadores para a característica área de olho de lombo 82

Tabela 6 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação e validação de um esquema de validação cruzada, considerando marcadores selecionados por HPDs de 95% em 100 análises, com modelos ajustados a grupos de marcadores selecionados com determinadas frequências mínimas nas 100 análises, para a característica área de olho de lombo 86

Tabela 7 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para torneios com grupos aleatórios considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 107
Tabela 8 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 109
Tabela 9 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 111
Tabela 10 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 113

Tabela 11 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para o Lasso Bayesiano considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 115
Tabela 12 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para o Lasso Bayesiano considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos 116

SUMÁRIO

1 INTRODUÇÃO 22
2 REFERENCIAL TEÓRICO 26
2.1 SNPs 26
2.2 Estudos de associação genômica ampla (GWAS) 27
2.3 Seleção Genômica Ampla (GWS) 29
2.4 Lasso Bayesiano 32
2.5 *Tournament Screening* 37
3 MATERIAL E MÉTODOS 40
3.1 Dados 40
3.2 Seleção de marcadores utilizando torneios 50
3.3 Seleção de marcadores com maiores módulos das estimativas pelo Lasso Bayesiano 52
3.4 Obtenção do modelo final 52
3.5 Verificação dos efeitos da multicolinearidade para SNPs próximos 52
3.6 Situações consideradas no estudo de simulação 53
3.7 Avaliação dos marcadores selecionados 53
3.8 Avaliação da predição de valores genéticos 54
3.9 Análises com fenótipos reais de área de olho de lombo (AOL) 55
4 RESULTADOS E DISCUSSÃO 57
4.1 Verificação da redução dos efeitos da multicolinearidade por meio da formação de grupos de marcadores 57
4.2 Influência do tamanho dos grupos na predição dos valores genéticos 59
4.3 Avaliação da capacidade de predição de valores genéticos sem utilizar validação cruzada considerando torneios com grupos de 25 marcadores 62
4.4 Avaliação da capacidade de predição de valores genéticos utilizando validação cruzada considerando torneios com grupos de 25 marcadores 64
4.5 Seleção de marcadores considerando torneios com grupos de 25 marcadores 74
4.6 Resultados das análises com fenótipos reais considerando torneios com grupos de 25 marcadores 79
5 CONCLUSÃO 88
REFERÊNCIAS 90
APÊNDICES 93

1 INTRODUÇÃO

O DNA é o material genético dos organismos vivos, assim, variações fenotípicas entre indivíduos de uma população podem ser reflexo de diferenças nas sequências de nucleotídeos do DNA desses indivíduos. O desenvolvimento da genética molecular permitiu a identificação de regiões polimórficas do DNA, algumas das quais estão associadas às alterações no desempenho de indivíduos para características de interesse.

Com o desenvolvimento dos marcadores moleculares criou-se a expectativa de que informações genotípicas desses marcadores, uma vez correlacionadas com características fenotípicas de interesse, pudessem ser utilizadas na obtenção e seleção de indivíduos com maior valor genético (ALMEIDA, 2013).

Avanços tecnológicos para automação do processo de genotipagem permitiram o desenvolvimento de novas classes de marcadores moleculares, dentre os quais se destacamos polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphism* - SNP). Uma das principais vantagens deste tipo de marcador é a possibilidade de genotipar simultaneamente centenas de milhares de SNPs amplamente distribuídos pelo genoma, de tal modo que é de se esperar que parte destes marcadores esteja em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci* - QTL).

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, pode-se fazer testes de hipóteses e predições (LI; SILLANPÄÄ, 2012). 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 genéticos genômicos

(*Genomic Breeding Value*- GBV) de indivíduos candidatos à seleção em programas de melhoramento.

A seleção genômica ampla (*Genome-Wide Selection* - GWS), proposta por Meuwissen, Hayes e Goddard (2001), baseia-se na estimação simultânea dos efeitos de milhares de marcadores, que são utilizados para predizer os GBVs de indivíduos e, com base nesses valores, identificar indivíduos geneticamente superiores. Na GWS os efeitos de todos os marcadores são estimados simultaneamente sem a necessidade prévia de identificar os marcadores com efeitos significativos (RESENDE et al., 2008).

O grande número de marcadores, no entanto, implica em problemas de multicolinearidade (diferentes marcadores com o mesmo perfil genotípico) e de dimensionalidade (número de marcadores muito maior que o número de indivíduos genotipados) (SILVA et al., 2013). Neste tipo de situação, o método dos mínimos quadrados ordinários se torna inaplicável, sendo necessárias outras metodologias estatísticas. Dentre os métodos estatísticos utilizados em GWS, um dos de maior destaque é o Lasso Bayesiano, que é uma formulação bayesiana do método Lasso. Resumidamente, o método Lasso (*Least Absolute Shrinkage and Selection Operator*) se caracteriza como um método de regressão penalizada, inicialmente proposto por Tibshirani (1996), e sua formulação bayesiana (Lasso Bayesiano) foi apresentada por Park e Casella (2008). O Lasso Bayesiano tem sido utilizado em vários estudos de associação (LI et al., 2011; YI; XU, 2008) e de seleção genômica (DE LOS CAMPOS et al., 2009; LEGARRA et al., 2011; SILVA et al., 2011).

Meuwissen (2007) chama atenção para o fato de que a utilização de milhares de marcadores pode fazer com que variações devidas ao erro sejam explicadas pelos efeitos dos marcadores. Neste caso, a capacidade preditiva do modelo na amostra de estimação (indivíduos que foram utilizados no ajuste do modelo) apresentará bom desempenho, mas, em contrapartida, não apresentará

bom desempenho em uma amostra de validação (indivíduos que não foram utilizados no ajuste do modelo). Para avaliar a capacidade de predição do modelo, pode-se utilizar a técnica de validação cruzada. Resende et al. (2012) propuseram a utilização de subconjuntos com diferentes números de marcadores (100, 250, 500, 1.000,...) que obtiveram maiores módulos das estimativas no modelo completo (modelo contendo todos os marcadores) e, por meio de validações cruzadas, baseando-se nas correlações entre valores genéticos preditos e fenótipos observados na amostra de validação, verifica-se o número aproximado de marcadores que proporciona melhor capacidade preditiva. A escolha desses subconjuntos de marcadores baseia-se, portanto, na classificação dos marcadores com maiores módulos das estimativas no modelo completo.

Esse tipo de abordagem, no entanto, demanda muito tempo e esforço computacional, sem garantias de resolver o problema da multicolinearidade que é muito comum neste tipo de estudo. Uma metodologia alternativa que pode ser utilizada para selecionar marcadores e apresenta melhor desempenho computacional é a seleção de variáveis por torneios (CHEN; CHEN, 2009). Esta metodologia consiste em selecionar variáveis por meio de torneios de análises entre as variáveis regressoras em que, em cada etapa do torneio, uma determinada quantidade de variáveis com melhores estimativas são selecionadas.

Uma hipótese que surge da literatura é que realizar a seleção de marcadores por meio de torneios pode ser um modo de amenizar o problema da multicolinearidade, pois, os sorteios dos marcadores para formação dos grupos diminuem as ocorrências de marcadores com o mesmo perfil genotípico. Outra vantagem dessa metodologia é que ela permite uma abordagem computacional mais eficiente, pois, as análises em cada grupo são extremamente rápidas se comparadas com uma análise utilizando todos os marcadores e, além disso, segundo Beleti Junior (2013), aplicações com grande quantidade de processamento podem obter melhor desempenho se implementadas sob modelos

que possibilitem multiprocessamento, ou processamento paralelo. Como as análises em cada grupo são independentes umas das outras, elas podem ser executadas simultaneamente por meio de programação paralela utilizando um computador com vários núcleos (*cores*) ou utilizando um *cluster*.

Alves (2014) propôs uma metodologia para seleção de marcadores por meio de torneios de análises de regressão, tendo como motivação a metodologia de torneios proposta por Chen e Chen (2009). A ideia central da metodologia proposta por Alves (2014) é formar grupos aleatórios de marcadores e submeter cada grupo a uma análise de regressão linear múltipla clássica. O marcador que obtiver o maior valor-*p* em cada grupo é eliminado e os marcadores que não foram eliminados formam novos grupos aleatórios para a próxima etapa do torneio. O processo é repetido até que o número de marcadores seja reduzido a um nível desejável.

Considerando a metodologia de torneios proposta por Alves (2014) e levando em conta que existe um grande número de SNPs em todos os cromossomos, surge a hipótese de que condicionar a formação dos grupos dos torneios à estrutura de cromossomos pode reduzir ainda mais os efeitos da multicolinearidade nas análises de regressão e fornecer mais segurança para a seleção de marcadores.

O objetivo do presente trabalho foi verificar se o modo como os grupos são formados (aleatórios ou condicionados aos cromossomos) em torneios de análises entre marcadores tem influência sobre a seleção de marcadores e sobre a capacidade de predição de valores genéticos. Ambas as estratégias serão comparadas ao Lasso Bayesiano, tomado como padrão em estudos de simulação e na análise de dados reais.

2 REFERENCIAL TEÓRICO

2.1 SNPs

Um polimorfismo de nucleotídeo único (SNP) é uma alteração em uma determinada sequência do DNA em que indivíduos de uma população diferem nesta sequência por uma única base (Figura 1).

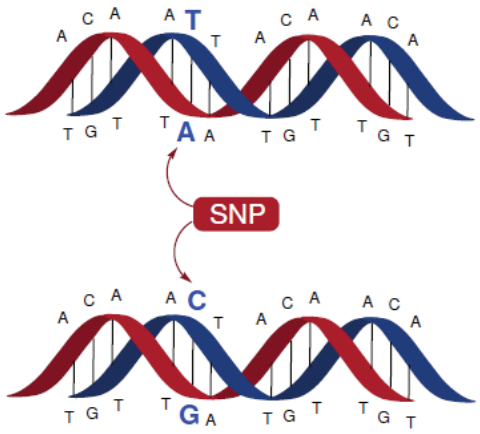

Figura 1 Single Nucleotide Polimorphism (SNP)
Fonte: Agnihotram (2012)

Segundo Manolio, Brooks e Collins (2008) a maioria dos SNPs são bialélicos, ou seja, tem apenas duas variantes, ou alelos. Assim, denotando os alelos de um SNP por A e B, então, em organismos diploides, o genótipo de um SNP pode ser AA, AB ou BB.

Após o desenvolvimento de tecnologias para o sequenciamento do genoma, que possibilitaram a identificação de SNPs, foram desenvolvidos *chips* de alta densidade que permitem a genotipagem simultânea de milhares de SNPs amplamente distribuídos ao longo do genoma.

2.2 Estudos de associação genômica ampla (GWAS)

Estudos com marcadores moleculares no melhoramento genético podem ser divididos em duas linhas: a detecção de marcadores associados à QTLs e mapeamento destes; e o uso dos marcadores nos programas de seleção genética por meio da seleção auxiliada por marcadores (*Marker Assisted Selection* - MAS) e seleção genômica ampla (*Genome-Wide Selection* - GWS) (RESENDE et al., 2012).

A GWAS (*Genome-Wide Association Studies*) procura associações entre *locos* e caráter fenotípico em nível populacional, por meio de testes de hipóteses visando detectar efeitos com significância estatística.

Associações entre SNPs e fenótipos podem ser identificadas por meio de regressões sobre marcadores individuais. O efeito do SNP é considerado significativo se tiver valor-*p* abaixo de um determinado nível de significância e, como muitos marcadores são testados, é utilizado algum procedimento para testes múltiplos para controlar a taxa de erro tipo I.

Testar associação de um SNP por vez pode ser uma opção sensata quando a característica é controlada por um ou poucos *locos*, entretanto, existem fortes evidências de que muitas características importantes são controladas por um grande número de genes (JANSS et al., 2012). Assim, a alteração no fenótipo é, geralmente, devida a múltiplos SNPs em várias posições no cromossomo com efeitos individuais pequenos, mas que, conjuntamente, tem um grande efeito sobre o fenótipo. Nesta situação, um modelo de regressão múltipla, em que cada SNP é uma variável regressora, pode descrever de forma mais adequada a relação entre fenótipos e SNPs. A equação (1) descreve esta relação:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad (1)$$

Em que:

a) y_i $(i = 1, 2, \ldots, n)$ é o valor do fenótipo do indivíduo i;

b) μ é a média geral, uma constante inerente a todas as observações;

c) x_{ij} é o genótipo do SNP j $(j = 1, 2, \ldots, p)$, do indivíduo i, codificado como $0, 1$ ou 2 com base no número de cópias de um dos alelos do SNP;

d) β_j é o efeito principal do SNP j;

e) ε_i é o erro aleatório $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

Para exemplificar a codificação do genótipo do SNP, considere o número de cópias do alelo A; se o genótipo do SNP for BB (zero cópias do alelo A) então $x_{ij} = 0$, se o genótipo for AB (uma cópia do alelo A) então $x_{ij} = 1$ e se o genótipo for AA (duas cópias do alelo A) então $x_{ij} = 2$.

O problema de selecionar SNPs associados à característica pode então ser tratado como um problema de seleção de variáveis. No entanto, a presença de forte multicolinearidade (diferentes marcadores com o mesmo perfil genotípico) e a alta dimensionalidade dos marcadores ($p \gg n$), tornam os métodos de seleção de variáveis convencionais inaplicáveis ou ineficientes.

O Lasso Bayesiano é uma metodologia que tem sido utilizada em vários estudos de associação genômica (LI et al., 2011; YI; XU, 2008). O Lasso (*Least Absolute Shrinkage and Selection Operator*), proposto por Tibshirani (1996), se caracteriza como um método de regressão penalizada. Park e Casella (2008) propuseram uma formulação bayesiana do método Lasso denominada Lasso Bayesiano, em que o estimador Lasso é interpretado como uma estimativa da moda *a posteriori* em um contexto bayesiano, através da distribuição *a priori*

Laplace (Exponencial Dupla). Embora este método não realize seleção de variáveis automaticamente, intervalos de credibilidade para as estimativas dos efeitos dos marcadores podem ser utilizados para esta finalidade, sendo selecionados os marcadores cujos intervalos não contêm o zero.

Marcadores significativos detectados por GWAS também podem ser utilizados para predição de valores genéticos genômicos. No entanto, para a maioria das características, SNPs detectados por GWAS explicam apenas uma pequena fração da herdabilidade (YANG et al., 2011).

2.3 Seleção Genômica Ampla (GWS)

A seleção assistida por marcadores (*Marker Assisted Selection* - MAS) baseia-se na detecção, mapeamento e utilização de QTLs de grande efeito na identificação de indivíduos geneticamente superiores (RESENDE et al., 2008). Entretanto, a natureza poligênica de caracteres quantitativos e a alta influência ambiental permitem a detecção de apenas um pequeno número de QTLs de grandes efeitos que não explicam suficientemente a variação genética.

Com o desenvolvimento de tecnologias de genotipagem em larga escala, Meuwissen, Hayes e Goddard (2001) propuseram uma metodologia para avaliação genética cujo objetivo era capturar os efeitos de todos os *locos* e explicar quase a totalidade da variância genética de uma característica quantitativa. Esta metodologia ficou conhecida como seleção genômica ampla (GWS - *Genome-Wide Selection*). Segundo Togashi, Lin e Yamazaki (2011), a GWS é basicamente uma extensão da tradicional seleção assistida por marcadores, porém com um número muito maior de marcadores.

Na seleção genômica ampla, são utilizados milhares de marcadores, cobrindo amplamente o genoma, de forma que todos os QTLs estejam em desequilíbrio de ligação com algum marcador. Assim, os efeitos dos marcadores

podem ser estimados simultaneamente sem a necessidade prévia de identificar os marcadores com efeitos significativos e de mapear QTLs, como no caso da MAS (RESENDE et al., 2008). As estimativas dos efeitos dos marcadores são usadas para fornecer o valor genético genômico (*Genomic Breeding Value* - GBV) dos indivíduos candidatos à seleção.

Esse tipo de estudo, no entanto, apresenta dois grandes obstáculos para a análise estatística, sendo um deles a multicolinearidade (diferentes marcadores com o mesmo perfil genotípico) e o outro, a dimensionalidade ($p \gg n$) (SILVA et al., 2013). A multicolinearidade se refere à situação em que há uma relação linear exata ou aproximadamente exata entre as variáveis regressoras, que tem como consequências coeficientes indeterminados e erros padrão não definidos (colinearidade perfeita) ou erros padrão grandes e estimativas imprecisas dos coeficientes (colinearidade alta, mas não perfeita). Além disso, a estimação dos efeitos de um número elevado de marcadores a partir de um número limitado de indivíduos impossibilita o ajuste de todos os efeitos simultaneamente por mínimos quadrados ordinários, necessitando de outras metodologias de análise.

De um modo geral, os métodos estatísticos mais utilizados em GWS são RR-BLUP, G-BLUP, Bayes A, Bayes B e Lasso Bayesiano. Além dessas, outras metodologias que se baseiam em redução de dimensionalidade, como regressão via quadrados mínimos parciais, regressão via componentes principais e regressão via componentes independentes vêm sendo utilizadas neste tipo de estudo. O método de regressão via quadrados mínimos parciais permite ainda, uma abordagem multivariada, considerando múltiplos fenótipos (AZEVEDO et al., 2013). Dentre os vários métodos utilizados, o Lasso Bayesiano é um dos de maior destaque e vem sendo utilizado em vários estudos de associação (LI et al., 2011; YI; XU, 2008) e de seleção genômica (DE LOS CAMPOS et al., 2009; LEGARRA et al., 2011; SILVA et al., 2011).

Um dos principais objetivos da GWS é fazer predições sobre valores genéticos de indivíduos que não tiveram seus fenótipos avaliados diretamente, ou seja, fazer predições para indivíduos apenas genotipados (SILVA et al., 2011). Dessa forma, a predição dos valores genéticos e a seleção dos indivíduos podem ser realizadas em fases muito juvenis, acelerando o processo de melhoramento genético.

Segundo Resende et al. (2012), na seleção genômica ampla, primeiramente, os efeitos dos marcadores são estimados com base em dados genotípicos e fenotípicos de uma população de estimação, que consiste de indivíduos genotipados (grande número de SNPs) e com seus fenótipos avaliados para a característica de interesse. A partir das estimativas dos marcadores é obtido um modelo para predição dos valores genéticos dos indivíduos. Este modelo é aplicado em uma população de validação, que consiste de indivíduos genotipados e fenotipados, masque não foram utilizados na obtenção do modelo, e contempla um número de indivíduos menor que o da população de estimação.Os valores genéticos preditos ($X\hat{\beta}$) na população de validação são submetidos à análise de correlação com os valores fenotípicos observados nesta população (y). Como a população de validação não foi envolvida na estimação dos efeitos dos marcadores, os erros dos valores genéticos preditos e dos valores fenotípicos observados são independentes e toda correlação entre eles é de natureza predominantemente genética e equivale à capacidade da GWS em predizer os fenótipos. Após a verificação da acurácia da predição, o modelo é aplicado em uma população de seleção para predição dos valores genéticos dos indivíduos candidatos à seleção. Esta população consiste de indivíduos candidatos à seleção que foram genotipados, mas que não necessitam ter os seus fenótipos avaliados. Os indivíduos dessa população são então selecionados no programa de melhoramento com base em seus valores genéticos preditos.

No entanto, a utilização de milhares de marcadores no modelo pode conduzir a um sobreajuste (*overfitting*), ou seja, variações que na realidade são devidas ao erro experimental e podem indevidamente ser explicadas pelos efeitos dos marcadores. Segundo Meuwissen (2007), é necessária a utilização de validação cruzada para contornar esse problema. A técnica de validação cruzada consiste em dividir o conjunto de indivíduos em k subconjuntos. Desses k subconjuntos, um subconjunto é removido para ser utilizado na validação do modelo (amostra de validação) e os $k-1$ subconjuntos restantes (amostra de estimação) são utilizados no ajuste do modelo. O processo de validação cruzada é repetido k vezes, de modo que cada um dos k subconjuntos seja utilizado uma vez como amostra de validação. Como critério para validação do modelo, pode ser utilizado, por exemplo, a correlação entre os fenótipos observados e os valores genéticos preditos na amostra de validação. O resultado final desse processo é a média das correlações obtidas nas k análises.

2.4 Lasso Bayesiano

Tibshirani (1996) propôs um método de encolhimento (*shrinkage*) denominado Lasso (*Least Absolute Shrinkage and Selection Operator*) que minimiza a soma de quadrados residuais, restringindo a soma dos valores absolutos dos coeficientes de regressão $\sum_{j=1}^{p}|\beta_j| \leq t$ para $t \geq 0$, se a variável resposta e as variáveis regressoras estão padronizadas. Essa restrição permite que algumas estimativas dos coeficientes de regressão sejam exatamente zero, realizando simultaneamente um procedimento de *shrinkage* e seleção de modelos.

A estimativa do Lasso pode ser obtida por

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

em que $\lambda \geq 0$.

O estimador Lasso pode ser interpretado como uma estimativa da moda *a posteriori* quando os parâmetros de regressão têm *prioris Laplace* (Exponencial Dupla) independentes (PARK; CASELLA, 2008;TIBSHIRANI, 1996). Uma variável aleatória tem distribuição exponencial dupla com parâmetros μ e λ, isto é, $Y \sim ED(\mu,\lambda)$, se sua função densidade de probabilidade é dada por

$$f(y;\mu,\lambda)=\frac{\lambda}{2}e^{-\lambda|y-\mu|},\ -\infty<y<\infty,\ -\infty<\mu<\infty,\quad \lambda>0$$

Em que μ e λ são parâmetros de locação e escala, respectivamente. Na Figura 2, são apresentados os gráficos das funções densidade da distribuição exponencial dupla com parâmetros $\mu=0$ e $\lambda=1$ e da distribuição normal padrão.

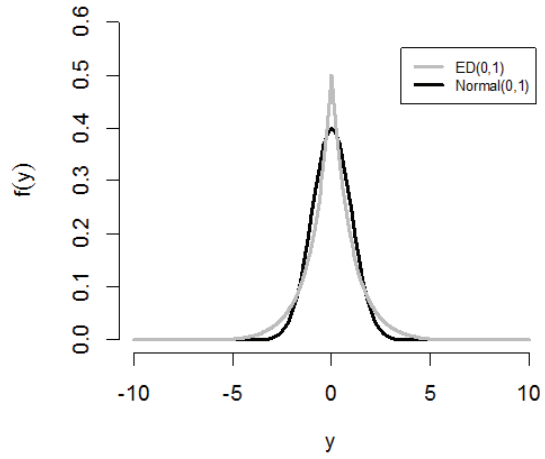

Figura 2 Gráficos das funções densidade das distribuições Exponencial Dupla e Normal

Observa-se que na distribuição exponencial dupla há uma maior probabilidade em torno do zero em comparação à distribuição normal com o mesmo fator de escala. Pode-se observar também que a probabilidade de ocorrer valores grandes (positivos ou negativos) é maior na distribuição exponencial dupla do que na distribuição normal. Dessa forma, a distribuição exponencial dupla pode ser utilizada como distribuição *a priori* para os efeitos dos marcadores de modo que cada coeficiente β_j tenha uma alta probabilidade de estar próximo de zero,mas ao mesmo tempo tenha chance de ter um grande efeito (YI; XU, 2008), ou seja,

$$\pi(\beta) = \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|}.$$

Além disso, a distribuição Exponencial Dupla pode ser expressa como uma mistura escalar de distribuições normais com variâncias com distribuições exponenciais independentes (ANDREWS; MALLOWS, 1974), isto é,

$$\frac{\lambda}{2}e^{-\lambda|\beta_j|} = \int_0^{\infty}\left[\frac{\exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right)}{\sqrt{2\pi\tau_j^2}}\right]\left[\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right)\right]d\tau_j^2 \qquad (2)$$

Em que β_j é o efeito desconhecido do *j*-ésimo marcador e τ_j^2 é a variância associada a β_j, à qual será atribuída uma distribuição *a priori*. Na equação (2), tem-se $\beta_j|\tau_j^2 \sim N(0, \tau_j^2)$ combinada com $\tau_j^2 \sim Exp(\lambda^2/2)$.

Motivados por essa variação da distribuição exponencial dupla, Park e Casella (2008) propuseram usá-la como distribuição *a priori* em um modelo hierárquico (Lasso Bayesiano).O amostrador de Gibbs é utilizado para obter as estimativas dos parâmetros de regressão $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ do modelo (1), e explora a representação da distribuição Laplace da equação (2). As distribuições a priori do Lasso Bayesiano são especificadas como (SUN; IBRAHIM; ZOU, 2010):

$$\pi(\mu) \propto 1,$$

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

$$\pi(\beta_j|\tau_j^2) \sim N(0, \tau_j^2) = \frac{1}{\sqrt{2\pi\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right),$$

$$\pi\left(\tau_j^2|\frac{\lambda^2}{2}\right) \sim Exp\left(\frac{\lambda^2}{2}\right) = \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right),$$

$$\pi\left(\frac{\lambda^2}{2}\right) \sim Gama(s,r) = \frac{r^2}{\Gamma(s)}\left(\frac{\lambda^2}{2}\right)^{s-1} \exp\left(-r\frac{\lambda^2}{2}\right),$$

Em que $j = 1, \dots, p$ indicam os p marcadores. Também é modelada a distribuição de $\lambda^2/2$ ao invés de λ, e na *priori* de $\lambda^2/2$, s e r são hiperparâmetros de forma e escala, respectivamente. Podem ser consideradas, ainda, duas opções de *prioris* para $\pi(\beta_j|\tau_j^2)$, 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 $(\mu, \beta, \sigma^2, \tau^2, \lambda \mid y)$ pode ser expressa como:

$$\pi(\mu, \beta, \sigma^2, \tau^2, \lambda \mid y) \propto$$

$$\propto \prod_{i=1}^{n} \pi(y_i|\mu, X, \beta, \sigma^2)\pi(\mu)\pi(\sigma^2) \prod_{j=1}^{p} \pi(\beta_j|\tau_j^2)\pi\left(\tau_j^2|\frac{\lambda^2}{2}\right)\pi\left(\frac{\lambda^2}{2}\right)$$

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

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. A partir dessa amostra são obtidas estimativas dos parâmetros de regressão $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$. Geralmente são utilizadas médias *a posteriori* para estimar β, cujas estimativas assumem valores próximos de zero, mas não exatamente zero.

2.5 *Tournament Screening*

Chen e Chen (2009) propuseram um método para seleção de modelos em que primeiramente é realizado um procedimento de redução de dimensionalidade para, somente depois, aplicar um método de seleção de modelos utilizando o conjunto reduzido de dados. O método de Chen e Chen (2009) se baseia em selecionar variáveis por meio de torneios de análises (*tournament screening*).

Seja S^1 o conjunto dos números inteiros de 1 a p. Seja S_0 o grupo (subconjunto) de S^1 correspondente aos componentes não nulos de β (coeficientes das variáveis causais), isto é, $\beta_j \neq 0 \Leftrightarrow j \in S_0$. Para qualquer grupo $S \in S^1$, seja $v(S)$ o número de elementos de S. Em particular, seja $v_0 = v(S_0)$. Seja $X(S)$ a submatriz correspondendo às colunas de X com índice em S, e por $\beta(S)$ os correspondentes componentes de β. A função de log-verossimilhança negativa penalizada é definida como:

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

Em que $p_\lambda(.)$ é uma função de penalidade regulada pelo parâmetro λ. A função de penalidade pode ser tomada como $p_\lambda(|\beta_j|) = \lambda|\beta|$, a penalidade L_1 usada no Lasso de Tibshirani (1996), ou ainda pode ser usada a função de penalidade SCAD usada por Fan e Li (2001). Ajustando o valor de λ, praticamente qualquer número de componentes de $\beta(S)$ podem ser estimados como zero.

Seja n_g o número de elementos de cada grupo, tal que $n_g < n$, e 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 com efeitos não nulos dentro do grupo, e ao mesmo tempo pequeno o suficiente para reduzir o número de variáveis eficientemente. Caso se tenha alguma noção do tamanho de v_0 (número total de variáveis com efeitos não nulos), k pode ser escolhido como $2v_0$ ou $3v_0$.

O procedimento de seleção de variáveis por meio de torneios, denominado pelos autores como *Tournament Screening* (TS) é apresentado de forma detalhada a seguir:

Estágio 1. Particiona-se S^1 em grupos aleatórios disjuntos de tamanhos aproximadamente iguais n_g de modo que $S^1 = S_{11} \cup ... \cup S_{1J_1}$, em que J_1 é o maior número inteiro tal que $n_g J_1 \leq p$. Para cada grupo S_{1j}, minimiza-se $l_p(\beta(S), \sigma^2|\lambda)$ ajustando λ para que sejam estimados apenas k componentes $\hat{\beta}(S_{1j})$ 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}^*, $j = 1, 2, \ldots, J_1$ para formar o grupo de todas as variáveis selecionadas no primeiro estágio S^2. O número de variáveis é reduzido para kJ_1 neste estágio.

Estágio 2. É repetido o processo do Estágio 1 com S^1 substituído por S^2. O número de variáveis é reduzido para kJ_2 em que J_2 é o maior número inteiro tal que $n_g J_2 \leq KJ_1$.

Próximos estágios. O processo acima continua até que o número de variáveis seja reduzido para k.

O procedimento TS descrito acima inicia com uma partição aleatória do conjunto de variáveis e, a redução da dimensão das variáveis é geralmente dependente da partição inicial. Contudo, segundo Chen e Chen (2009), a seleção de variáveis por torneios é capaz de selecionar as variáveis com efeitos não nulos e eliminar apenas variáveis com efeitos nulos. Fan e Lv (2008) se referem

à capacidade de uma metodologia de reter as variáveis com efeitos não nulos como *"sure screening property"*. Dessa forma, os conjuntos finais das variáveis selecionadas, resultantes de torneios com diferentes partições, devem diferir apenas quanto às variáveis com efeitos nulos.

Seleção final de variáveis. 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 estendido (*Extended Bayes Information Criterion* - EBIC).

3 MATERIAL E MÉTODOS

3.1 Dados

a) Genótipos dos SNPs

Os genótipos dos SNPs utilizados neste estudo são dados reais de bovinos, cedidos pela Embrapa Pecuária Sudeste. Foram genotipados 400 animais da raça Canchim com *chips* BovineHD BeadChip (Illumina Inc., San Diego, CA).

Primeiramente foi realizado o controle de qualidade dos dados em que foram realizadas filtragens de amostras (animais) e de SNPs. Foram eliminadas amostras com taxa de genotipagem (*call rate*) menor que 98%, com desvio na heterozigozidade em relação à média dentre as amostras maior que três desvios padrões, e *outliers* identificados por gráfico de escalonamento multidimensional (MDS). Foram eliminados SNPs com genótipos faltantes (não foi utilizado nenhum método de imputação),com frequência alélica menor que 3%,com genótipos fixos na população (100% de ocorrência de um único genótipo), e, foram eliminados marcadores redundantes (com genótipos idênticos para todas as amostras) de modo que foi mantido apenas um deles. A verificação de marcadores redundantes foi feita apenas para os adjacentes entre si devido à inviabilidade computacional de verificar todas as combinações possíveis de pares de marcadores em um conjunto de quase 700.000 marcadores. Após o controle de qualidade permaneceram no estudo 384 animais e 526.493 SNPs distribuídos por 32 cromossomos (Tabela 1).

Tabela 1 Quantidade de SNPs por cromossomo após o controle de qualidade

Cromossomo	Quantidade de SNPs
0	727
1	32.587
2	28.215
3	25.311
4	24.930
5	24.700
6	25.281
7	23.352
8	23.559
9	22.421
10	22.217
11	22.825
12	18.139
13	16.466
14	17.386
15	17.100
16	17.315
17	15.941
18	14.084
19	13.531
20	15.449
21	14.999
22	13.438
23	10.917
24	13.528
25	9.392
26	10.928
27	9.503
28	9.484
29	10.482
X	2.269
Y	17

b) Fenótipos reais

A característica fenotípica real estudada foi área de olho de lombo (AOL). Antes de realizar as análises os dados fenotípicos foram corrigidos para efeitos de grupos de contemporâneos (GC). Para formação dos grupos de contemporâneos, foram utilizados 987 bovinos, dentre os quais estão inclusos os 400 animais genotipados, em que os grupos de contemporâneos foram formados pelas variáveis: ano de nascimento, fazenda, grupo genético e sexo. Os fenótipos corrigidos são, portanto, os resíduos de um modelo de efeitos fixos da variável AOL em função das demais variáveis citadas. Para realização das análises com marcadores SNP, foram utilizados apenas os fenótipos corrigidos dos 384 animais genotipados que permaneceram após o processo de controle de qualidade dos dados. Na Tabela 2, são apresentadas as estatísticas descritivas dos dados originais e corrigidos para efeitos de grupos de contemporâneos da característica área de olho de lombo dos 384 animais. Na Figura 3, são apresentados os dados de fenótipos corrigidos de AOL e, na Figura 4, pode-se observar que estes dados apresentam distribuição aproximadamente normal.

Tabela 2 Estatísticas descritivas dos dados originais e corrigidos para efeitos de grupos de contemporâneos (GC) da característica área de olho de lombo (AOL), de 384 bovinos da raça Canchim

	Tipo de dados	
Estatística Descritiva	AOL (cm^2)	AOL corrigido para GC
Mínimo	19,91	-28,92
1º Quartil	40,14	-6,80
Mediana	47,16	-0,57
Média	47,49	0,27
3º Quartil	54,67	6,83
Máximo	75,30	30,67
Desvio padrão	10,15	9,19

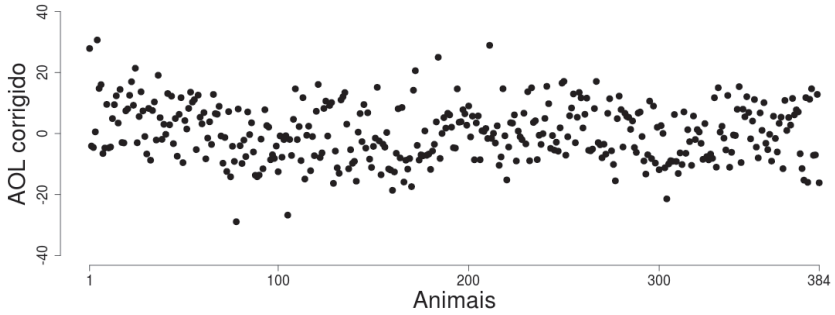

Figura 3 Dados de área de olho de lombo corrigidos para efeitos de grupos de contemporâneos de 384 bovinos da raça Canchim

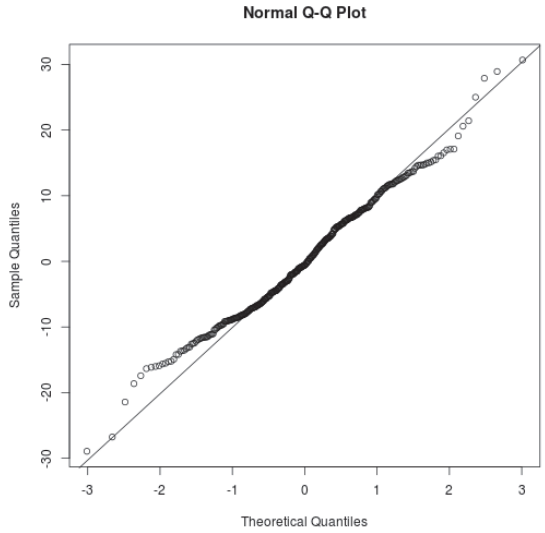

Figura 4 Normal QQ-Plot para dados de AOL corrigidos para efeitos de grupos de contemporâneos de 384 bovinos da raça Canchim

c) Efeitos simulados de SNPs

Para verificar se a metodologia de torneios proposta é capaz de selecionar os SNPs com efeitos não nulos, ou SNPs próximos destes, é necessário conhecer os verdadeiros valores dos efeitos dos SNPs. Além disso, com base nos verdadeiros efeitos e na matriz de genótipos dos SNPs podem-se obter os verdadeiros valores genéticos dos indivíduos, que são necessários para avaliar as metodologias quanto à capacidade de predição de valores genéticos.

Para simular os efeitos dos marcadores foi considerado o mapa real de SNPs dos 384 bovinos da raça Canchim. Como no estudo de simulação foram realizadas 100 repetições de cada tipo de análise, então, por questões de tempo e de custo computacional, o conjunto de marcadores utilizado no estudo de simulação foi composto por apenas uma parte dos SNPs originais, totalizando 11.812 SNPs, os quais foram provenientes dos cromossomos 20 ao 29 (Tabela 3), selecionados a cada dez SNPs.

Tabela 3 Quantidade de SNPs por cromossomo no conjunto de dados reduzido utilizado no estudo de simulação

Cromossomo	Quantidade de SNPs
20	1.545
21	1.500
22	1.344
23	1.092
24	1.353
25	939
26	1.093
27	950
28	948
29	1.048
Total	11.812

Foram simulados vetores de efeitos genéticos aditivos de SNPs (β) contendo uma determinada quantidade de SNPs com efeitos não nulos e o restante com efeitos nulos. Foram simulados vetores de efeitos dos SNPs considerando três diferentes cenários:

a) 48 SNPs com efeitos não nulos, distribuídos entre quatro cromossomos, próximos entre si no cromossomo (Figura 5);
b) 48 SNPs com efeitos não nulos, distribuídos entre oito cromossomos, um pouco mais dispersos entre si no cromossomo (Figura 6);
c) 250 SNPs com efeitos não nulos, distribuídos por todos os cromossomos, dispersos entre si no cromossomo (Figura 7).

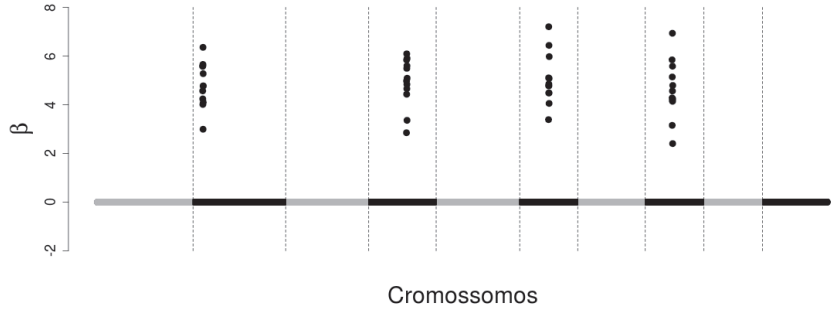

Figura 5 Efeitos simulados de 48 SNPs com efeitos não nulos, distribuídos entre 4 cromossomos, próximos entre si no cromossomo

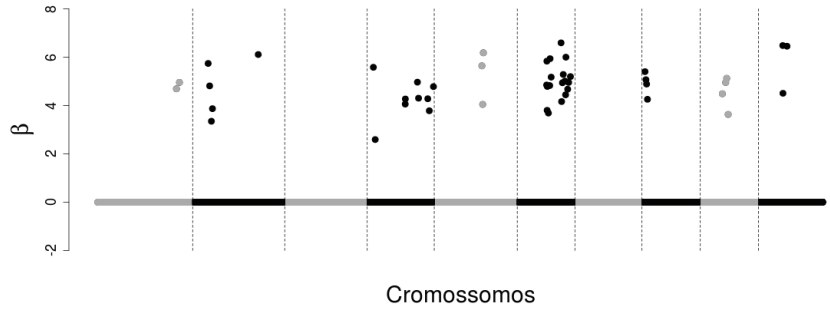

Figura 6 Efeitos simulados de 48 SNPs com efeitos não nulos, distribuídos entre oito cromossomos, um pouco mais dispersos entre si no cromossomo

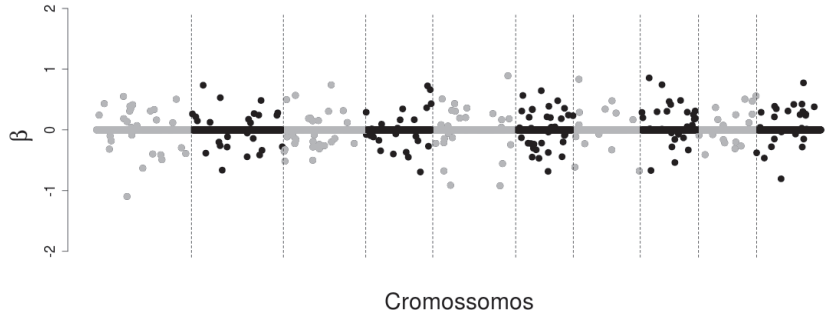

Figura 7 Efeitos simulados de 250 SNPs com efeitos não nulos, distribuídos por todos os cromossomos, dispersos entre si no cromossomo

d) Fenótipos Simulados

Para simular vetores de fenótipos dos animais foram utilizados os vetores de efeitos simulados dos SNPs (β) e a matriz de incidência (X) dos genótipos dos 11.812 SNPs, em que, a matriz de incidência é proveniente do

conjunto de genótipos reais e suas colunas são referentes aos SNPs selecionados para o estudo de simulação.

Primeiramente foram calculados os verdadeiros valores genéticos dos animais, $GBV = X\beta$. Nas Figuras 8, 9 e 10, podem ser observados os gráficos de dispersão dos GBVs obtidos a partir dos efeitos de marcadores simulados nas situações de 48 SNPs não nulos agrupados, 48 SNPs não nulos dispersos e 250 SNPs não nulos dispersos, respectivamente. De posse dos GBVs foram simulados vetores de fenótipos dos animais de acordo com a Equação (3):

$$y = 1\mu + X\beta + \varepsilon \tag{3}$$

Em que $\mu = 100$, e ε é o vetor de efeitos residuais que tem distribuição normal com média zero e variância σ_ε^2.

Foram simulados vetores de fenótipos considerando diferentes herdabilidades (h^2): 0,25, 0,5 e 1,0. Para isso, o vetor de resíduos ε foi simulado segundo uma distribuição normal com média zero e uma variância σ_ε^2 compatível com a herdabilidade desejada. Para obter σ_ε^2, foi considerada a Equação (4):

$$h^2 = \frac{\sigma_{GBV}^2}{\sigma_{GBV}^2 + \sigma_\varepsilon^2} \tag{4}$$

Em que σ_{GBV}^2 é a variância genética aditiva (variância de $X\beta$). Dessa forma, para $h^2 = 0{,}25$ foi considerado $\sigma_\varepsilon^2 = 3 \times \sigma_{GBV}^2$ e, para $h^2 = 0{,}50$ foi considerado $\sigma_\varepsilon^2 = \sigma_{GBV}^2$. Para simular y com herdabilidade $h^2 = 1{,}0$, não foi utilizado nenhum vetor de resíduos (ε) na Equação (3), ou seja, para $h^2 = 1{,}0$, foi considerado $y = 1\mu + X\beta$. A herdabilidade $h^2 = 1{,}0$ foi utilizada para avaliar as metodologias na ausência de erro.

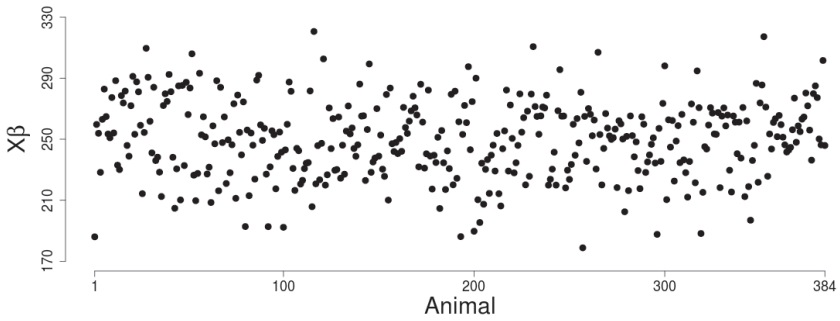

Figura 8 GBVs simulados em uma situação de 48 SNPs não nulos agrupados

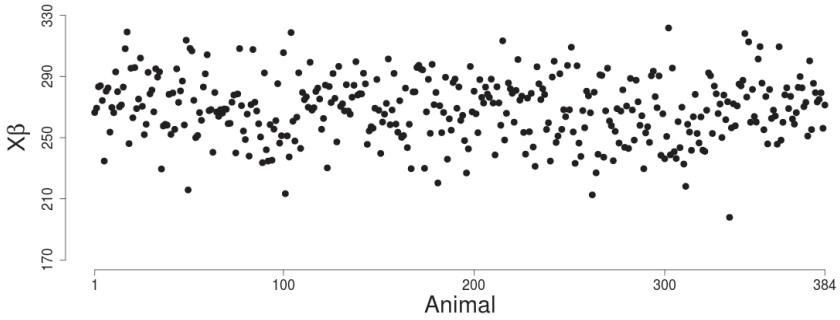

Figura 9 GBVs simulados em uma situação de 48 SNPs não nulos dispersos

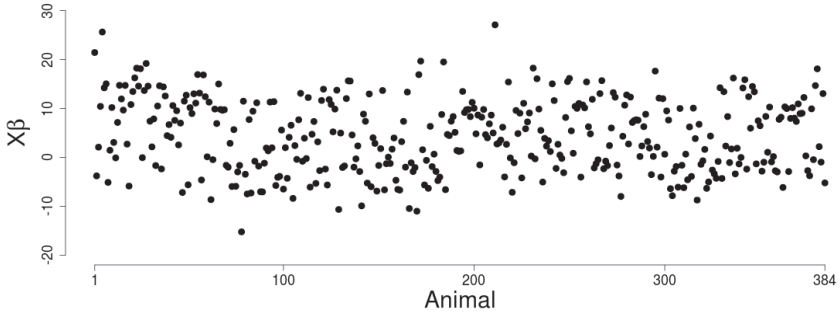

Figura 10 GBVs simulados em uma situação de 250 SNPs não nulos dispersos

Nas Figuras 11, 12 e 13, são apresentados os gráficos de dispersão $X\beta$ *versus* y para todos os cenários considerados.

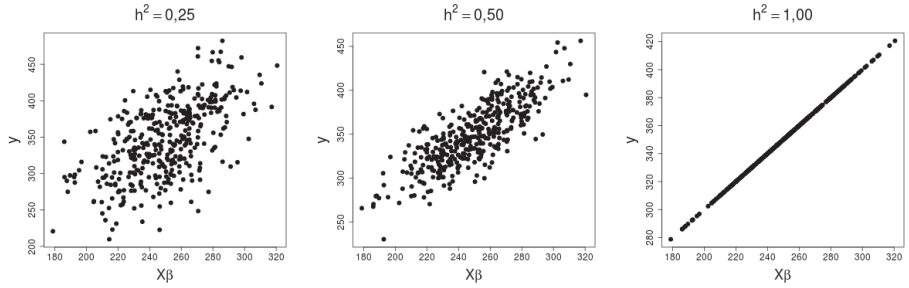

Figura 11 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 48 SNPs não nulos agrupados, considerando diferentes herdabilidades

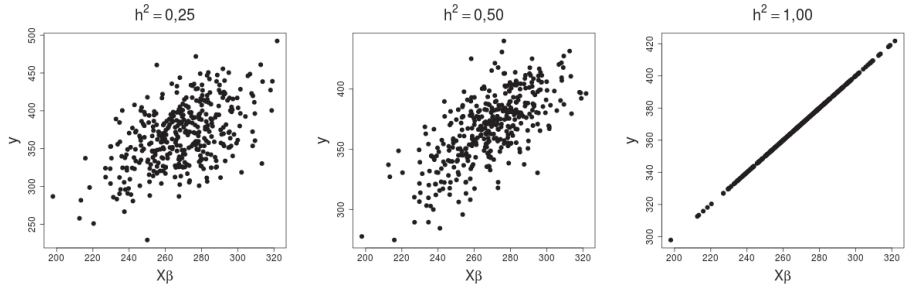

Figura 12 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 48 SNPs não nulos dispersos, considerando diferentes herdabilidades

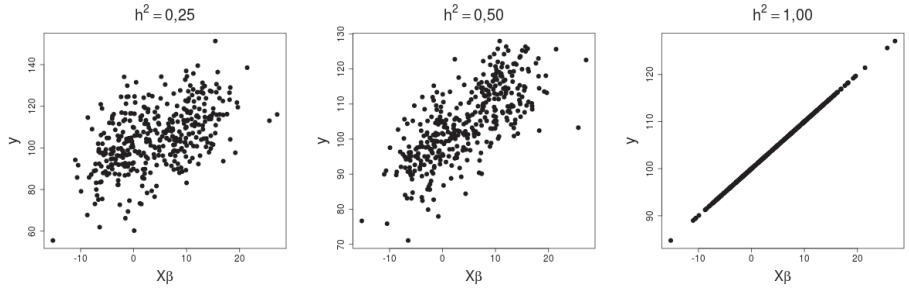

Figura 13 Gráficos de dispersão $X\beta$ *versus* y, simulados em uma situação de 250 SNPs não nulos dispersos, considerando diferentes herdabilidades

A importância de se utilizar fenótipos simulados é que, nesta situação, os verdadeiros GBVs dos indivíduos são conhecidos possibilitando verificar se a metodologia utilizada consegue predizer os GBVs corretamente.

3.2 Seleção de marcadores utilizando torneios

O primeiro passo para a realização do torneio é a partição do conjunto de todos os marcadores S^1 em grupos menores $S_{11} \cup S_{12} \cup ... \cup S_{1J_1}$, cada grupo com tamanho n_g, $n_g < p$, em que p é o número total de marcadores. A seguir são apresentadas as etapas que compõem a metodologia proposta para seleção de marcadores por torneios.

Etapa 1: É feita a divisão do número total de marcadores S^1 em grupos menores $S_{11} \cup S_{12} \cup ... \cup S_{1J_1}$, cada grupo com tamanho $n_g < p$. Para os marcadores de cada grupo S_{1j} é ajustado um modelo utilizando regressão linear múltipla clássica em que cada marcador é uma variável regressora. Obtidas as estimativas dos efeitos dos marcadores, $\beta(S_{1j})$, é eliminado o marcador cuja estimativa tem o maior valor-*p* ou, no caso de colinearidade perfeita, é eliminado (aleatoriamente) um dos marcadores cujo efeito não pôde ser estimado. Os marcadores selecionados (que não foram eliminados) de todos os grupos são reunidos 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.

Durante as etapas do torneio, são armazenadas as sequências de eliminação dos marcadores gerando uma classificação dos marcadores, de modo que os mais importantes são os que restaram ao final do torneio seguidos pelos marcadores eliminados na seguinte ordem: dos últimos marcadores eliminados no torneio até os que foram eliminados primeiro. Dessa forma, supondo que o número final de marcadores selecionados pelo torneio seja pequeno e que se queira selecionar um número maior de marcadores não é necessário realizar outro torneio, bastando agregar aos marcadores selecionados à quantidade adicional de marcadores provenientes dos marcadores excluídos de acordo com a classificação gerada pela ordem de eliminação dos mesmos. Vale ressaltar que esta forma de classificação não é individual para cada marcador, mas sim para conjuntos de marcadores que foram eliminados em uma mesma etapa do torneio.

Foram realizados torneios com grupos aleatórios e com grupos condicionados aos cromossomos. Para os grupos aleatórios, são sorteados aproximadamente n_g marcadores, independente de quais cromossomos eles sejam. Para os grupos condicionados aos cromossomos, cada grupo é formado por marcadores de quase todos os cromossomos, sorteados proporcionalmente ao número total de marcadores de cada cromossomo, formando grupos com tamanho aproximadamente n_g.

Para reduzir o tempo de análise pela metodologia de torneios, foram implementadas rotinas que possibilitam processamento paralelo, de modo que as análises em cada grupo foram executadas simultaneamente, uma em cada *core* do computador. Para implementação da programação paralela, foi utilizada a biblioteca *multicore* (URBANEK, 2014) do *software* R (R CORE TEAM, 2014). No APÊNDICE D, está disponível um código R reproduzível correspondente à metodologia de torneios com grupos aleatórios utilizando programação paralela.

3.3 Seleção de marcadores com maiores módulos das estimativas pelo Lasso Bayesiano

Foram ajustados modelos com todos os marcadores utilizando o Lasso Bayesiano e, determinadas quantidades demarcadores, com maiores módulos das estimativas dos efeitos no modelo completo, foram selecionados e utilizados para o ajuste de novos modelos para predição de valores genéticos.

3.4 Obtenção do modelo final

Ao final da etapa de seleção de marcadores, pelos torneios ou pelo Lasso Bayesiano, foram ajustados modelos aos marcadores selecionados utilizando o Lasso Bayesiano. O ajuste do modelo considerou uma cadeia de Markov de 4.000 iterações, "burn-in" de 2.000 e "thin" de 20 iterações, resultando em uma amostra de tamanho 100.

3.5 Verificação dos efeitos da multicolinearidade para SNPs próximos

Para verificar se marcadores próximos entre si são mais afetados pela multicolinearidade do que marcadores distantes foram feitas análises, utilizando regressão linear múltipla clássica, com os dados de genótipos dos SNPs do Cromossomo 1 do conjunto original de dados. Foram realizadas regressões lineares múltiplas considerando grupos de 100 SNPs tomados sequencialmente no cromossomo e considerando grupos de SNPs tomados aleatoriamente no cromossomo. A influência da multicolinearidade foi verificada pelas ocorrências de colinearidades perfeitas entre marcadores, observadas por meio de coeficientes que não puderam ser estimados nas análises.

3.6 Situações consideradas no estudo de simulação

Foram consideradas as seguintes situações no estudo de simulação:

a) Diferentes cenários para os efeitos dos SNPs: 48 SNPs com efeitos não nulos agrupados, 48 SNPs com efeitos não nulos dispersos, e 250 SNPs com efeitos não nulos dispersos;
b) Diferentes herdabilidades: $0{,}25$, $0{,}5$ e $1{,}0$;
c) Diferentes tamanho dos grupos: 25, 50 e 100;
d) Diferentes metodologias: Lasso Bayesiano, torneios com grupos aleatórios e torneios com grupos condicionados aos cromossomos.

3.7 Avaliação dos marcadores selecionados

Para avaliar as metodologias quanto à capacidade de selecionar marcadores corretamente, ou seja, selecionar os SNPs com efeitos não nulos (ou próximos destes), foram considerados apenas os 100 SNPs que permaneceram ao final de um torneio ou os 100 SNPs com maiores módulos das estimativas caso tenha sido utilizado o Lasso Bayesiano sem torneios. A capacidade de selecionar marcadores corretamente foi avaliada por meio da combinação de dois tipos de gráficos em que um deles representa os módulos dos efeitos simulados (para visualizar as posições dos SNPs com efeitos não nulos nos cromossomos) e o outro representa a frequência com que cada SNP foi selecionado em 100 análises.

3.8 Avaliação da predição de valores genéticos

Para avaliação da capacidade de predição dos valores genéticos dos indivíduos, foram calculadas as correlações entre GBVs preditos e simulados, $cor(X\hat{\beta}, X\beta)$, e as correlações entre GBVs preditos e fenótipos, $cor(X\hat{\beta}, y)$.

Foram ajustados modelos para predição dos GBVs considerando diferentes números de marcadores selecionados, de acordo com a classificação dos marcadores no torneio ou no Lasso Bayesiano. A classificação dos marcadores foi obtida pela ordem de eliminação destes no torneio ou pelo módulo das estimativas de seus efeitos no modelo completo (com todos os marcadores) no Lasso Bayesiano. De posse desta classificação, foram ajustados modelos para predição dos GBVs, utilizando o Lasso Bayesiano, para diferentes números de marcadores: 100, 250, 500, 1.000, 2.000, 4.000, 8.000 e 11.812, e, então, foram calculadas as correlações $cor(X\hat{\beta}, X\beta)$ e $cor(X\hat{\beta}, y)$.

Para avaliar a capacidade de predição dos GBVs para animais que não foram utilizados no ajuste do modelo, foram feitas análises utilizando um esquema de validação cruzada. Dessa forma, as correlações entre GBVs preditos e simulados e entre GBVs preditos e fenótipos foram obtidas considerando duas situações:

a) sem validação cruzada;
b) com validação cruzada.

A validação cruzada foi realizada dividindo-se o conjunto de dados dos 384 animais em oito subconjuntos de 48 animais. Em cada etapa da validação cruzada um subconjunto foi removido da análise (amostra de validação) para ser utilizado na validação da mesma e, o subconjunto maior, formado pelos outros sete (amostra de estimação), foi utilizado para estimação dos efeitos dos

marcadores. O modelo obtido na amostra de estimação foi aplicado na amostra de validação para predizer os GBVs dos indivíduos desta amostra e, foram então, calculadas as correlações $cor(X\hat{\beta}, y)$ e $cor(X\hat{\beta}, X\beta)$. O processo foi repetido oito vezes, tal que, em cada uma delas um dos oito subconjuntos foi utilizado como amostra de validação, e, os resultados finais foram as correlações médias das oito análises. Foram também calculadas as médias das correlações obtidas nas amostras de estimação.

3.9 Análises com fenótipos reais de área de olho de lombo (AOL)

Foram selecionados marcadores utilizando torneios com grupos aleatórios, torneios com grupos condicionados aos cromossomos e Lasso Bayesiano sem torneios para fenótipos reais da característica área de olho de lombo (AOL) corrigidos para efeitos de grupos de contemporâneos. Os torneios foram realizados com o tamanho de grupo que obteve melhor desempenho no estudo de simulação. Foram calculadas correlações entre GBVs preditos e fenótipos para as amostras de estimação e de validação em um esquema de validação cruzada.

Foram realizados 100 torneios com grupos aleatórios em que, em cada torneio, 100 marcadores foram selecionados e, foram utilizados intervalos de credibilidade HPD (Highest Posterior Density) de 95% para realizar um procedimento de seleção final de marcadores.

Foram ajustados modelos utilizando o Lasso Bayesiano aos marcadores selecionados (por HPDs) com maiores frequências em 100 análises. Os modelos foram ajustados variando-se a frequência mínima com que os marcadores utilizados foram selecionados nas 100 análises. Os ajustes dos modelos foram feitos utilizado um esquema de validação cruzada com oito grupos de 48

animais. Foram calculadas as correlações entre GBVs preditos e fenótipos para as amostras de estimação e validação.

4 RESULTADOS E DISCUSSÃO

4.1 Verificação da redução dos efeitos da multicolinearidade por meio da formação de grupos de marcadores

Com o objetivo de verificar se a formação de grupos aleatórios (ou condicionados à estrutura de cromossomos) nos torneios realmente seria uma forma de contornar os efeitos da multicolinearidade, foram realizadas regressões lineares múltiplas, com grupos de 100 marcadores considerando duas situações:

a) SNPs tomados sequencialmente no cromossomo e;
b) SNPs tomados aleatoriamente no cromossomo.

Na Tabela 4, são apresentadas medidas descritivas dos coeficientes de regressão e de seus erros padrões e, o número de coeficientes que não puderam ser estimados devido às colinearidades perfeitas. Observa-se que nas análises com SNPs sequenciais, a multicolinearidade é tão alta que ocorreram colinearidades perfeitas impedindo que alguns coeficientes pudessem ser estimados. Nas análises com grupos aleatórios, todos os coeficientes puderam ser estimados e, além disso, os erros padrões das estimativas foram pequenos indicando não haver presença de forte multicolinearidade nestes grupos, uma vez que um dos efeitos da multicolinearidade é produzir erros padrões grandes.

Como em todas as análises com SNPs tomados sequencialmente, a multicolinearidade foi tão forte que houve colinearidades perfeitas, fazendo com que alguns coeficientes não pudessem ser estimados (Tabela 4), não foi utilizado nenhum critério adicional para verificação da multicolinearidade.

Na Figura 37 (APÊNDICE - A), é apresentado o resultado parcial de uma análise utilizando 100 SNPs tomados sequencialmente no cromossomo, em

que se pode observar mais claramente os efeitos da multicolinearidade sobre uma análise de regressão linear múltipla clássica quando são utilizados SNPs próximos.

Tabela 4 Medidas descritivas dos coeficientes de regressão e de seus erros padrões obtidos em regressões lineares múltiplas, com grupos de SNPs próximos (localizados sequencialmente) no cromossomo 1, grupos formados por SNPs tomados aleatoriamente no cromossomo 1, e quantidade de coeficientes que não puderam ser estimados devido às colinearidades perfeitas

SNPs	Coeficiente			Erro padrão			Casos não estimados
	Mínimo	Média	Máximo	Mínimo	Média	Máximo	
1 a 100	−30,89	1,14	46,91	1,8	9,95	34,95	25
101 a 200	−71,39	0,47	55,04	1,61	7,11	31,48	23
201 a 300	−86,25	0,33	75,19	1,87	12,02	53,77	28
301 a 400	−64,31	1,24	62,93	1,36	22,88	217,9	22
401 a 500	−44,03	1,13	109,2	1,82	19,21	117,8	29
Aleatórios	−2,83	0,42	5,58	0,59	0,85	1,71	0
Aleatórios	−1,68	0,39	2,65	0,6	0,86	1,59	0
Aleatórios	−1,71	0,33	3,31	0,58	0,79	1,49	0
Aleatórios	−1,94	0,37	3,49	0,56	0,91	3,94	0
Aleatórios	−2,02	0,43	4,47	0,54	0,82	2,58	0

Verificou-se, portanto, que a formação de grupos aleatórios é uma forma de amenizar os efeitos da multicolinearidade em análises de regressão linear múltipla clássica utilizadas nos grupos de um torneio. No entanto, apesar de constatada a influência da multicolinearidade sobre as análises nos grupos, é necessário, ainda, verificar se ela tem influência sobre a seleção dos marcadores e sobre a predição dos valores genéticos dos indivíduos.

4.2 Influência do tamanho dos grupos na predição dos valores genéticos

Para verificara influência do tamanho dos grupos na predição dos valores genéticos, foram feitos gráficos das correlações médias (ou seja, médias das correlações obtidas em 100 análises) entre GBVs preditos ($X\hat{\beta}$) esimulados ($X\beta$), em que os GBVs preditos foram obtidos a partir de SNPs selecionados em torneios com diferentes tamanhos de grupos.

Na Figura 14, são apresentadas as correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados à estrutura de cromossomos, considerando três tamanhos de grupos (25, 50 e 100), oito diferentes números de SNPs selecionados (100, 250, 500, 1.000, 2.000, 4.000, 8.000 e 11.812) e três herdabilidades (0,25, 0,5 e 1,0), em um cenário de 250 SNPs com efeitos não nulos dispersos.

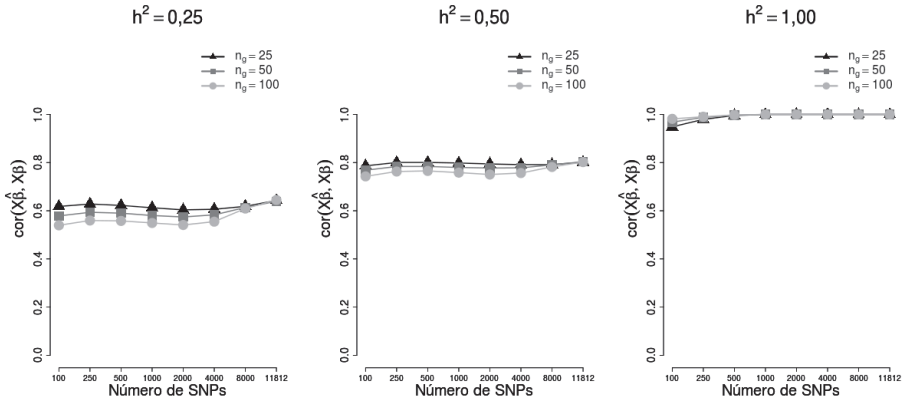

Figura 14 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Pode-se observar pela Figura 14 que, para a herdabilidade $h^2 = 0{,}25$, as correlações entre os GBVs preditos e simulados, $cor(X\hat{\beta}, X\beta)$, foram um pouco maiores para torneios com grupos de tamanho 25, enquanto para herdabilidades maiores a influência do tamanho dos grupos diminuiu. Observa-se também que a influência do tamanho dos grupos depende do número de SNPs selecionados, pois, quando são selecionados poucos SNPs (próximo de 100), a influência do tamanho dos grupos é maior do que quando são selecionados muitos SNPs (mais de 4.000). Esta diminuição da influência do tamanho dos grupos com o aumento dos SNPs selecionados se deve ao fato de que selecionar muitos SNPs equivale a eliminar poucos SNPs pelos torneios, ou seja, para números grandes de SNPs selecionados, os torneios foram subutilizados e, por isso o tamanho dos grupos não influencia muito. Note ainda que quanto mais aumenta o número de SNPs selecionados mais próximas as correlações ficam das obtidas no modelo completo (modelo com 11.812 SNPs).

Nas Figuras 38 a 42 (APÊNDICE B), são apresentadas as correlações entre GBVs preditos e simulados para diferentes tamanhos de grupos contemplando as demais situações consideradas no estudo como diferentes tipos de formação dos grupos no torneio (aleatórios e condicionados aos cromossomos) e diferentes tipos de efeitos simulados dos SNPs. Podem-se observar nestas figuras comportamentos semelhantes ao observado na Figura 14.

Também foram feitos gráficos das correlações médias entre GBVs preditos ($X\hat{\beta}$) e fenótipos (y) para comparar o efeito de diferentes tamanhos de grupos. Na Figura 15, são apresentadas as correlações entre $X\hat{\beta}$ e y para as mesmas situações consideradas na Figura 14. Observa-se que quando se considera as correlações entre $X\hat{\beta}$ e y (Figura 15) ocorre o inverso do que quando se considera as correlações entre $X\hat{\beta}$ e $X\beta$ (Figura 14), ou seja, tamanhos de grupos maiores proporcionam maiores correlações.

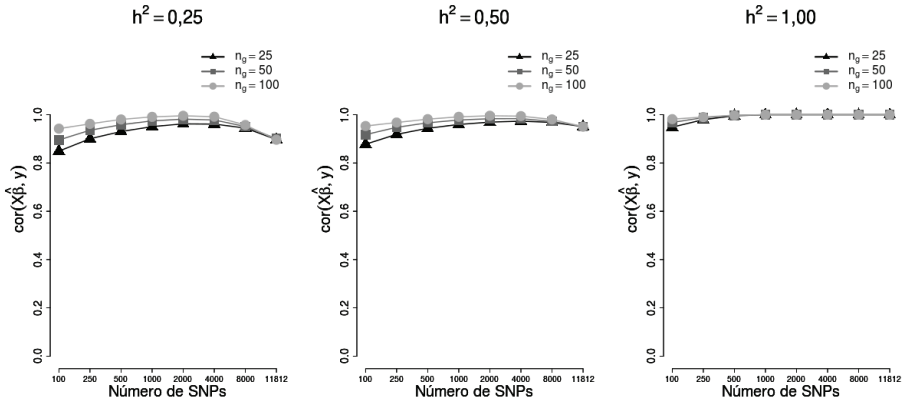

Figura 15 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Nas Figuras 43 a 47 (APÊNDICE B), são apresentadas as correlações entre GBVs preditos e fenótipos para diferentes tamanhos de grupos contemplando as demais situações consideradas no estudo e, podem-se observar por estas figuras comportamentos semelhantes ao observado na Figura 15.

Os fenótipos (y), no entanto, são compostos por valores genéticos adicionados de erros, assim, como o objetivo é predizer valores genéticos ($X\beta$), foram considerados tamanhos de grupos que proporcionam maiores correlações entre $X\hat{\beta}$ e $X\beta$ e, como tamanhos de grupos menores proporcionaram maiores correlações entre GBVs preditos e simulados, os torneios cujos resultados são apresentados nas próximas seções utilizaram grupos de tamanho 25. Nas Tabelas 7 a 10 (APÊNDICE C), pode-se observar correlações para outros tamanhos de grupos além de 25. Nestas tabelas são apresentadas as correlações entre GBVs preditos e fenótipos e entre GBVs preditos e simulados para amostras de estimação e validação de um esquema de validação cruzada considerando torneios com grupos aleatórios e torneios com grupos condicionados aos cromossomos em um cenário de 250 SNPs com efeitos não nulos dispersos. Pode-se observar que, mesmo utilizando validação cruzada, tamanhos de grupos menores proporcionaram maiores correlações entre GBVs preditos e simulados na maioria das situações e que tamanhos de grupos maiores proporcionaram maiores correlações entre GBVs preditos e fenótipos na maioria das situações.

4.3 Avaliação da capacidade de predição de valores genéticos sem utilizar validação cruzada considerando torneios com grupos de 25 marcadores

Nas Figuras 16 e 17, são apresentadas as correlações médias (médias das correlações em 100 análises) entre GBVs preditos e simulados e entre GBVs preditos e fenótipos, respectivamente, para diferentes metodologias,

considerando diferentes números de marcadores selecionados e diferentes herdabilidades para um cenário de 48 SNPs com efeitos não nulos agrupados.

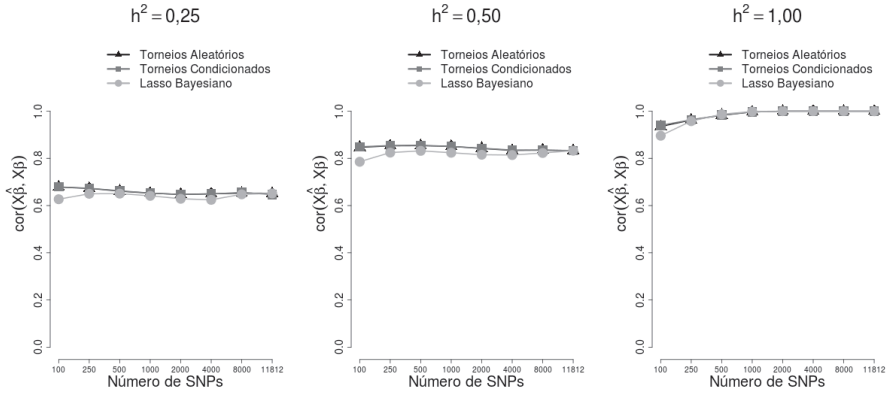

Figura 16 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

Observa-se que, para números pequenos de marcadores selecionados (próximo de 100), tanto as correlações entre GBVs preditos e simulados quanto entre GBVs preditos e fenótipos foram maiores utilizando torneios do que utilizando o Lasso Bayesiano.

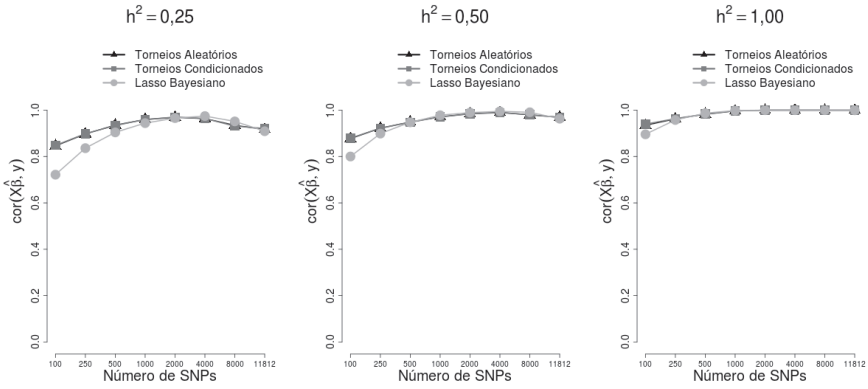

Figura 17 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

Pode-se observar também que não houve grandes diferenças entre as correlações obtidas de torneios com grupos aleatórios e torneios com grupos condicionados aos cromossomos. Isto sugere que tanto os torneios com grupos aleatórios quanto os torneios com grupos condicionados aos cromossomos devem estar selecionando conjuntos semelhantes de marcadores.

Nas Figuras 48 a 51 (APÊNDICE B), estão representadas as correlações entre GBVs preditos e simulados e entre GBVs preditos e fenótipos comparando diferentes metodologias para as demais situações consideradas no estudo de simulação.

4.4 Avaliação da capacidade de predição de valores genéticos utilizando validação cruzada considerando torneios com grupos de 25 marcadores

Nas Figuras 18 a 20, são apresentadas as correlações médias (médias das correlações em 100 análises) entre GBVs preditos e simulados, $cor\left(X\hat{\beta}, X\beta\right)$, para amostras de estimação em um esquema de validação cruzada, comparando

diferentes metodologias. De um modo geral, para as amostras de estimação, as correlações entre GBVs preditos e simulados obtiveram resultados muito próximos para todas as metodologias na maioria das situações consideradas no estudo de simulação. Nota-se uma pequena diferençadas correlações utilizando torneios para as correlações utilizando Lasso Bayesiano no cenário de 250 SNPs com efeitos não nulos dispersos (Figura 20). Para a herdabilidade 0,25 e com mais de 1.000 marcadores selecionados, o Lasso Bayesiano obteve correlações um pouco maiores que os torneios, para a herdabilidade 0,5 e com mais de 500 marcadores, os torneios obtiveram correlações um pouco maiores que o Lasso Bayesiano e, para a herdabilidade 1,0 e com 100 marcadores, os torneios obtiveram correlações maiores que o Lasso Bayesiano.

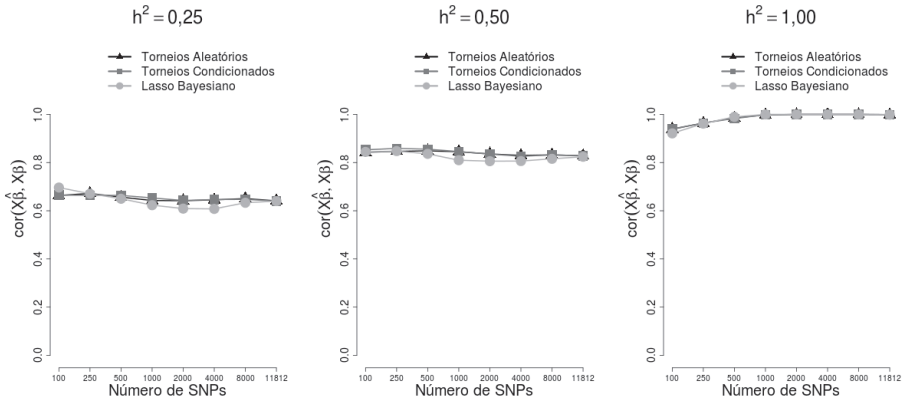

Figura 18 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

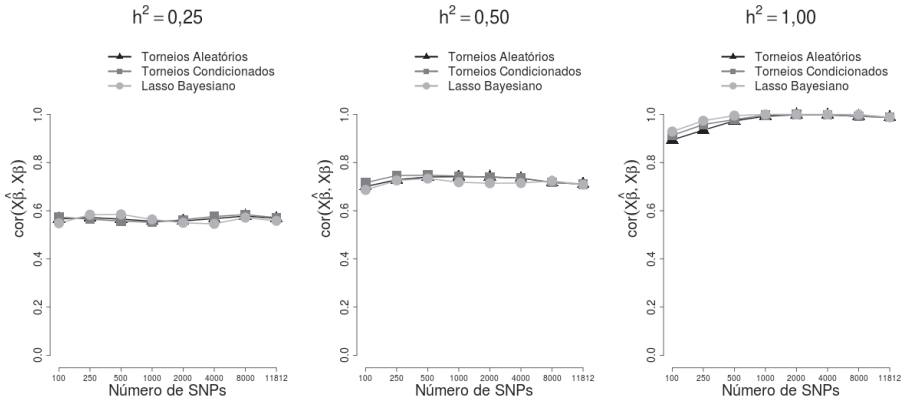

Figura 19 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

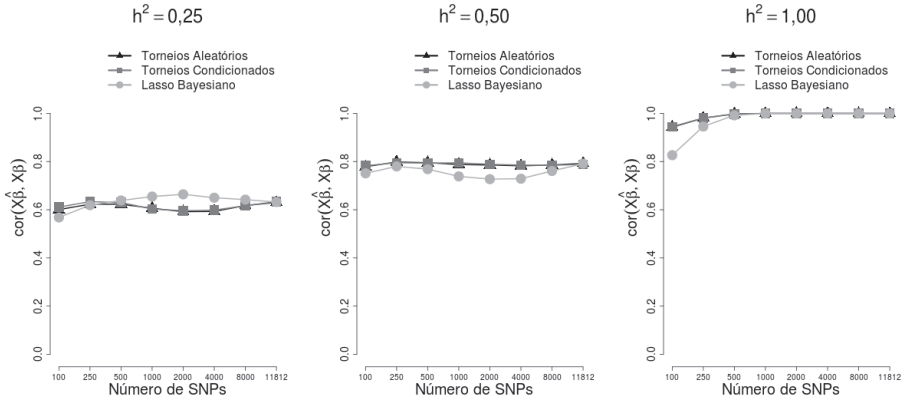

Figura 20 Correlações médias entre GBVs preditos e simulados em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Nas Figuras 21 a 23, são apresentadas as correlações médias entre GBVs preditos e fenótipos, $cor(X\hat{\beta}, y)$, para amostras de estimação em um esquema

de validação cruzada, comparando diferentes metodologias. Nota-se uma pequena diferença das correlações utilizando torneios para as correlações utilizando Lasso Bayesiano para 100 marcadores selecionados com herdabilidade 0,25 nos cenário de 48 SNPs com efeitos não nulos agrupados (Figura 21) e 48 SNPs com efeitos não nulos dispersos (Figura 22). Para o cenário de 250 SNPs com efeitos não nulos dispersos (Figura 23), nota-se uma diferença mais acentuada entre as correlações utilizando torneios e as correlações utilizando o Lasso Bayesiano para as herdabilidades 0,25 e 1,0. No entanto, apesar de as correlações entre GBVs preditos e fenótipos para a herdabilidade 0,25 no cenário de 250 SNPs com efeitos não nulos dispersos terem sido expressivamente maiores para os torneios do que para o Lasso Bayesiano, não se observou tais diferenças para o mesmo cenário e mesma herdabilidade com relação às correlações entre GBVs preditos e simulados (Figura 20).

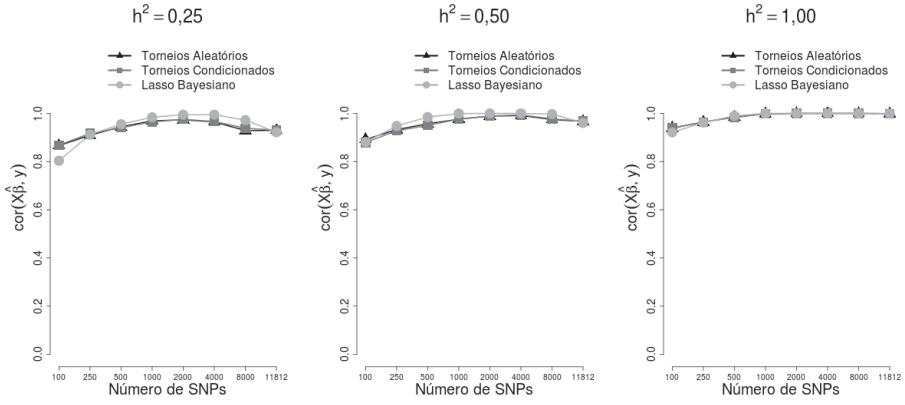

Figura 21 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada,para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

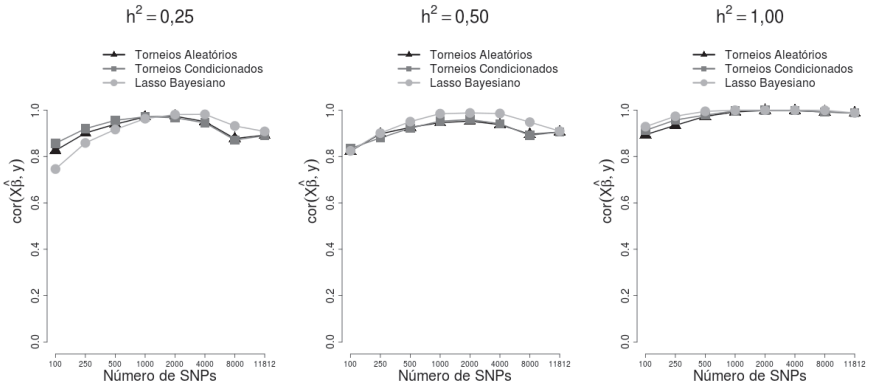

Figura 22 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

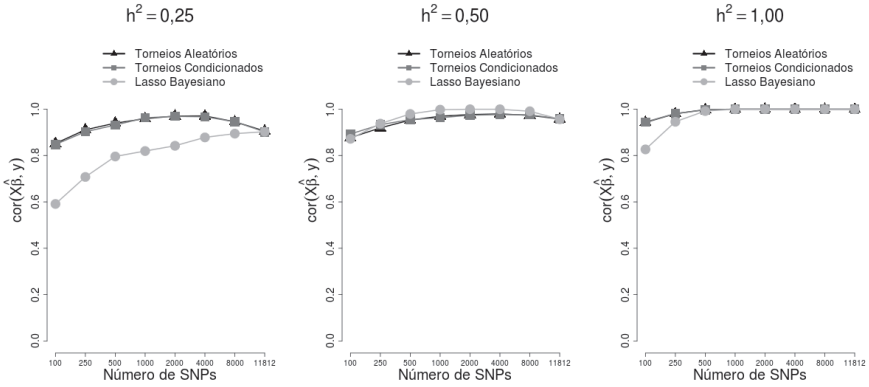

Figura 23 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Nas Figuras 24 a 26, são apresentadas as correlações entre GBVs preditos e simulados, $cor(X\hat{\beta}, X\beta)$, para amostras de validação em um esquema de validação cruzada, comparando diferentes metodologias. Nos cenários de 48

SNPs com efeitos não nulos agrupados (Figura 24) e 48 SNPs com efeitos não nulos dispersos (Figura 25) para a herdabilidade 0,25 não houve grandes diferenças entre as correlações obtidas pelas diferentes metodologias, para a herdabilidade 0,50 as correlações quando se utiliza de 4.000 marcadores foram um pouco maiores com o Lasso Bayesiano. Para a herdabilidade 1,0, as correlações no cenário de 48 SNPs com efeitos não nulos agrupados (Figura 24) foram maiores com o Lasso Bayesiano quando se utiliza mais de 4.000 marcadores e, para o cenário de 48 SNPs com efeitos não nulos dispersos (Figura 25) as correlações foram maiores com o Lasso Bayesiano para todos os números de SNPs selecionados.

Para o cenário de 250 SNPs com efeitos não nulos dispersos (Figura 26), observa-se que as correlações obtidas pelos torneios são expressivamente maiores que as correlações obtidas pelo Lasso Bayesiano para as herdabilidades 0,25 e 1,0 para quase todos os números de SNPs selecionados, já para a herdabilidade 0,5, praticamente não houve diferença entre as metodologias.

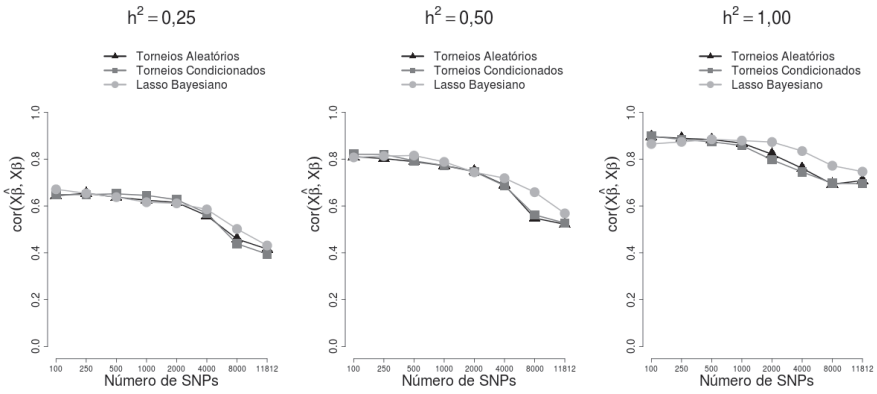

Figura 24 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

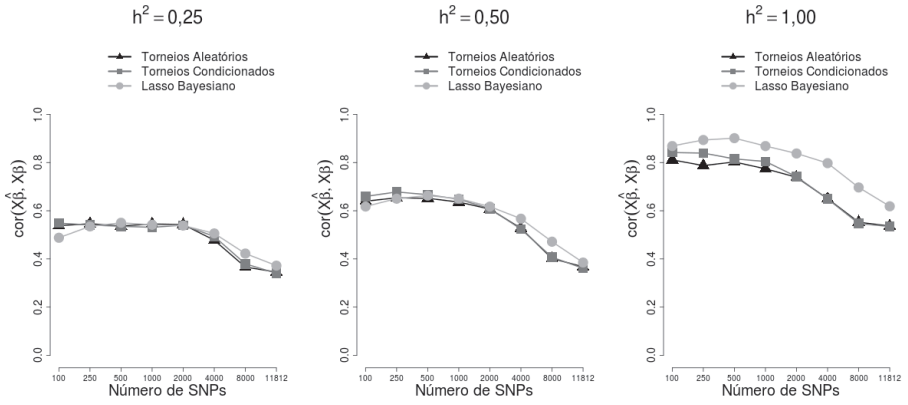

Figura 25 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

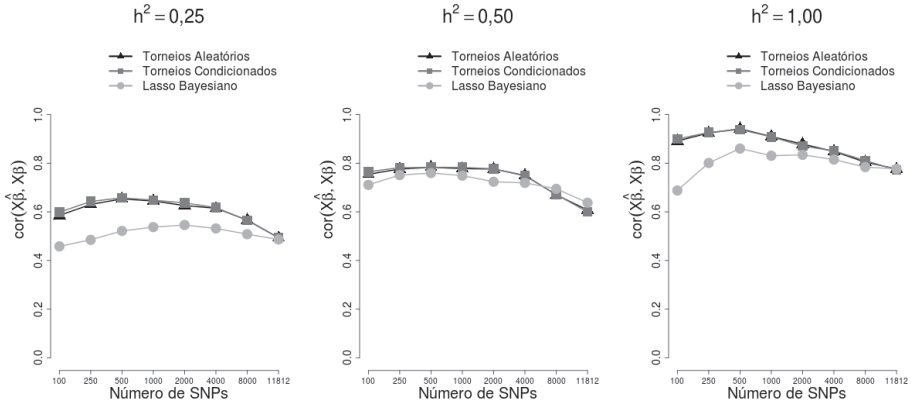

Figura 26 Correlações médias entre GBVs preditos e simulados em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Nas Figuras 27 a 29, são apresentadas as correlações entre GBVs preditos e fenótipos, $cor(X\hat{\beta}, y)$, para amostras de validação em um esquema de validação cruzada, comparando diferentes metodologias.

No cenário de 48 SNPs com efeitos não nulos agrupados (Figura 27), observa-se que para a herdabilidade 0,25 as correlações obtidas nos torneios são maiores que as correlações obtidas no Lasso Bayesiano quando são selecionados poucos marcadores (próximo de 100) e, com o aumento do número de marcadores selecionados as correlações obtidas nos torneios diminuem rapidamente sendo ultrapassadas pelas correlações obtidas no Lasso Bayesiano. No cenário de 48 SNPs com efeitos não nulos dispersos (Figura 28) para a herdabilidade 0,25, as correlações obtidas nos torneios são maiores que as correlações obtidas no Lasso Bayesiano quando são selecionados até aproximadamente 2.000 marcadores e, a partir daí, as correlações obtidas nos torneios diminuem rapidamente sendo ultrapassadas pelas correlações obtidas no Lasso Bayesiano. Para as herdabilidades 0,5 e 1,0, para os mesmos cenários mencionados, o Lasso Bayesiano obteve correlações superiores às obtidas nos torneios para a maioria dos números de marcadores selecionados (Figuras 27 e 28).

Para o cenário de 250 SNPs com efeitos não nulos dispersos (Figura 29) observa-se que, para a herdabilidade 0,25, houve uma diferença muito grande entre as correlações obtidas nos torneios e as correlações obtidas no Lasso Bayesiano para praticamente todos os números de marcadores selecionados, para a herdabilidade 1,0, também houve uma diferença expressiva entre as correlações obtidas nos torneios e as correlações obtidas no Lasso Bayesiano, enquanto para a herdabilidade 0,5, pratiamente não houve diferenças entre as correlações obtidas nas diferentes metodologias.

É importante notar que apesar de as correlações entre GBVs preditos e fenótipos terem sido muito maiores nos torneios do que no Lasso Bayesiano

para a herdabilidade 0,25 no cenário de 250 SNPs com efeitos não nulos dispersos (Figura 29), essa diferença não foi tão grande para as correlações entre GBVs preditos e simulados considerando a mesma herdabilidade e o mesmo cenário dos efeitos dos SNPs, como se pode observar na Figura 26. Além disso, as correlações entre GBVs preditos e simulados são uma medida direta da capacidade de predição dos valores genéticos, enquanto as correlações entre GBVs preditos e fenótipos são uma medida indireta da capacidade de predição dos valores genéticos.

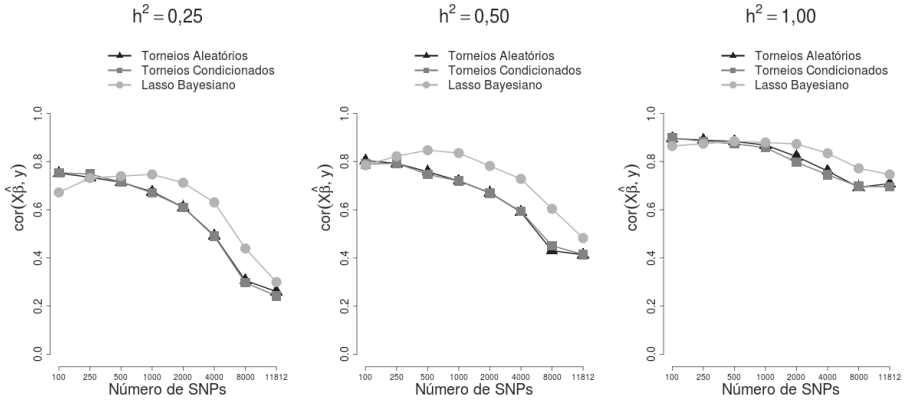

Figura 27 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

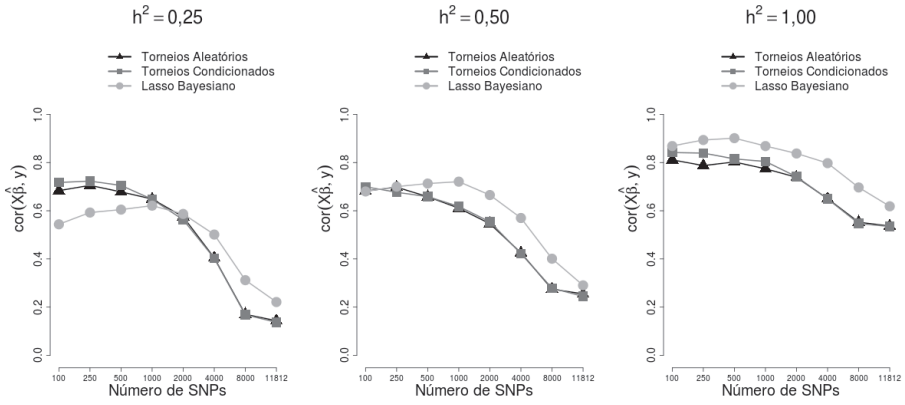

Figura 28 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

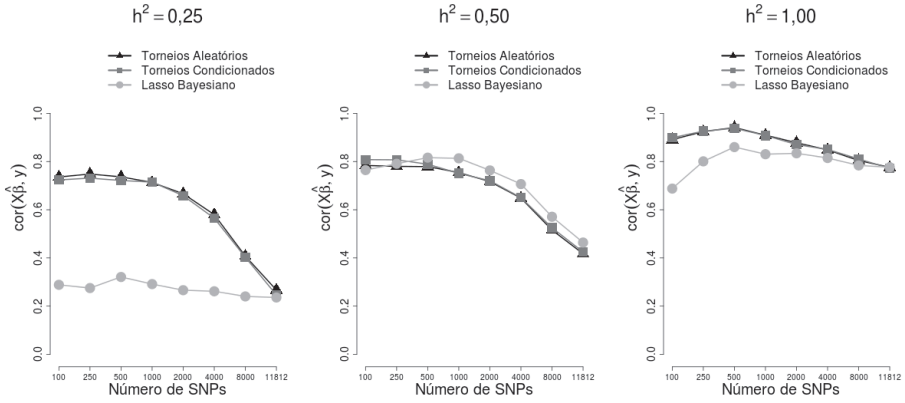

Figura 29 Correlações médias entre GBVs preditos e fenótipos em amostras de validação, num esquema de validação cruzada, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

Nas Tabelas 7 a 10 (APÊNDICE C), podem-se observar correlações para outros tamanhos de grupos dos torneios e, nas Tabelas 11 e 12, podem-se

observar correlações para o Lasso Bayesiano. Nessas tabelas são apresentadas as correlações entre GBVs preditos e fenótipos e entre GBVs preditos e simulados para amostras de estimação e validação de um esquema de validação cruzada considerando diferentes metodologias (torneios e Lasso Bayesiano) em um cenário de 250 SNPs com efeitos não nulos dispersos.

4.5 Seleção de marcadores considerando torneios com grupos de 25 marcadores

A capacidade de selecionar marcadores corretamente foi avaliada por meio da combinação de dois tipos de gráficos em que um deles representa os módulos dos efeitos simulados dos SNPs e o outro representa a frequência com que cada SNP foi selecionado em 100 análises para cada metodologia. Nas Figuras 30 a 32, podem-se observar os efeitos simulados dos marcadores para um cenário de 48 SNPs com efeitos não nulos agrupados e as frequências com que cada marcador foi selecionado em 100 análises de cada metodologia para as herdabilidades 0,25, 0,5 e 1,0, respectivamente.

Pode-se observar que, para a herdabilidade 0,25 (Figura 30), as frequências dos marcadores selecionados nas posições dos verdadeiros SNPs com efeitos não nulos ou próximos a estes estão muito próximas de 100 para os torneios (C e D), enquanto para o Lasso Bayesiano (B) essas frequências atingem no máximo 70, ou seja, para a maioria dos SNPs com efeitos não nulos as metodologias de torneios foram capazes de selecioná-los em quase todas as 100 repetições dessas análises, já o Lasso Bayesiano conseguiu selecioná-los em, no máximo, 70 análises. Observa-se ainda que, para a herdabilidade 0,25, todas as metodologias tiveram dificuldades em selecionar marcadores do último cromossomo que contém SNPs com efeitos não nulos, pois as frequências dos

marcadores selecionados referentes aos SNPs não nulos deste cromossomo foram relativamente baixas para todas as metodologias (próximas de 50).

Para as herdabilidades 0,5 (Figura 31) e 1,0 (Figura 32), observa-se que os torneios foram capazes de selecionar a maioria dos verdadeiros SNPs com efeitos não nulos na maioria das 100 análises (frequência próxima de 100), enquanto o Lasso Bayesiano selecionou SNPs com efeitos não nulos com frequência bem menor que os torneios.

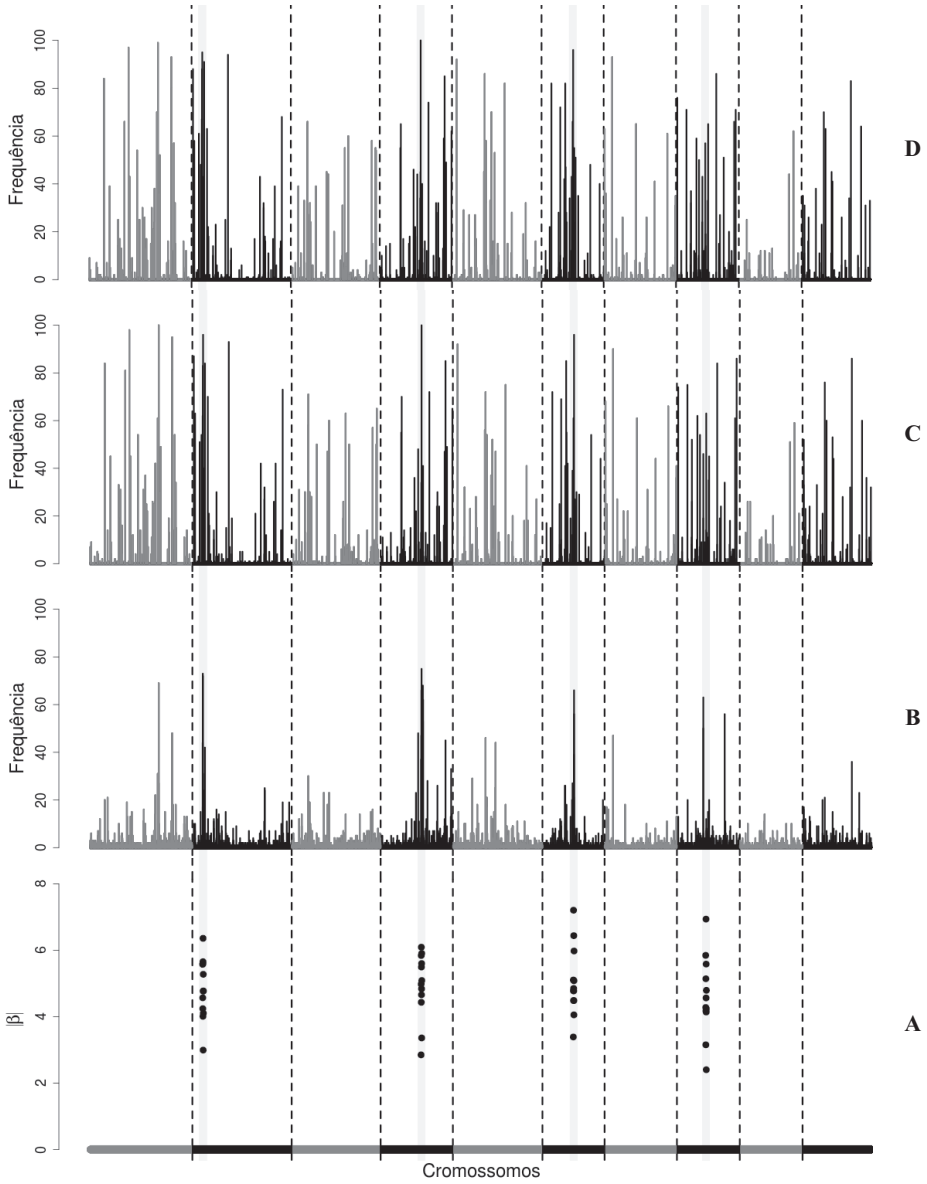

Figura 30 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 0,25

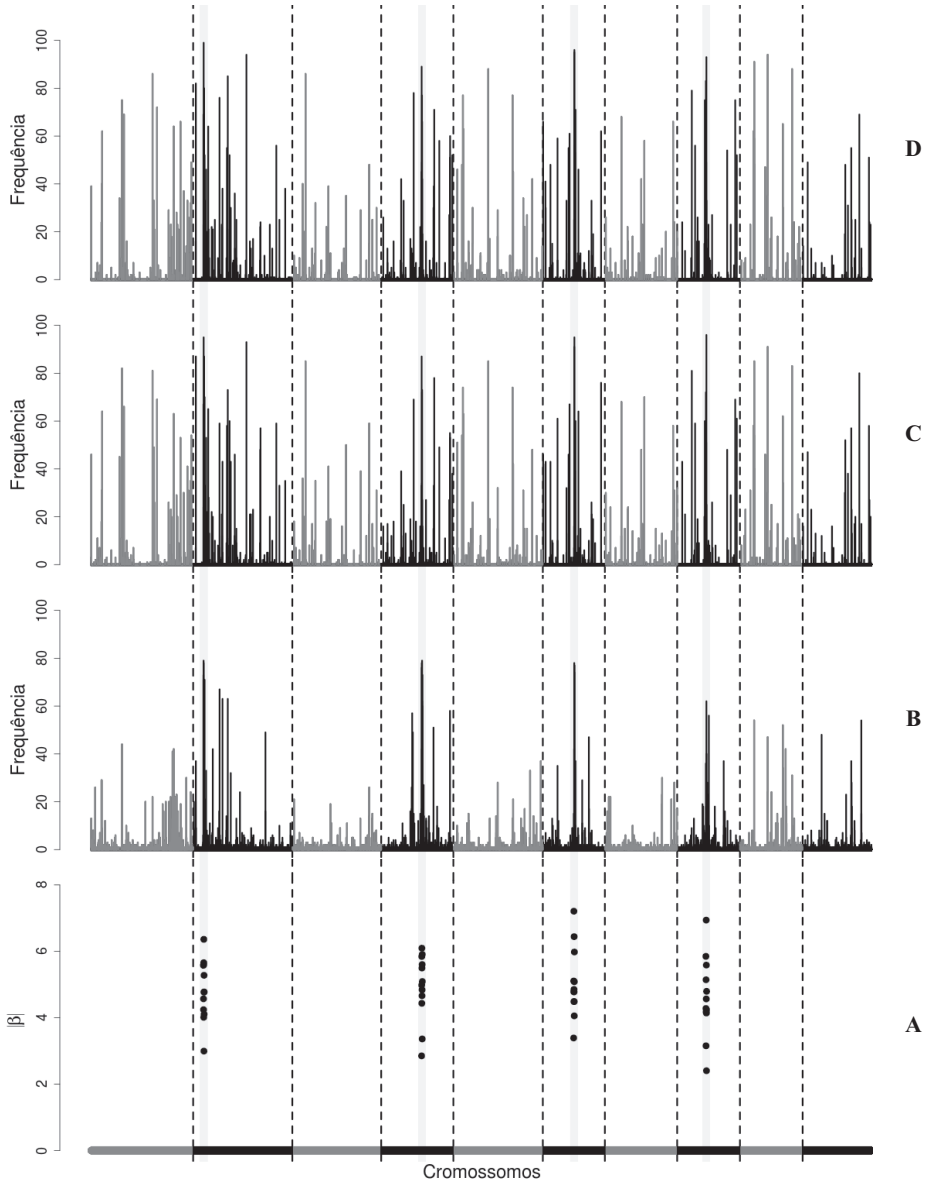

Figura 31 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 0,5

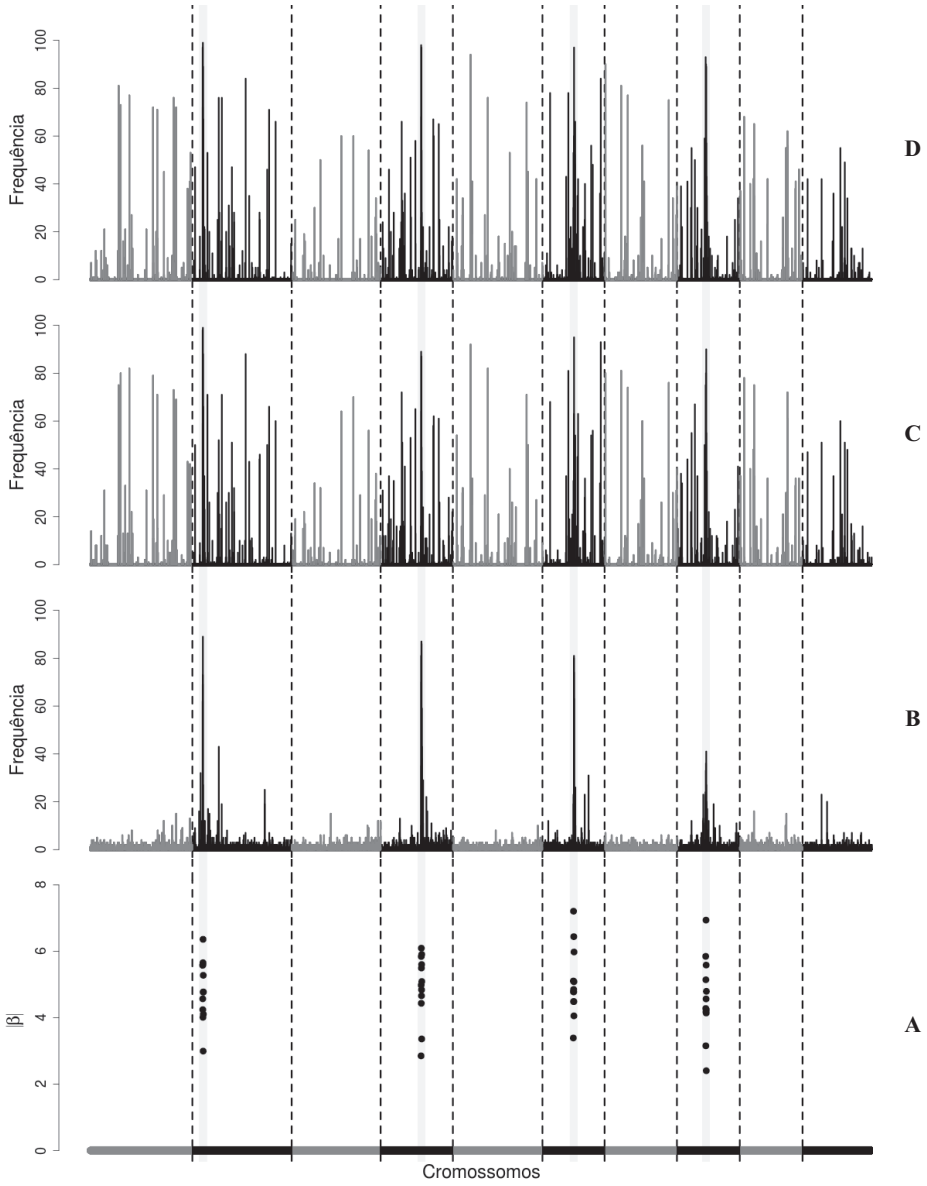

Figura 32 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos agrupados e com herdabilidade 1,0

Para todas as herdabilidades consideradas, as metodologias de torneios tiveram uma tendência em selecionar quase sempre os mesmos marcadores, tanto SNPs com efeitos não nulos quanto SNPs com efeitos nulos, ou seja, nas 100 repetições de cada análise, os torneios selecionaram determinados grupos de marcadores com alta frequência. O Lasso Bayesiano selecionou muitos marcadores diferentes com frequências baixas e, os marcadores com efeitos não nulos foram selecionados com frequências menores que as obtidas nos torneios.

Outro resultado muito interessante é que os marcadores selecionados pelos torneios com grupos aleatórios e os marcadores selecionados foram praticamente os mesmos. Comparando-se as frequências dos marcadores selecionados por torneios com grupos aleatórios (C) e por torneios com grupos condicionados aos cromossomos (D) nas Figuras 30 a 32, observa-se que SNPs nas mesmas posições dos cromossomos são selecionados em ambos os tipos de torneios e com frequências muito próximas.

O fato de ambos os tipos de torneios selecionarem grupos de marcadores muito parecidos explica o fato de as correlações entre GBVs preditos e simulados e entre GBVs preditos e fenótipos terem sido tão próximas para os dois tipos de torneios.

Resultados semelhantes aos obtidos nas Figuras 30 a 32 podem ser observados nas Figuras 52 a 57, que representam as frequências dos marcadores selecionados em 100 análises referentes aos cenários de 48 SNPs com efeitos não nulos dispersos e 250 SNPs com efeitos não nulos dispersos.

4.6 Resultados das análises com fenótipos reais considerando torneios com grupos de 25 marcadores

Na Figura 33, são apresentadas as correlações médias entre GBVs preditos e fenótipos para amostras de estimação (Figura 33a) e validação (Figura

33b), em um esquema de validação cruzada, para os fenótipos reais da característica área de olho de lombo (AOL) corrigidos para efeitos de grupos de contemporâneos. Observa-se que as metodologias de torneios apresentaram maiores correlações em relação ao Lasso Bayesiano, tanto nas amostras de estimação (Figura 33a) quanto nas amostras de validação (Figura 33b), sendo que, nas amostras de validação, essa diferença foi muito mais expressiva. Observa-se, também que com o aumento do número de marcadores selecionados, a diferença entre as correlações obtidas pelos torneios e as correlações obtidas pelo Lasso Bayesiano diminuem.

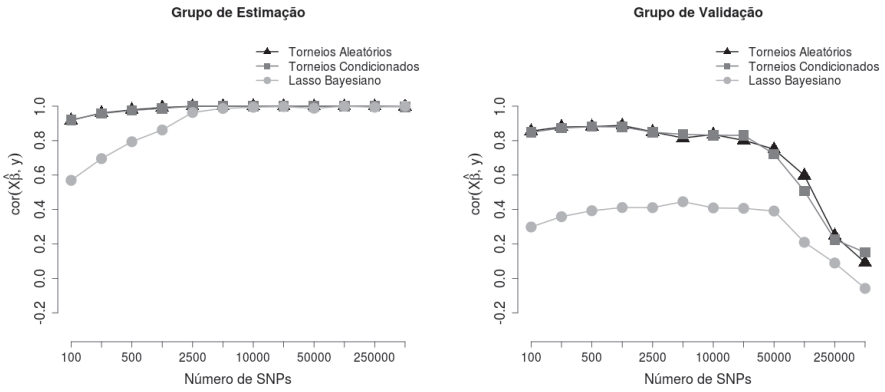

Figura 33 Correlações médias entre GBVs preditos e fenótipos para amostras de estimação (a) e validação (b), considerando diferentes metodologias e diferentes números de SNPs selecionados, para a característica área de olho de lombo

As correlações entre GBVs preditos e fenótipos, $cor(X\hat{\beta}, y)$, no entanto, são medidas indiretas da capacidade de predizer os valores genéticos. Para se avaliar de forma direta a capacidade de predizer os valores genéticos, seria necessário calcular as correlações entre os GBVs preditos e os verdadeiros GBVs (desconhecidos), $cor(X\hat{\beta}, X\beta)$. Contudo, note que o comportamento das

correlações entre GBVs preditos e fenótipos nas análises com fenótipos reais (Figura 33) foi muito semelhante ao obtido em uma das situações avaliadas no estudo de simulação. Esta situação é referente ao cenário de 250 SNPs com efeitos não nulos dispersos e herdabilidade 0,25, cujas correlações entre GBVs preditos e fenótipos para as amostras de estimação e validação são apresentados nas Figuras 23 (pg. 68) e 29 (pg. 73), respectivamente. Note que, tanto nas análises com fenótipos reais quanto nas análises com fenótipos simulados, houve uma diferença muito grande entre as correlações obtidas nos torneios e as correlações obtidas no Lasso Bayesiano, principalmente para números pequenos de marcadores selecionados. Como o comportamento das correlações entre GBVs preditos e fenótipos para as análises com fenótipos reais foi muito semelhante ao comportamento das correlações entre GBVs preditos e fenótipos para a referida situação do estudo de simulação, então, observando o comportamento das correlações entre os GBVs preditos e os GBVs simulados (conhecidos) do estudo de simulação pode-se ter um indicativo do comportamento das correlações entre os GBVs preditos e os verdadeiros GBVs (desconhecidos) das análises com fenótipos reais. Podem-se observar nas Figuras 20 (pg. 66) e 26 (pg. 70) as correlações entre GBVs preditos e simulados, $cor\left(X\hat{\beta}, X\beta\right)$, em amostras de estimação e validação, respectivamente, para um cenário de 250 SNPs com efeitos não nulos dispersos. Observa-se na Figura 20 que, para as amostras de estimação, quase não houve diferenças entre as correlações obtidas pelas diferentes metodologias, enquanto na Figura 26, observa-se que, para as amostras de validação, as correlações entre GBVs preditos e simulados obtidas nos torneios continuam sendo maiores que as obtidas no Lasso Bayesiano, porém, com uma diferença menor do que a observada quando foram consideradas as correlações entre GBVs preditos e fenótipos. Dessa forma, considerando as correlações entre GBVs preditos e fenótipos, $cor\left(X\hat{\beta}, X\beta\right)$, obtidas nas análises com fenótipos reais de área de olho

de lombo (Figura 33), espera-se que as correlações entre os GBVs preditos e os verdadeiros GBVs (desconhecidos), $cor(X\hat{\beta}, X\beta)$,continuem sendo maiores para os torneios do que para o Lasso Bayesiano, porém, com uma diferença menor do que a observada nas correlações entre GBVs preditos e fenótipos. Observa-se ainda que as correlações obtidas nas metodologias de torneios com diferentes formações de grupos (aleatórios e condicionados aos cromossomos) foram muito próximas entre si.

As análises para obtenção da classificação dos marcadores pelas metodologias de torneios e pelo Lasso Bayesiano foram executadas em um computador HP BL660c Gen8 com processador Intel® Xeon® E5-4617 (6 núcleos, 2,9 GHz, 15 MB, 130 W) e 128 GB de memória RAM. Os tempos de execução das análises são apresentados na Tabela 5.

Tabela 5 Tempo de execução das análises para obtenção da classificação dos marcadores para a característica área de olho de lombo

Metodologia	Tempo de Análise
Torneios com grupos aleatórios	11 minutos
Torneios com grupos condicionados	4 horas e 14 minutos
Lasso Bayesiano sem torneios	3 horas e 8 minutos

Observe que a metodologia de torneios com grupos aleatórios apresentou desempenho computacional muito superior às demais metodologias. A metodologia de torneios com grupos condicionados aos cromossomos foi a que demandou maior tempo de análise, sendo mais demorada que o Lasso Bayesiano.

Com relação à seleção de marcadores, na Figura 34, podem-se observar as frequências dos marcadores selecionados em 100 torneios com grupos aleatórios em que, em cada torneio, 100 marcadores foram selecionados. Como foi visto no estudo de simulação, na seção 4.5 (pg. 74), a metodologia de

torneios teve uma tendência em selecionar determinados grupos de marcadores com alta frequência em análises repetidas, porém, observou-se que esses grupos de marcadores selecionados com alta frequência eram compostos não somente pelos verdadeiros marcadores com efeitos não nulos, mas, também, por marcadores cujos verdadeiros efeitos são nulos. Dessa forma, dentre os marcadores selecionados com maiores frequências nos torneios para os dados reais de AOL (Figura 34) podem estar os verdadeiros marcadores com efeitos não nulos como também marcadores cujos efeitos são, na realidade, nulos.

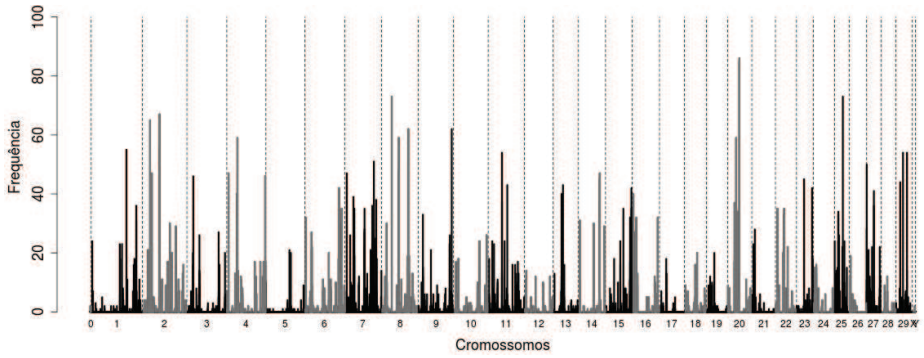

Figura 34 Frequências dos SNPs selecionados em 100 torneios com grupos aleatórios, com 100 SNPs selecionados em cada torneio, para a característica área de olho de lombo

Assim, para os dados reais de AOL, além da seleção de marcadores por torneios, foi aplicado ao conjunto reduzido de marcadores um procedimento adicional para seleção de marcadores. Como os efeitos dos marcadores selecionados pelo torneio são estimados utilizando o Lasso Bayesiano, então, a seleção final de marcadores foi realizada por meio de intervalos de credibilidade HPD (Highest Posterior Density) de 95%.

Na Figura 35, são apresentadas as estimativas dos efeitos de 100 marcadores selecionados em um dos torneios com grupos aleatórios, dentre os 100 torneios realizados, com HPDs de 95% para os marcadores cujos efeitos

diferem significativamente de zero (HPD não inclui o zero). Observa-se que após reduzir o número de marcadores de 526.493 para 100, utilizando a metodologia de torneios, os efeitos dos marcadores foram estimados pelo Lasso Bayesiano e foram calculados seus HPDs, sendo então selecionados 23 marcadores distribuídos dentre 15 cromossomos.

Para os marcadores selecionados em cada um dos 100 torneios, foi então aplicado um procedimento adicional de seleção de marcadores por meio de HPDs. O número de marcadores selecionados por HPDs nas 100 análises variou de 19 à 37 marcadores, sendo que a média foi de aproximadamente 28 marcadores selecionados.

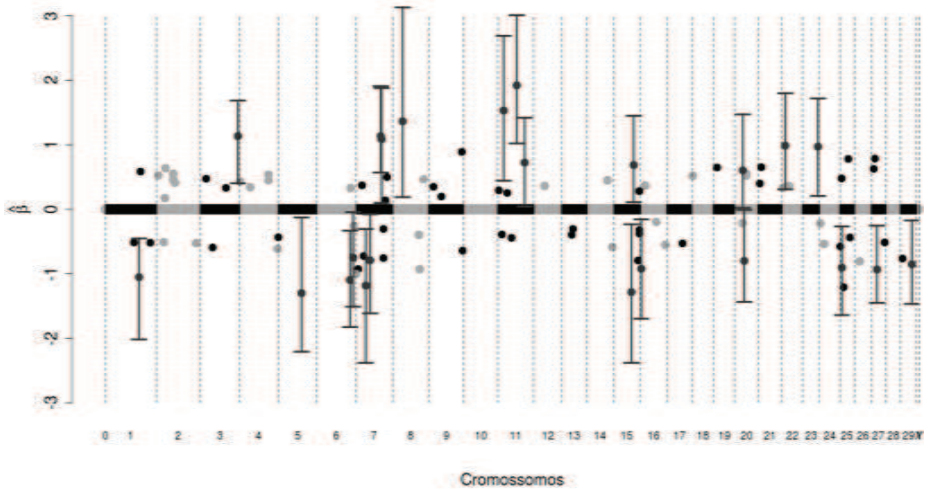

Figura 35 Estimativas dos efeitos de 100 SNPs selecionados em um torneio com grupos aleatórios, com HPDs de 95% para os marcadores cujos efeitos diferem significativamente de zero, para a característica área de olho de lombo

Para avaliar a capacidade de predição dos GBVs utilizando marcadores selecionados por HPDs foram ajustados modelos a estes marcadores com o Lasso Bayesiano em um esquema de validação cruzada (com oito grupos de 48

animais) e foram calculadas as correlações entre GBVs preditos e fenótipos para as amostras de estimação e validação. A correlações obtidas nas amostras de estimação nas 100 análises variaram de 0,7037 à 0,8719, com média 0,8084, e, para as amostras de validação as correlações variaram de 0,6559 à 0,8372, com média 0,7658.

Na Figura 36, são apresentadas as frequências dos marcadores selecionados por HPDs nas 100 análises. Note que os marcadores selecionados pelos HPDs diferem significativamente de zero ao nível de credibilidade de 95%. Dessa forma, é provável que um marcador selecionado por HPD, cuja frequência nas 100 análises é alta, seja realmente um marcador cujo verdadeiro efeito é não nulo ou que ele esteja em desequilíbrio de ligação com algum marcador cujo verdadeiro efeito é não nulo.

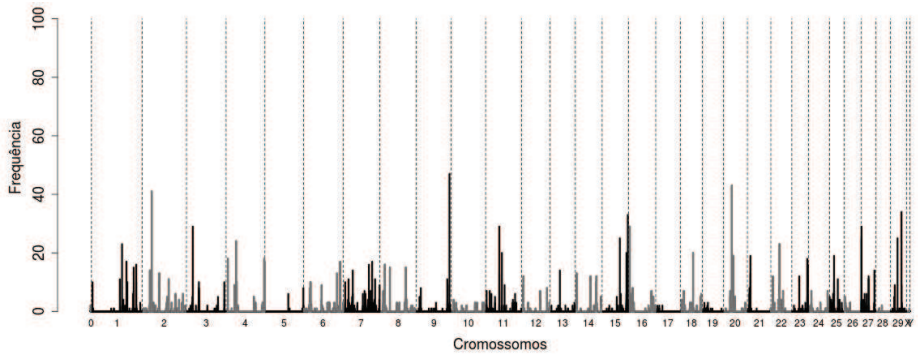

Figura 36 Frequências, em 100 análises, dos SNPs selecionados por HPDs de 95%, calculados para os SNPs previamente selecionados em torneios com grupos aleatórios para a característica área de olho de lombo

Na Tabela 6, são apresentadas as correlações médias entre GBVs preditos e fenótipos para amostras de validação e de estimação de um esquema de validação cruzada, em que, os modelos foram ajustados considerando diferentes frequências mínimas com que os marcadores foram selecionados nas 100 análises.

Tabela 6 Correlações médias entre GBVs preditos e fenótipos em amostras de estimação e validação de um esquema de validação cruzada, considerando marcadores selecionados por HPDs de 95% em 100 análises, com modelos ajustados a grupos de marcadores selecionados com determinadas frequências mínimas nas 100 análises, para a característica área de olho de lombo

Frequência mínima em 100 análises	Número de marcadores	Amostra de estimação	Amostra de validação
1	775	0,9946	0,9238
2	407	0,9853	0,9260
3	275	0,9772	0,9200
4	185	0,9683	0,9170
5	150	0,9616	0,9103
6	127	0,9567	0,9107
7	104	0,9482	0,9032
8	85	0,9333	0,8859
9	78	0,9273	0,8839
10	69	0,9192	0,8784
11	57	0,9039	0,8634
12	50	0,8883	0,8463
13	44	0,8638	0,8186
14	39	0,8379	0,7848
15	35	0,8104	0,7571
16	32	0,7947	0,7437
17	28	0,7793	0,7305
18	24	0,7523	0,7058
19	21	0,7263	0,6774
20	18	0,6925	0,6474

Na primeira coluna da Tabela 6, é apresentada a frequência mínima com que os marcadores foram selecionados, na segunda coluna, é apresentado os número de marcadores selecionados com a respectiva frequência mínima e, na terceira e quarta colunas, são apresentadas as correlações entre GBVs preditos e fenótipos nas amostras de estimação e validação, respectivamente. Por exemplo, na primeira linha, houve 775 marcadores selecionados com frequência mínima

igual a um, ou seja, 775 marcadores selecionados ao menos uma vez nas 100 análises (todos os marcadores selecionados nas 100 análises) e que, as correlações obtidas para estes marcadores nas amostras de estimação e de validação foram de 0,9946 e 0,9238, respectivamente.

Pode-se ver que as correlações na amostra de validação atingem valores acima de 0,90 quando são utilizados marcadores com frequências mínimas variando de um a sete. Especificamente para marcadores com frequência mínima igual a sete, foram selecionados 104 marcadores, que é próximo dos 100 marcadores selecionados em um torneio individual, e a correlação obtida na amostra de validação foi de 0,9032, que foi maior que a correlação obtida na amostra de validação para um torneio individual (próxima de 0,85) como se pode ver na Figura 33b (pg. 80). Além disso, utilizando os 50 marcadores com frequência mínima igual a 12, a correlação obtida na amostra de validação foi de 0,8463, que é bem próxima da correlação obtida na amostra de validação para um torneio individual com 100 marcadores selecionados - Figura 33b (pg. 80). Para marcadores selecionados com frequências maiores, as correlações diminuem consideravelmente.

5 CONCLUSÃO

Torneios com grupos aleatórios e torneios com grupos condicionados aos cromossomos tendem a selecionar consistentemente grupos de marcadores muito semelhantes. Consequentemente terão capacidade de predição de valores genéticos também semelhantes.

Os torneios apresentaram resultados superiores ao Lasso Bayesiano tanto na seleção de marcadores quanto nas predições para a maioria das situações consideradas no estudo de simulação. O Lasso Bayesiano apresentou uma tendência em selecionar marcadores diferentes em cada análise. Da mesma forma, os torneios apresentaram resultados superiores ao Lasso Bayesiano nas análises com dados reais de área de olho de Lombo (AOL).

Conclui-se que os torneios são uma boa estratégia de seleção de marcadores e devem ser adotados em conjunção a análises posteriores mais sofisticadas.

Como uma conjectura final, se os torneios tendem a selecionar várias vezes os mesmos marcadores em análises repetidas enquanto o Lasso Bayesiano tende a selecionar marcadores diferentes em análises repetidas e, como os torneios apresentaram maior capacidade de predição de valores genéticos do que o Lasso Bayesiano, imagina-se que os valores genéticos preditos a partir de marcadores selecionados por torneios têm maior tendência a resultar de combinações de marcadores associados à característica, enquanto os valores genéticos preditos a partir de marcadores selecionados pelo Lasso Bayesiano resultam de segregações convenientemente associadas ao caráter, mesmo que provindas de marcadores que nada têm a ver com o caráter.

A utilização de um procedimento adicional de seleção de marcadores aplicado aos marcadores pré-selecionados por torneios pode auxiliar na seleção de marcadores associados à característica e melhorar a qualidade das predições

dos valores genéticos. A utilização de intervalos de credibilidade para o Lasso Bayesiano, realizado após a seleção de marcadores por meio de torneios, associada à repetição da análise de um grande número de vezes fornece maiores evidências sobre os verdadeiros marcadores associados à característica e, com base no número de vezes com que os marcadores foram selecionados em repetidas análises, ou seja, na frequência com que foram selecionados em várias análises, podem-se utilizar grupos de marcadores selecionados com maiores frequências nestas análises para melhorar predições de valores genéticos e potencialmente identificar genes realmente associados ao caráter de interesse.

A metodologia de torneios com grupos aleatórios foi muito mais eficiente que as demais em relação ao tempo de execução da análise, sendo aproximadamente 17 vezes mais rápida que o Lasso Bayesiano e aproximadamente 23 vezes mais rápida que a metodologia de torneios com grupos condicionados aos cromossomos.

Como os torneios com grupos condicionados aos cromossomos não apresentaram nenhuma vantagem em relação aos torneios com grupos aleatórios quanto à seleção de marcadores ou quanto à predição de valores genéticos e, além disso, apresentaram desvantagem computacional, então a metodologia de torneios com grupos aleatórios é recomendada como a melhor das metodologias avaliadas.

REFERÊNCIAS

AGNIHOTRAM, T. **Statistical analysis of target SNPs and their association with phenotypes**. 2012. 59 p. Dissertation (Master in Statistical) - Swiss Federal Institute of Technology Zurich, Zurich, 2012.

ALMEIDA, I. F. de. **Métodos de estimação e validação na seleção genômica na presença ou ausência de correção de fenótipos**. 2013. 55 p. Tese (Doutorado em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa, MG, 2013.

ALVES, R. R. **Seleção por torneios nas estimativas de associação entre marcadores SNPs e fenótipos**. 2014. 70 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2014.

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. **Journal Royal Statistical Society Series B**, London, v. 36, p. 99-102, 1974.

AZEVEDO, C. F. et al. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, Brasília, v. 48, n. 6, p. 619-626, jun. 2013.

BELETI JUNIOR, C. R. **Paralelização de aplicações na plataforma R**. 2013. 95 p. Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Maringá, Maringá, 2013.

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. Disponível em: <http://link.springer.com/article/10.1007/s11425-009-0089-4>. Acesso em: 10 mar. 2014.

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

FAN, J.; LI, R. Variable selection via non-concave penalized likelihood and its oracle properties. **Journal of the American Statistical Association**, Madison, v. 96, n. 456, p. 1348-1360, Dec. 2001.

FAN, J.; LV, J. Sure independence screening for ultrahigh dimensional feature space. **Journal of the Royal Statistical Society Series B-Statistical Methodology**, London, v. 70, p. 849-911, Nov. 2008.

JANSS, L. et al. Inferences from genomic models in stratified populations. **Genetics**, Austin, v. 192, n. 2, p. 693-694, Oct. 2012.

LEGARRA, A. et al. Improved Lasso for genomic selection. **Genetics Research**, Cambridge, v. 93, n. 1, p. 77-87, Feb. 2011.

LI, J. et al. The Bayesian lasso for genome-wide association studies. **Bioinformatics**, Oxford, v. 27, n. 4, p. 516-523, Feb. 2011.

LI, Z.; SILLANPÄÄ, M. J. Overview of Lasso-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and Applied Genetics**, Berlin, v. 125, n. 3, p. 419-435, Aug. 2012.

MANOLIO, T. A.; BROOKS, L. D.; COLLINS, F. S.A HapMap harvest of insights into the genetics of common disease. **Journal of Clinical Investigation**, Ann Arbor, v. 118, n. 5, p. 1590-1605, May 2008.

MEUWISSEN, T. Genomic selection: marker assisted selection on a genome wide scale. **Journal of Animal Breeding and Genetics**, New York, v. 124, n. 6, p. 321-322, Dec. 2007.

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, Apr. 2001.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, Madison, v. 103, n. 482, p. 681-686, June 2008.

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

RESENDE, M. D. V. et al. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n. 56, p. 63-77, 2008.

RESENDE, M. D. V. et al. **Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória (RR) e estatística espacial**. Viçosa, MG: UFV, 2012. 291 p.

SILVA, F. F. et al. A note on accuracy of Bayesian Lasso regression in GWS. **Livestock Science**, New York, v. 142, n. 1/3, p. 310-314, Dec. 2011.

SILVA, F. F. et al. Seleção genômica ampla para curvas de crescimento. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 65, n. 5, p. 1519-1526, Out. 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, May. 2010.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistical Society Series B-Methodological**, London, v. 58, n. 1, p. 267-288, 1996.

TOGASHI, K.; LIN, C. Y.; YAMAZAKI, T. The efficiency of genome-wide selection for genetic improvement of net merit. **Journal of Animal Science**, Champaign, v. 89, n. 10, p. 2972-2980, Oct. 2011.

URBANEK, S. **Multicore:** parallel processing of R code on machines with multiple cores or CPUs. Disponível em: <http://cran.r-project.org/package=multicore>. Acesso em: 10 maio 2014.

YANG, J. et al. GCTA: a tool for genome-wide complex trait analysis. **American Journal of Human Genetics**, Cambridge, v. 88, n. 1, p. 76-82, Jan. 2011.

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

APÊNDICES

APÊNDICE A - Efeitos da multicolinearidade sobre análises utilizando regressão linear múltipla clássica

```
Call:
lm(formula = y ~ -1 + X[,i:(i+99)])

Residuals:
Min      1Q  Median      3Q     Max
-15.4650  -3.3161  -0.1227   3.5732  20.3680

Coefficients: (22 not defined because of singularities)
                                  Estimate  Std. Error  t value  Pr(>|t|)
X[,i:(i+99)]Hapmap43437-BTA-101873   14.678      21.754    0.675    0.5003
X[,i:(i+99)]BovineHD0100000042        0.1461     44.316    0.033    0.9737
X[,i:(i+99)]BovineHD0100000046        0.8019     38.997    0.206    0.8372
X[,i:(i+99)]BovineHD0100000048       -0.3221     41.268   -0.078    0.9378
X[,i:(i+99)]BovineHD0100000049       19.511      26.457    0.737    0.4614
X[,i:(i+99)]BovineHD0100000051      -14.717      30.929   -0.476    0.6345
X[,i:(i+99)]BovineHD0100000054        0.3098     16.799    0.184    0.8538
X[,i:(i+99)]BovineHD0100046368      170.105     280.932    0.606    0.5453
X[,i:(i+99)]BovineHD0100000057       23.746      44.449    0.534    0.5936
X[,i:(i+99)]BovineHD0100000058      189.112     279.194    0.677    0.4987
X[,i:(i+99)]BovineHD0100000060      241.376     195.566    1.234    0.2181
X[,i:(i+99)]BovineHD0100000061      273.234     185.641    1.472    0.1421
X[,i:(i+99)]BovineHD0100000063      236.896     314.344    0.754    0.4517
                  ⋮                    ⋮           ⋮        ⋮        ⋮
X[,i:(i+99)]BovineHD0100000083      -24.472      21.374   -1.145    0.2531
X[,i:(i+99)]BovineHD0100000084           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000086      -44.545     103.126   -0.432    0.6661
X[,i:(i+99)]BovineHD0100000087       10.248      98.700    0.104    0.9174
X[,i:(i+99)]BovineHD0100000089           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000090      -56.458      80.936   -0.698    0.4860
X[,i:(i+99)]BovineHD0100000091      -69.513      67.475   -1.030    0.3037
X[,i:(i+99)]BovineHD0100000092           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000094     -168.112     132.163   -1.272    0.2043
X[,i:(i+99)]BovineHD0100000095       -0.1745     42.550   -0.041    0.9673
X[,i:(i+99)]BovineHD0100000098           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000099           NA          NA       NA        NA
X[,i:(i+99)]ARS-BFGL-NGS-105096          NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000101      162.285     111.751    1.452    0.1475
X[,i:(i+99)]BovineHD0100000102      144.966      75.380    1.923    0.0554   .
X[,i:(i+99)]BovineHD0100000103           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000104           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000105     -159.814      65.916   -2.425    0.0159   *
X[,i:(i+99)]BovineHD0100000106           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000107           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000108           NA          NA       NA        NA
X[,i:(i+99)]BovineHD0100000111       21.067      57.190    0.368    0.7128
X[,i:(i+99)]BovineHD0100000115       25.085      55.015    0.456    0.6487
X[,i:(i+99)]BovineHD0100000117      106.123      49.171    2.158    0.0317   *
X[,i:(i+99)]BovineHD0100000121     -113.921      55.842   -2.040    0.0422   *
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 37 Resultado de uma regressão linear múltipla clássica utilizando 100 SNPs dispostos sequencialmente no cromossomo 1

APÊNDICE B - Gráficos

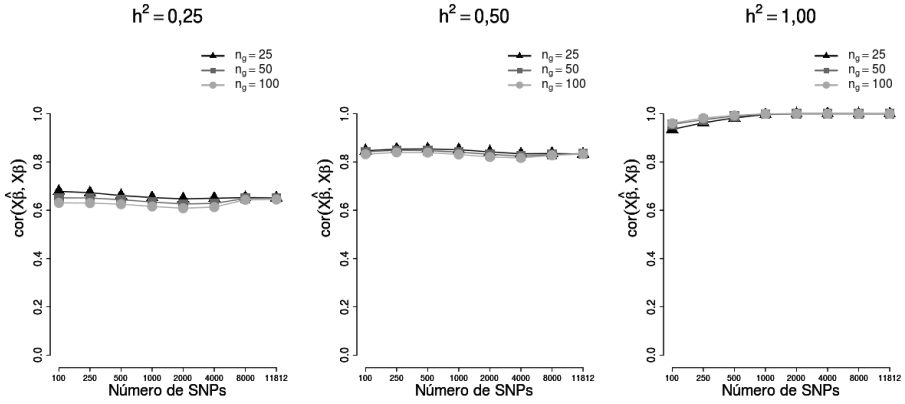

Figura 38 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

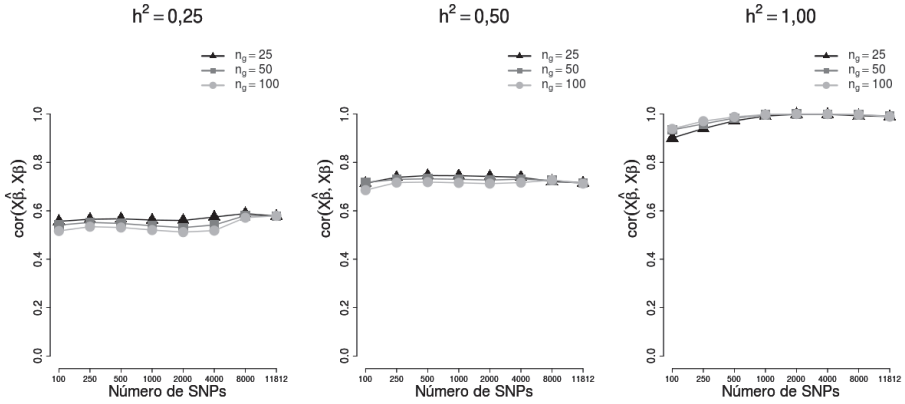

Figura 39 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

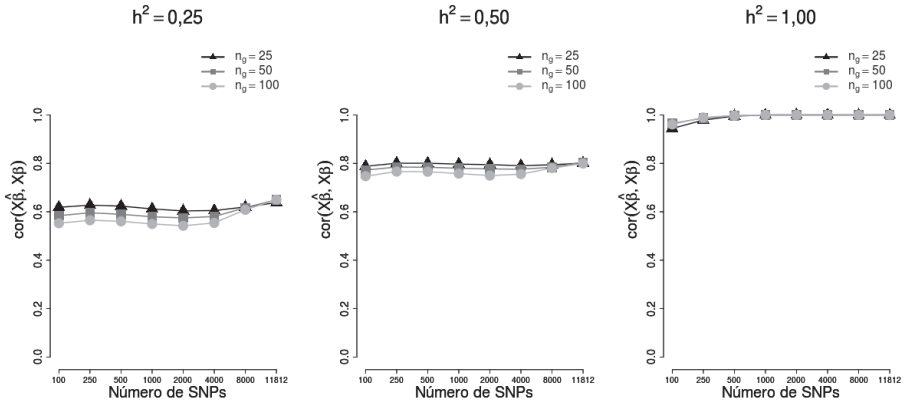

Figura 40 Correlações médias entre GBVs preditos e simulados para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

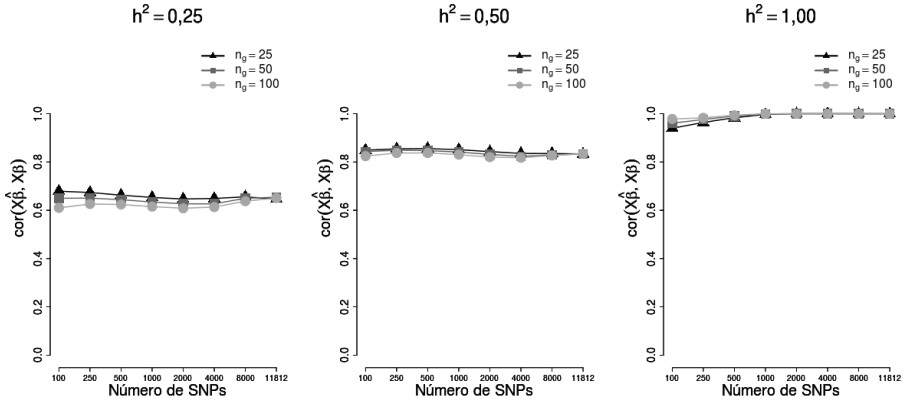

Figura 41 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

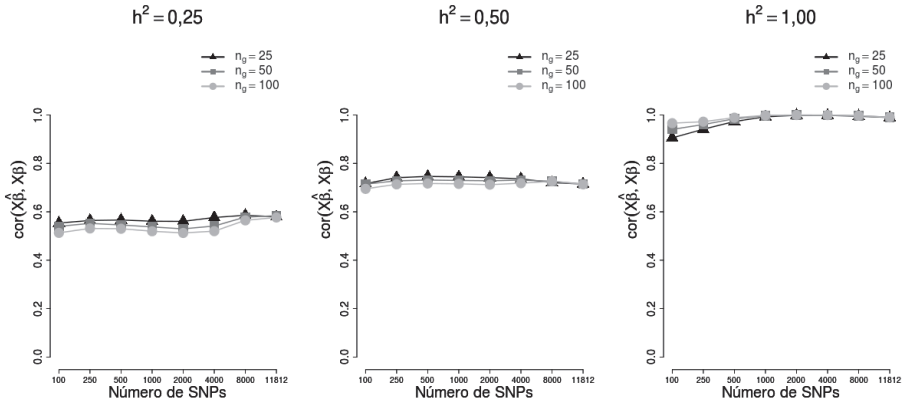

Figura 42 Correlações médias entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

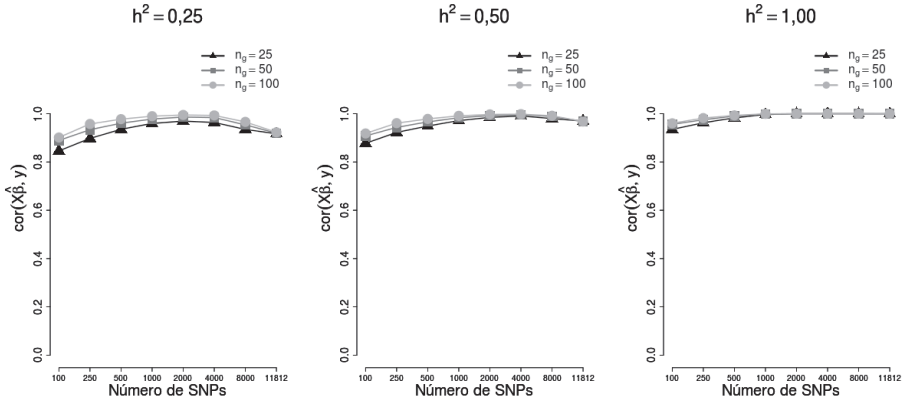

Figura 43 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

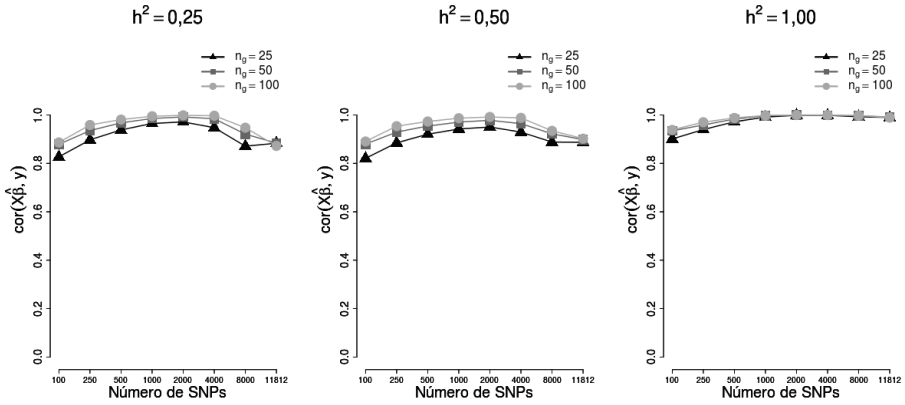

Figura 44 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

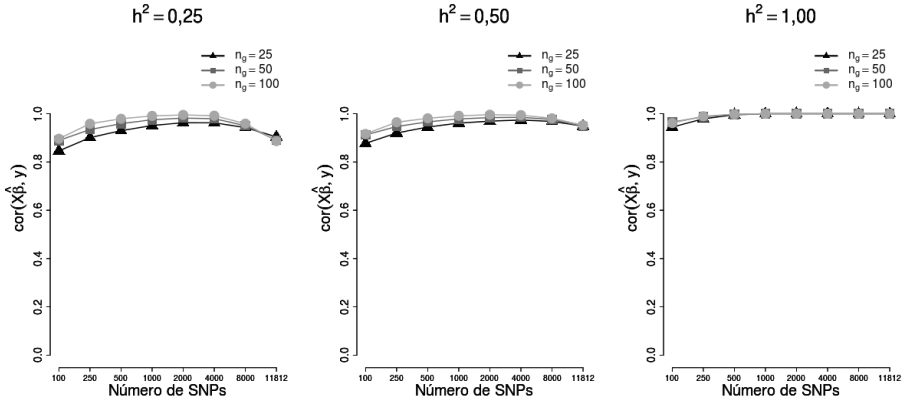

Figura 45 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos aleatórios, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

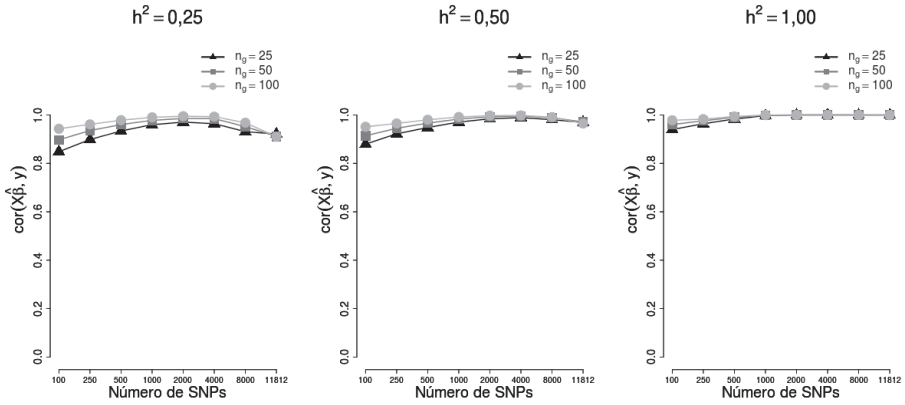

Figura 46 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos agrupados

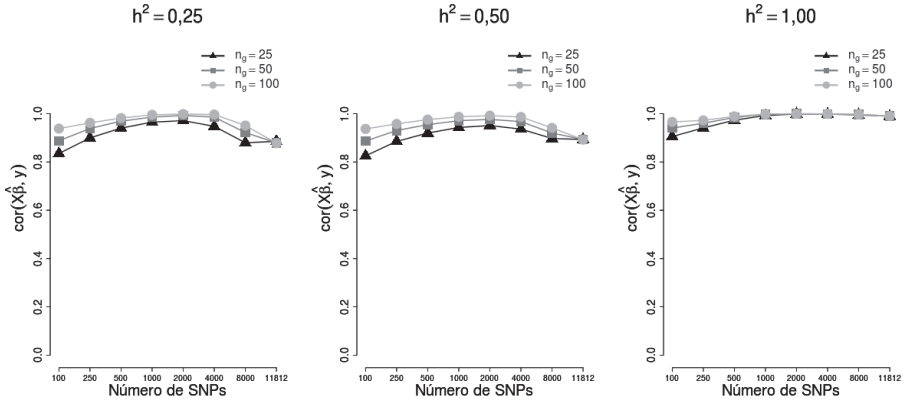

Figura 47 Correlações médias entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos, considerando diferentes tamanhos de grupos, diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

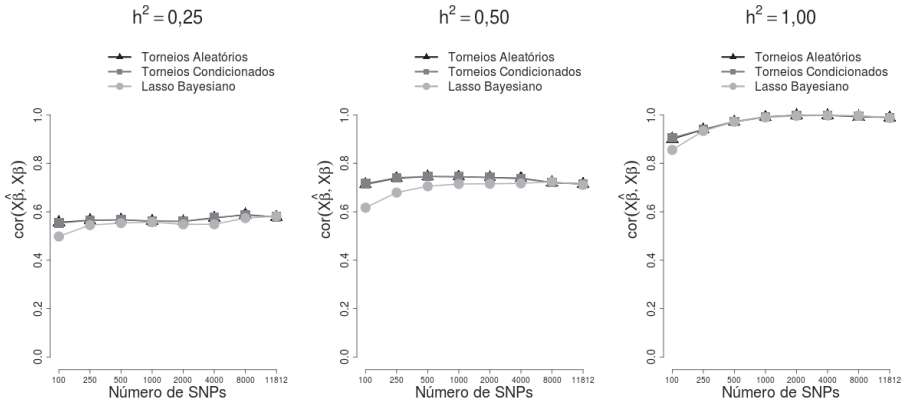

Figura 48 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

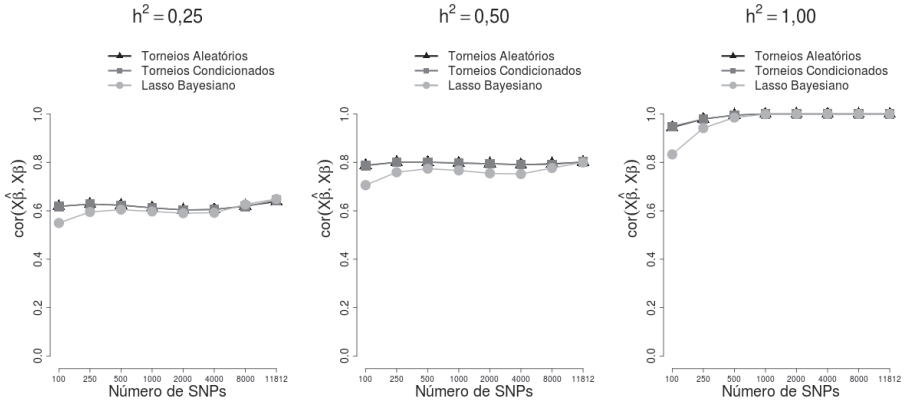

Figura 49 Correlações médias entre GBVs preditos e simulados, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

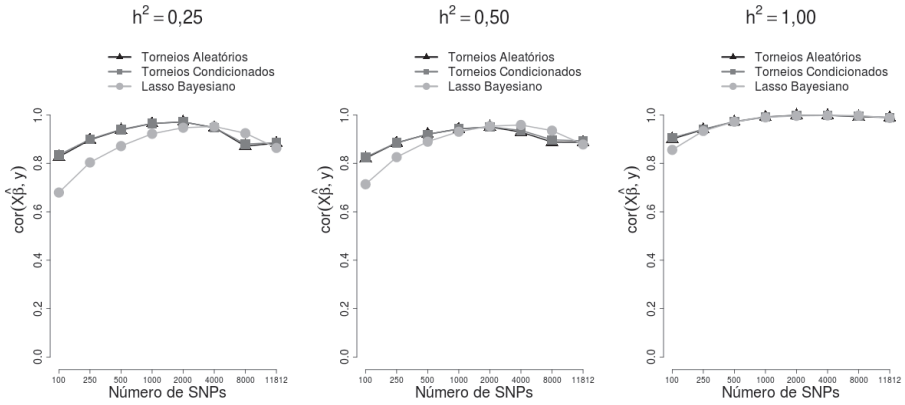

Figura 50 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 48 SNPs com efeitos não nulos dispersos

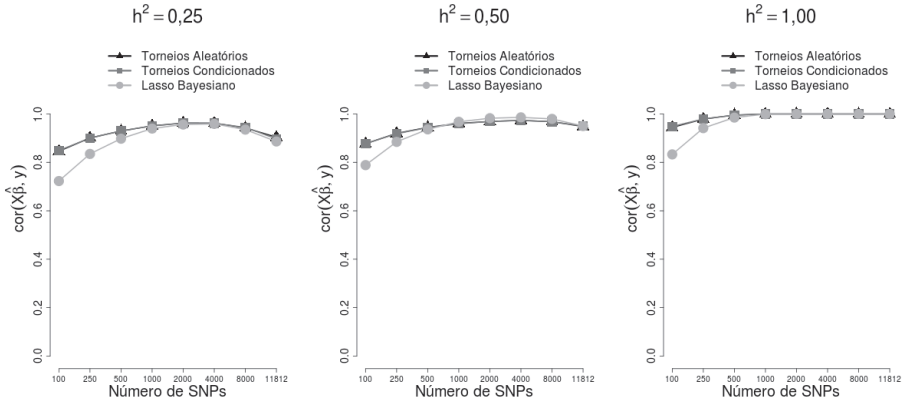

Figura 51 Correlações médias entre GBVs preditos e fenótipos, para diferentes metodologias, considerando diferentes números de SNPs selecionados e diferentes herdabilidades, em um cenário de 250 SNPs com efeitos não nulos dispersos

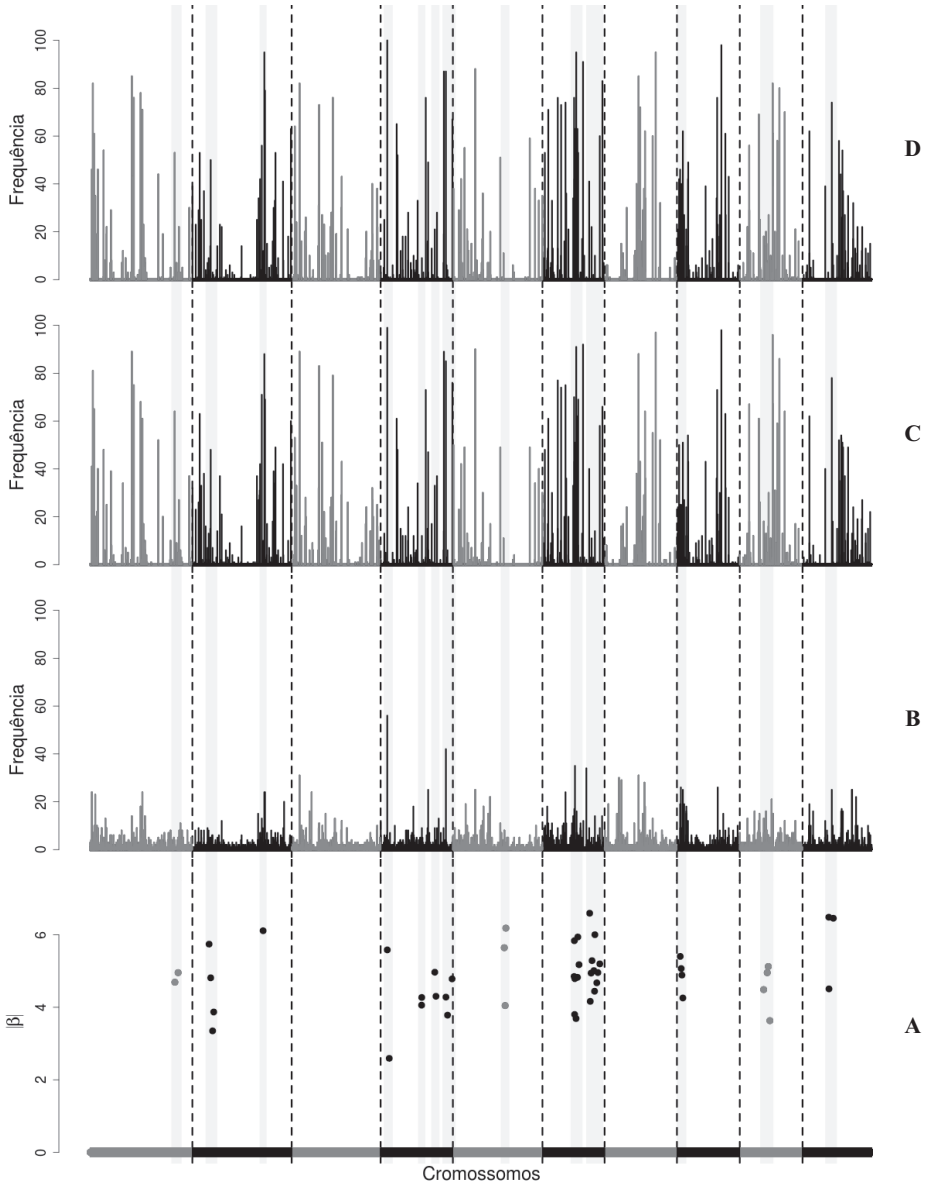

Figura 52 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 0,25

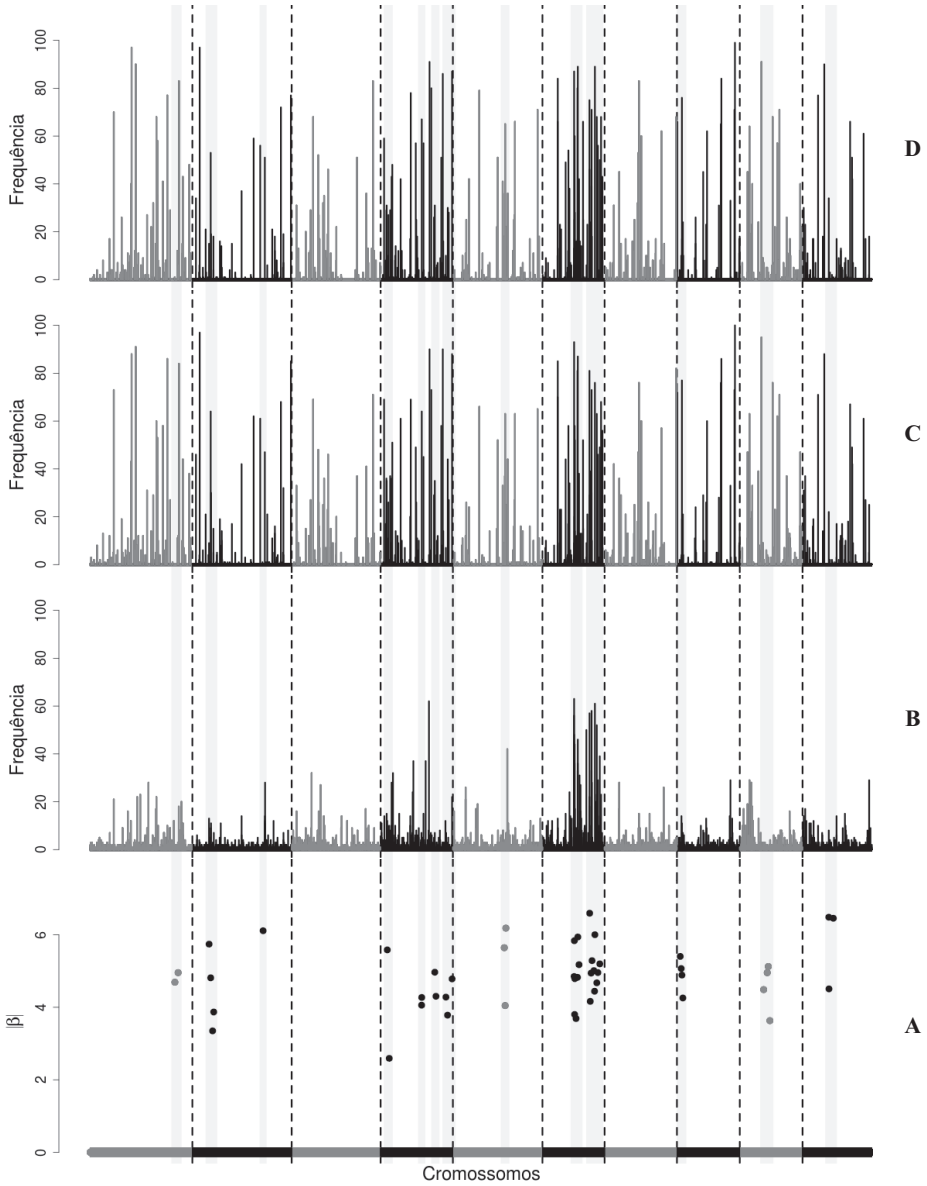

Figura 53 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 0,5

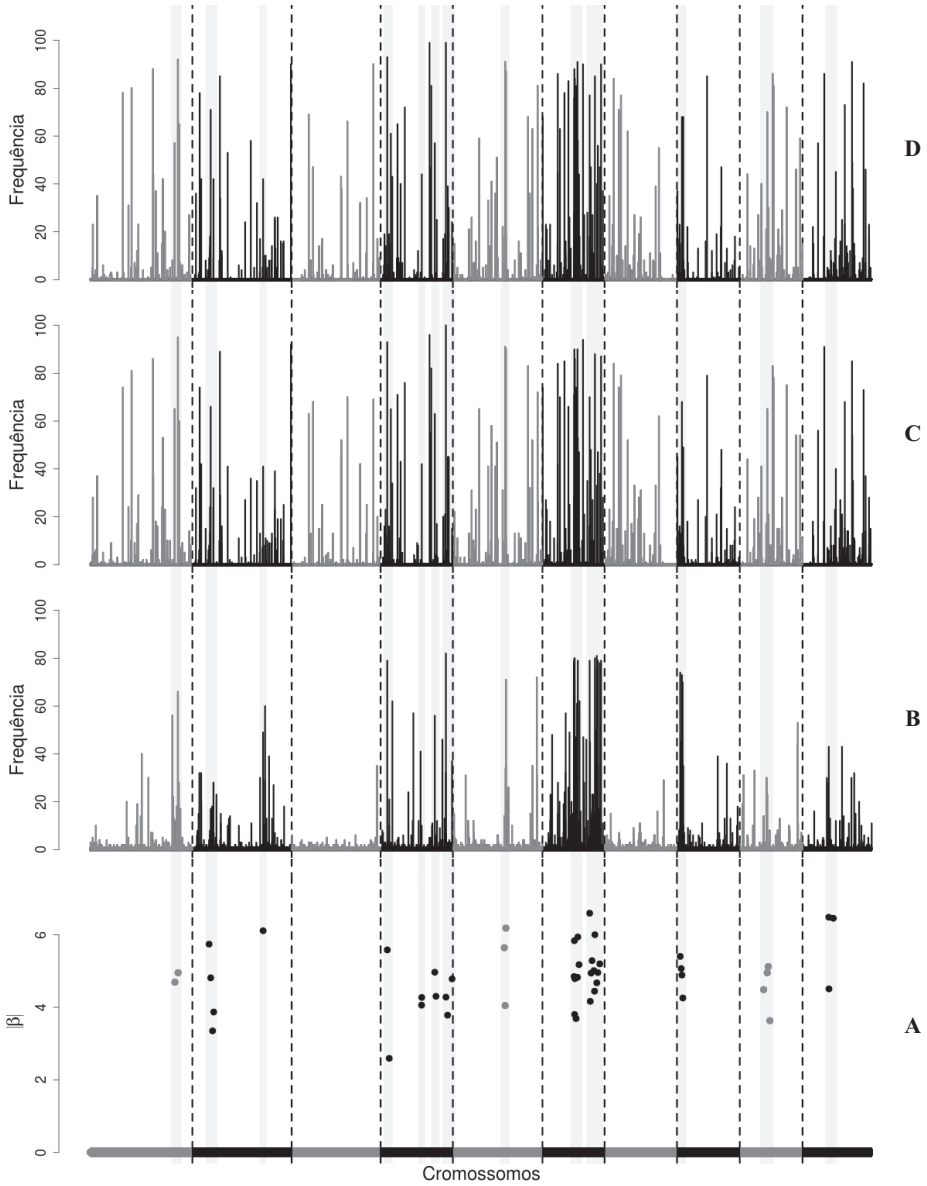

Figura 54 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 48 SNPs com efeitos não nulos dispersos e com herdabilidade 1,0

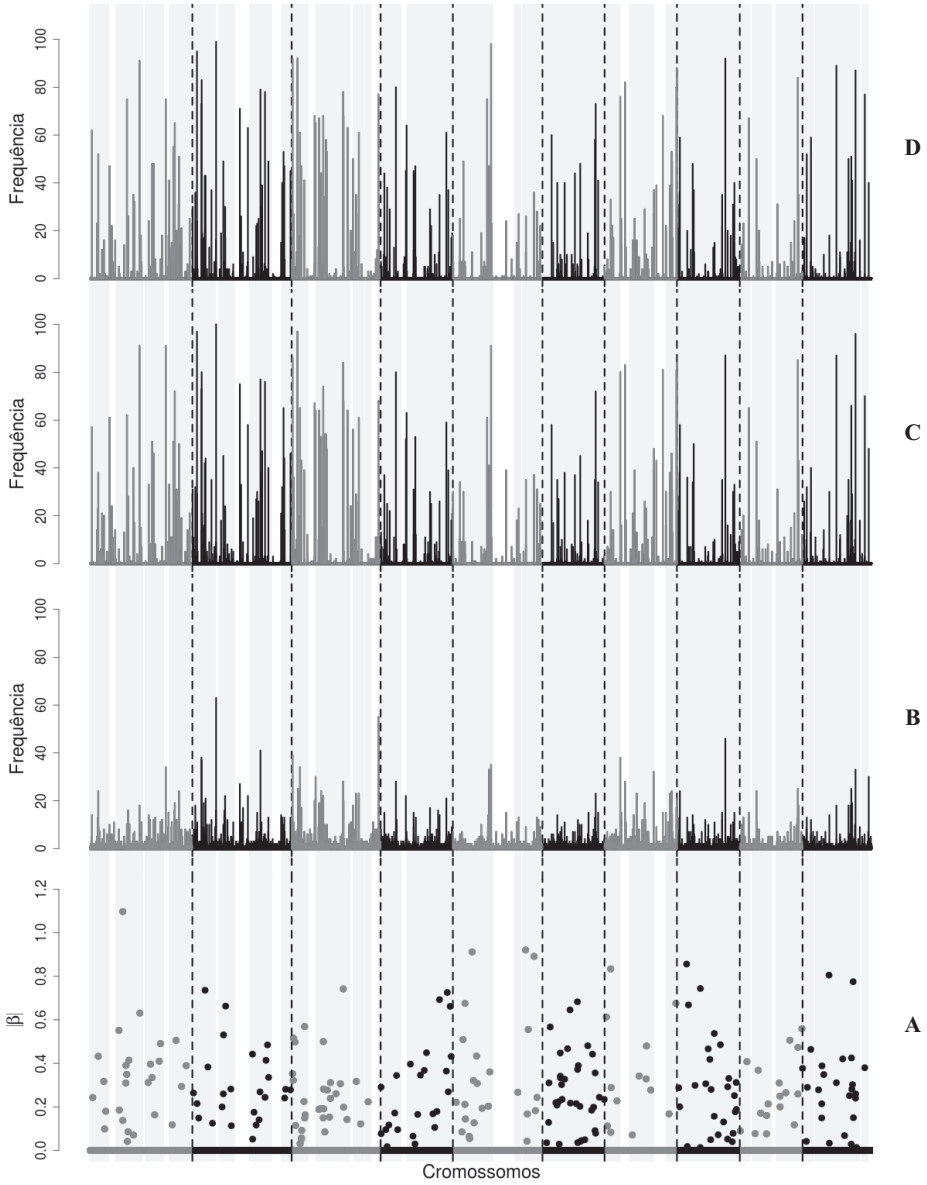

Figura 55 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 0,25

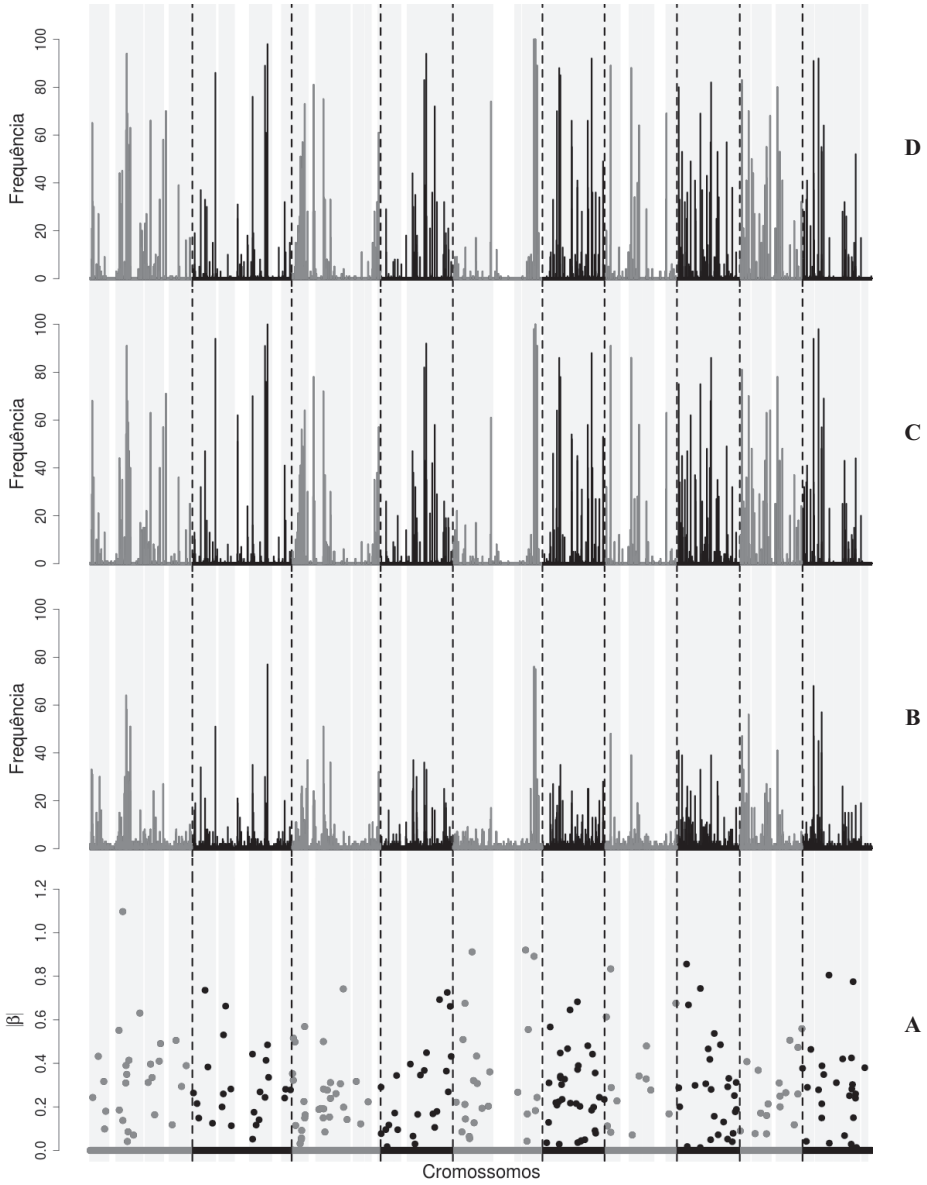

Figura 56 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 0,5

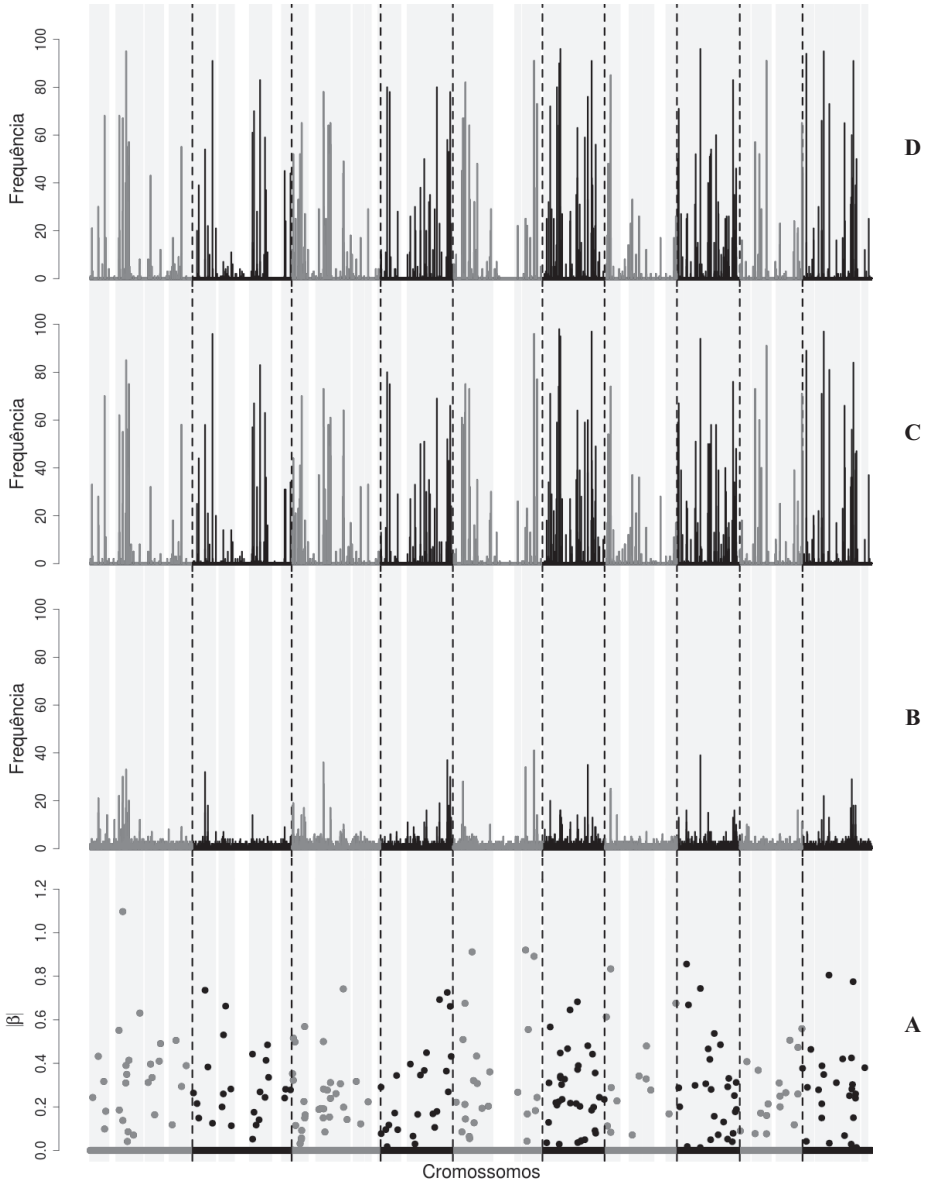

Figura 57 Módulos dos efeitos simulados dos SNPs (A) e frequências dos marcadores selecionados em 100 análises utilizando: (B) Lasso Bayesiano; (C) torneios com grupos aleatórios; (D) torneios com grupos condicionados aos cromossomos, em um cenário de 250 SNPs com efeitos não nulos dispersos e com herdabilidade 1,0

APÊNDICE C - Tabelas

Tabela 7 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para torneios com grupos aleatórios considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Estimação	*25*	*100*	*85,21 ± 0,05*	*87,70 ± 0,04*	*94,26 ± 0,02*
		250	*91,10 ± 0,06*	*91,93 ± 0,06*	*98,05 ± 0,02*
		500	*93,98 ± 0,07*	*95,29 ± 0,07*	*99,79 ± 0,01*
		1000	*95,91 ± 0,10*	*96,97 ± 0,10*	*100,00 ± 0,00*
		2000	*97,04 ± 0,21*	*97,67 ± 0,17*	*100,00 ± 0,00*
		4000	*97,09 ± 0,49*	*97,98 ± 0,29*	*100,00 ± 0,00*
		8000	*94,58 ± 0,85*	*97,33 ± 0,62*	*99,99 ± 0,00*
		11812	*90,57 ± 2,17*	*95,79 ± 1,01*	*99,99 ± 0,01*
	50	*100*	*88,80 ± 0,04*	*91,61 ± 0,03*	*96,29 ± 0,01*
		250	*93,49 ± 0,05*	*95,25 ± 0,05*	*98,79 ± 0,01*
		500	*95,70 ± 0,07*	*97,08 ± 0,06*	*99,72 ± 0,01*
		1000	*97,88 ± 0,07*	*98,41 ± 0,09*	*100,00 ± 0,00*
		2000	*98,95 ± 0,13*	*99,00 ± 0,10*	*100,00 ± 0,00*
		4000	*98,59 ± 0,22*	*98,71 ± 0,20*	*100,00 ± 0,00*
		8000	*95,64 ± 0,87*	*97,96 ± 0,38*	*99,99 ± 0,00*
		11812	*90,96 ± 2,07*	*96,03 ± 1,05*	*99,99 ± 0,01*
	100	*100*	*89,79 ± 0,05*	*92,26 ± 0,03*	*96,49 ± 0,01*
		250	*96,11 ± 0,04*	*96,88 ± 0,03*	*99,08 ± 0,01*
		500	*97,84 ± 0,04*	*98,40 ± 0,03*	*99,95 ± 0,01*
		1000	*99,47 ± 0,05*	*99,48 ± 0,04*	*100,00 ± 0,00*
		2000	*99,76 ± 0,06*	*99,73 ± 0,06*	*100,00 ± 0,00*
		4000	*99,55 ± 0,10*	*99,67 ± 0,08*	*100,00 ± 0,00*
		8000	*96,24 ± 0,80*	*98,28 ± 0,41*	*100,00 ± 0,00*
		11812	*91,17 ± 2,17*	*95,72 ± 0,96*	*99,99 ± 0,01*

continua

Tabela 7 conclusão

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Validação	*25*	*100*	*73,61 ± 0,99*	*78,37 ± 0,71*	*89,11 ± 0,51*
		250	*74,83 ± 1,11*	*78,04 ± 0,72*	*92,42 ± 0,36*
		500	*73,66 ± 0,95*	*77,92 ± 0,88*	*94,14 ± 0,32*
		1000	*71,33 ± 0,96*	*75,45 ± 0,99*	*90,96 ± 0,90*
		2000	*66,65 ± 1,48*	*71,68 ± 1,30*	*87,84 ± 1,28*
		4000	*57,97 ± 2,39*	*64,83 ± 2,16*	*84,84 ± 1,24*
		8000	*40,76 ± 3,96*	*51,59 ± 4,64*	*80,43 ± 2,12*
		11812	*26,68 ± 4,78*	*41,74 ± 5,23*	*77,63 ± 2,07*
	50	*100*	*78,62 ± 0,71*	*84,05 ± 0,48*	*92,82 ± 0,29*
		250	*78,64 ± 0,76*	*83,96 ± 0,71*	*94,21 ± 0,28*
		500	*77,19 ± 0,95*	*82,14 ± 0,85*	*94,15 ± 0,33*
		1000	*75,94 ± 1,06*	*79,49 ± 0,85*	*91,60 ± 0,91*
		2000	*71,44 ± 1,83*	*75,26 ± 1,20*	*88,59 ± 1,03*
		4000	*62,52 ± 2,23*	*67,49 ± 2,72*	*85,96 ± 1,74*
		8000	*41,48 ± 4,50*	*53,65 ± 3,61*	*82,18 ± 2,08*
		11812	*26,56 ± 3,92*	*42,42 ± 4,53*	*79,23 ± 3,53*
	100	*100*	*79,05 ± 0,98*	*85,10 ± 0,56*	*93,29 ± 0,24*
		250	*82,55 ± 0,87*	*86,55 ± 0,54*	*95,23 ± 0,18*
		500	*80,34 ± 1,13*	*84,60 ± 0,76*	*95,59 ± 0,27*
		1000	*79,24 ± 1,31*	*81,81 ± 0,95*	*91,93 ± 0,62*
		2000	*72,87 ± 1,79*	*76,15 ± 1,44*	*89,71 ± 0,78*
		4000	*63,71 ± 2,73*	*68,57 ± 2,77*	*86,69 ± 1,61*
		8000	*42,09 ± 3,80*	*53,93 ± 4,19*	*82,92 ± 1,68*
		11812	*29,50 ± 5,44*	*44,53 ± 4,35*	*78,41 ± 3,37*

Tabela 8 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para torneios com grupos aleatórios considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Estimação	*25*	*100*	*60,09 ± 0,09*	*77,88 ± 0,07*	*94,26 ± 0,02*
		250	*62,32 ± 0,08*	*79,85 ± 0,06*	*98,05 ± 0,02*
		500	*62,30 ± 0,12*	*79,54 ± 0,07*	*99,79 ± 0,01*
		1000	*60,63 ± 0,14*	*78,80 ± 0,10*	*100,00 ± 0,00*
		2000	*59,22 ± 0,36*	*78,58 ± 0,27*	*100,00 ± 0,00*
		4000	*59,31 ± 0,94*	*78,25 ± 0,54*	*100,00 ± 0,00*
		8000	*61,77 ± 1,20*	*78,59 ± 1,04*	*99,99 ± 0,00*
		11812	*63,06 ± 2,27*	*79,28 ± 1,31*	*99,99 ± 0,01*
	50	*100*	*58,05 ± 0,10*	*79,79 ± 0,07*	*96,29 ± 0,01*
		250	*59,47 ± 0,08*	*78,22 ± 0,06*	*98,79 ± 0,01*
		500	*59,15 ± 0,10*	*78,05 ± 0,07*	*99,72 ± 0,01*
		1000	*57,65 ± 0,12*	*77,17 ± 0,17*	*100,00 ± 0,00*
		2000	*55,96 ± 0,39*	*76,56 ± 0,26*	*100,00 ± 0,00*
		4000	*56,81 ± 0,58*	*77,14 ± 0,47*	*100,00 ± 0,00*
		8000	*60,63 ± 1,47*	*77,89 ± 0,71*	*99,99 ± 0,00*
		11812	*62,74 ± 2,26*	*78,96 ± 1,44*	*99,99 ± 0,01*
	100	*100*	*54,26 ± 0,08*	*75,26 ± 0,07*	*96,49 ± 0,01*
		250	*57,46 ± 0,09*	*76,93 ± 0,05*	*99,08 ± 0,01*
		500	*56,48 ± 0,09*	*76,30 ± 0,05*	*99,95 ± 0,01*
		1000	*54,27 ± 0,19*	*75,13 ± 0,14*	*100,00 ± 0,00*
		2000	*53,27 ± 0,30*	*74,38 ± 0,31*	*100,00 ± 0,00*
		4000	*54,26 ± 0,43*	*74,59 ± 0,40*	*100,00 ± 0,00*
		8000	*59,95 ± 1,45*	*77,43 ± 0,86*	*100,00 ± 0,00*
		11812	*62,43 ± 2,37*	*79,13 ± 1,19*	*99,99 ± 0,01*

continua

Tabela 8 conclusão

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Validação	*25*	*100*	$58{,}53 \pm 0{,}78$	$75{,}43 \pm 0{,}72$	$89{,}11 \pm 0{,}51$
		250	$63{,}05 \pm 0{,}79$	$77{,}60 \pm 0{,}70$	$92{,}42 \pm 0{,}36$
		500	$65{,}30 \pm 0{,}78$	$78{,}31 \pm 0{,}79$	$94{,}14 \pm 0{,}32$
		1000	$64{,}46 \pm 0{,}84$	$77{,}77 \pm 0{,}78$	$90{,}96 \pm 0{,}90$
		2000	$62{,}60 \pm 1{,}18$	$77{,}61 \pm 1{,}40$	$87{,}84 \pm 1{,}28$
		4000	$61{,}49 \pm 2{,}42$	$74{,}97 \pm 2{,}53$	$84{,}84 \pm 1{,}24$
		8000	$56{,}62 \pm 4{,}26$	$66{,}99 \pm 5{,}66$	$80{,}43 \pm 2{,}12$
		11812	$49{,}32 \pm 6{,}23$	$60{,}75 \pm 6{,}29$	$77{,}63 \pm 2{,}07$
	50	*100*	$57{,}10 \pm 0{,}75$	$79{,}02 \pm 0{,}59$	$92{,}82 \pm 0{,}29$
		250	$60{,}23 \pm 0{,}73$	$77{,}26 \pm 0{,}71$	$94{,}21 \pm 0{,}28$
		500	$61{,}37 \pm 0{,}80$	$78{,}06 \pm 0{,}76$	$94{,}15 \pm 0{,}33$
		1000	$61{,}90 \pm 0{,}72$	$77{,}97 \pm 0{,}85$	$91{,}60 \pm 0{,}91$
		2000	$61{,}00 \pm 1{,}32$	$77{,}85 \pm 0{,}99$	$88{,}59 \pm 1{,}03$
		4000	$60{,}95 \pm 1{,}67$	$75{,}47 \pm 2{,}61$	$85{,}96 \pm 1{,}74$
		8000	$55{,}81 \pm 5{,}23$	$68{,}16 \pm 4{,}05$	$82{,}18 \pm 2{,}08$
		11812	$48{,}26 \pm 5{,}82$	$60{,}25 \pm 6{,}10$	$79{,}23 \pm 3{,}53$
	100	*100*	$52{,}02 \pm 0{,}92$	$72{,}93 \pm 0{,}63$	$93{,}29 \pm 0{,}24$
		250	$58{,}59 \pm 0{,}69$	$76{,}02 \pm 0{,}69$	$95{,}23 \pm 0{,}18$
		500	$58{,}69 \pm 1{,}02$	$76{,}40 \pm 0{,}89$	$95{,}59 \pm 0{,}27$
		1000	$58{,}88 \pm 0{,}96$	$77{,}22 \pm 0{,}96$	$91{,}93 \pm 0{,}62$
		2000	$58{,}12 \pm 1{,}93$	$76{,}04 \pm 1{,}70$	$89{,}71 \pm 0{,}78$
		4000	$58{,}78 \pm 2{,}52$	$73{,}81 \pm 2{,}75$	$86{,}69 \pm 1{,}61$
		8000	$55{,}56 \pm 4{,}20$	$67{,}66 \pm 5{,}23$	$82{,}92 \pm 1{,}68$
		11812	$50{,}03 \pm 5{,}61$	$62{,}15 \pm 5{,}06$	$78{,}41 \pm 3{,}37$

Tabela 9 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para torneios com grupos condicionados aos cromossomos considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Estimação	*25*	*100*	*84,77 ± 0,06*	*89,29 ± 0,03*	*94,49 ± 0,02*
		250	*90,25 ± 0,06*	*93,30 ± 0,05*	*98,09 ± 0,02*
		500	*93,14 ± 0,10*	*95,50 ± 0,07*	*99,69 ± 0,01*
		1000	*96,29 ± 0,11*	*96,33 ± 0,09*	*100,00 ± 0,00*
		2000	*96,91 ± 0,19*	*97,40 ± 0,17*	*100,00 ± 0,00*
		4000	*96,74 ± 0,46*	*97,67 ± 0,24*	*100,00 ± 0,00*
		8000	*94,58 ± 1,06*	*97,55 ± 0,47*	*100,00 ± 0,00*
		11812	*90,20 ± 1,88*	*95,88 ± 1,01*	*99,99 ± 0,01*
	50	*100*	*89,55 ± 0,04*	*91,74 ± 0,03*	*96,89 ± 0,01*
		250	*94,08 ± 0,04*	*95,33 ± 0,04*	*98,65 ± 0,01*
		500	*96,20 ± 0,05*	*96,96 ± 0,05*	*99,73 ± 0,01*
		1000	*97,76 ± 0,05*	*98,66 ± 0,06*	*100,00 ± 0,00*
		2000	*99,05 ± 0,13*	*98,83 ± 0,14*	*100,00 ± 0,00*
		4000	*98,71 ± 0,19*	*98,72 ± 0,18*	*100,00 ± 0,00*
		8000	*95,39 ± 0,71*	*97,58 ± 0,44*	*99,99 ± 0,00*
		11812	*90,51 ± 2,03*	*96,02 ± 1,07*	*99,99 ± 0,01*
	100	*100*	*94,61 ± 0,01*	*95,55 ± 0,01*	*98,07 ± 0,01*
		250	*96,44 ± 0,03*	*96,76 ± 0,04*	*99,02 ± 0,01*
		500	*98,45 ± 0,05*	*97,87 ± 0,05*	*99,85 ± 0,01*
		1000	*99,46 ± 0,06*	*99,29 ± 0,05*	*100,00 ± 0,00*
		2000	*99,80 ± 0,06*	*99,58 ± 0,07*	*100,00 ± 0,00*
		4000	*99,43 ± 0,11*	*99,51 ± 0,10*	*100,00 ± 0,00*
		8000	*96,17 ± 0,88*	*98,38 ± 0,38*	*100,00 ± 0,00*
		11812	*91,30 ± 2,37*	*95,95 ± 0,89*	*99,99 ± 0,01*

continua

Tabela 9 conclusão

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Validação	*25*	*100*	*72,41 ± 0,89*	*80,82 ± 0,67*	*89,93 ± 0,33*
		250	*73,13 ± 1,09*	*80,75 ± 0,71*	*92,62 ± 0,41*
		500	*72,17 ± 1,22*	*78,91 ± 0,66*	*93,84 ± 0,34*
		1000	*71,47 ± 1,17*	*75,10 ± 0,77*	*90,82 ± 1,00*
		2000	*65,85 ± 1,42*	*72,04 ± 0,85*	*87,05 ± 1,27*
		4000	*56,48 ± 2,69*	*65,06 ± 2,07*	*85,03 ± 1,46*
		8000	*40,27 ± 4,48*	*52,53 ± 4,45*	*81,06 ± 2,37*
		11812	*24,63 ± 4,39*	*42,47 ± 4,75*	*77,21 ± 3,43*
	50	*100*	*79,90 ± 0,71*	*84,33 ± 0,46*	*93,77 ± 0,23*
		250	*79,54 ± 0,73*	*84,14 ± 0,46*	*93,96 ± 0,26*
		500	*77,58 ± 0,83*	*81,93 ± 0,57*	*94,30 ± 0,27*
		1000	*75,01 ± 0,94*	*80,32 ± 0,77*	*91,28 ± 0,75*
		2000	*71,64 ± 1,45*	*75,16 ± 1,47*	*88,65 ± 1,23*
		4000	*61,96 ± 1,94*	*67,83 ± 2,20*	*85,05 ± 1,71*
		8000	*42,73 ± 2,95*	*54,77 ± 2,90*	*82,34 ± 2,00*
		11812	*25,65 ± 4,41*	*43,03 ± 4,84*	*79,04 ± 2,76*
	100	*100*	*88,38 ± 0,39*	*90,74 ± 0,40*	*96,11 ± 0,17*
		250	*83,68 ± 0,79*	*86,20 ± 0,63*	*94,83 ± 0,23*
		500	*82,13 ± 0,87*	*83,35 ± 0,66*	*94,54 ± 0,33*
		1000	*79,22 ± 0,99*	*81,73 ± 0,84*	*91,05 ± 0,83*
		2000	*72,86 ± 2,29*	*76,23 ± 1,48*	*89,17 ± 0,93*
		4000	*63,05 ± 2,11*	*69,03 ± 2,71*	*86,97 ± 1,57*
		8000	*41,85 ± 3,77*	*54,90 ± 3,81*	*83,20 ± 2,17*
		11812	*27,64 ± 6,19*	*44,79 ± 4,86*	*78,61 ± 3,26*

Tabela 10 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para torneios com grupos condicionados aos cromossomos considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Estimação	*25*	*100*	*61,11 ± 0,11*	*78,27 ± 0,05*	*94,49 ± 0,02*
		250	*63,37 ± 0,09*	*79,53 ± 0,06*	*98,09 ± 0,02*
		500	*62,99 ± 0,13*	*79,37 ± 0,07*	*99,69 ± 0,01*
		1000	*60,32 ± 0,15*	*79,49 ± 0,10*	*100,00 ± 0,00*
		2000	*59,62 ± 0,37*	*78,88 ± 0,20*	*100,00 ± 0,00*
		4000	*59,86 ± 0,93*	*78,54 ± 0,42*	*100,00 ± 0,00*
		8000	*61,69 ± 1,61*	*78,46 ± 0,88*	*100,00 ± 0,00*
		11812	*63,38 ± 2,01*	*78,89 ± 1,28*	*99,99 ± 0,01*
	50	*100*	*58,84 ± 0,10*	*76,74 ± 0,05*	*96,89 ± 0,01*
		250	*60,30 ± 0,11*	*78,32 ± 0,06*	*98,65 ± 0,01*
		500	*58,96 ± 0,09*	*78,48 ± 0,05*	*99,73 ± 0,01*
		1000	*57,70 ± 0,10*	*76,77 ± 0,11*	*100,00 ± 0,00*
		2000	*55,79 ± 0,40*	*76,73 ± 0,33*	*100,00 ± 0,00*
		4000	*56,55 ± 0,54*	*77,12 ± 0,43*	*100,00 ± 0,00*
		8000	*61,15 ± 1,14*	*78,44 ± 0,78*	*99,99 ± 0,00*
		11812	*63,12 ± 2,15*	*78,80 ± 1,30*	*99,99 ± 0,01*
	100	*100*	*54,79 ± 0,06*	*75,49 ± 0,04*	*98,07 ± 0,01*
		250	*55,79 ± 0,06*	*77,32 ± 0,05*	*99,02 ± 0,01*
		500	*55,85 ± 0,08*	*77,15 ± 0,09*	*99,85 ± 0,01*
		1000	*54,32 ± 0,20*	*75,54 ± 0,12*	*100,00 ± 0,00*
		2000	*53,06 ± 0,36*	*74,88 ± 0,28*	*100,00 ± 0,00*
		4000	*54,61 ± 0,40*	*75,17 ± 0,40*	*100,00 ± 0,00*
		8000	*59,84 ± 1,42*	*77,36 ± 0,89*	*100,00 ± 0,00*
		11812	*62,09 ± 2,35*	*78,94 ± 1,22*	*99,99 ± 0,01*

continua

Tabela 10 conclusão

Tipo de Amostra	Tamanho do grupo	Nº de SNPs selecionados	h^2		
			0,25	*0,5*	*1,0*
Validação	*25*	*100*	*59,86 ± 0,91*	*76,45 ± 0,69*	*89,93 ± 0,33*
		250	*64,26 ± 0,94*	*78,17 ± 0,67*	*92,62 ± 0,41*
		500	*65,79 ± 1,03*	*78,32 ± 0,61*	*93,84 ± 0,34*
		1000	*64,80 ± 1,06*	*78,27 ± 0,78*	*90,82 ± 1,00*
		2000	*63,72 ± 1,47*	*77,69 ± 0,79*	*87,05 ± 1,27*
		4000	*61,72 ± 2,65*	*74,78 ± 1,99*	*85,03 ± 1,46*
		8000	*56,49 ± 5,11*	*67,14 ± 5,50*	*81,06 ± 2,37*
		11812	*49,47 ± 6,03*	*59,95 ± 6,20*	*77,21 ± 3,43*
	50	*100*	*58,14 ± 0,86*	*74,76 ± 0,66*	*93,77 ± 0,23*
		250	*62,12 ± 1,16*	*77,55 ± 0,62*	*93,96 ± 0,26*
		500	*61,58 ± 0,99*	*78,74 ± 0,56*	*94,30 ± 0,27*
		1000	*61,72 ± 1,07*	*77,72 ± 0,60*	*91,28 ± 0,75*
		2000	*61,93 ± 1,24*	*77,09 ± 1,38*	*88,65 ± 1,23*
		4000	*60,46 ± 2,32*	*75,53 ± 2,09*	*85,05 ± 1,71*
		8000	*57,60 ± 3,79*	*69,49 ± 3,41*	*82,34 ± 2,00*
		11812	*49,56 ± 5,89*	*59,63 ± 6,18*	*79,04 ± 2,76*
	100	*100*	*53,96 ± 0,99*	*74,71 ± 0,60*	*96,11 ± 0,17*
		250	*55,59 ± 1,14*	*76,73 ± 0,74*	*94,83 ± 0,23*
		500	*58,76 ± 1,03*	*77,21 ± 0,77*	*94,54 ± 0,33*
		1000	*59,27 ± 1,34*	*77,11 ± 0,90*	*91,05 ± 0,83*
		2000	*57,24 ± 1,99*	*76,06 ± 1,36*	*89,17 ± 0,93*
		4000	*58,74 ± 2,43*	*74,32 ± 3,11*	*86,97 ± 1,57*
		8000	*55,16 ± 5,03*	*68,38 ± 4,54*	*83,20 ± 2,17*
		11812	*48,31 ± 6,85*	*62,22 ± 3,68*	*78,61 ± 3,26*

Tabela 11 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e fenótipos para o Lasso Bayesiano considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de amostra	Nº de SNPs selecionados	h^2		
		0,25	*0,5*	*1,0*
Estimação	*100*	$59{,}13 \pm 0{,}20$	$87{,}22 \pm 0{,}05$	$82{,}67 \pm 0{,}07$
	250	$70{,}77 \pm 0{,}32$	$93{,}76 \pm 0{,}05$	$94{,}58 \pm 0{,}06$
	500	$79{,}60 \pm 0{,}37$	$97{,}93 \pm 0{,}05$	$99{,}20 \pm 0{,}04$
	1000	$81{,}95 \pm 0{,}63$	$99{,}79 \pm 0{,}04$	$99{,}99 \pm 0{,}01$
	2000	$84{,}21 \pm 1{,}07$	$99{,}98 \pm 0{,}01$	$100{,}00 \pm 0{,}00$
	4000	$87{,}83 \pm 1{,}62$	$99{,}96 \pm 0{,}02$	$100{,}00 \pm 0{,}00$
	8000	$89{,}45 \pm 1{,}73$	$99{,}11 \pm 0{,}18$	$100{,}00 \pm 0{,}00$
	11812	$90{,}29 \pm 1{,}90$	$95{,}66 \pm 1{,}08$	$99{,}98 \pm 0{,}01$
Validação	*100*	$28{,}85 \pm 2{,}25$	$76{,}48 \pm 0{,}78$	$68{,}80 \pm 0{,}95$
	250	$27{,}50 \pm 2{,}55$	$79{,}11 \pm 0{,}92$	$80{,}10 \pm 0{,}90$
	500	$32{,}10 \pm 2{,}13$	$81{,}64 \pm 0{,}85$	$86{,}03 \pm 0{,}70$
	1000	$29{,}15 \pm 2{,}26$	$81{,}37 \pm 1{,}37$	$83{,}07 \pm 1{,}88$
	2000	$26{,}62 \pm 2{,}50$	$76{,}35 \pm 1{,}86$	$83{,}45 \pm 1{,}85$
	4000	$26{,}17 \pm 3{,}11$	$70{,}68 \pm 2{,}42$	$81{,}50 \pm 2{,}08$
	8000	$24{,}03 \pm 3{,}75$	$57{,}10 \pm 3{,}41$	$78{,}41 \pm 2{,}83$
	11812	$23{,}60 \pm 3{,}71$	$46{,}40 \pm 5{,}32$	$77{,}66 \pm 3{,}10$

Tabela 12 Médias e desvios padrões das correlações (em porcentagem) entre GBVs preditos e simulados para o Lasso Bayesiano considerando diferentes herdabilidades, diferentes números de SNPs selecionados, amostras de estimação e validação, para 100 repetições de cada análise em um cenário de 250 SNPs com efeitos não nulos dispersos

Tipo de amostra	Nº de SNPs selecionados	h^2		
		0,25	*0,5*	*1,0*
Estimação	*100*	$56{,}82 \pm 0{,}29$	$75{,}13 \pm 0{,}07$	$82{,}67 \pm 0{,}07$
	250	$61{,}91 \pm 0{,}27$	$78{,}00 \pm 0{,}08$	$94{,}58 \pm 0{,}06$
	500	$63{,}87 \pm 0{,}26$	$76{,}87 \pm 0{,}07$	$99{,}20 \pm 0{,}04$
	1000	$65{,}48 \pm 0{,}25$	$73{,}87 \pm 0{,}22$	$99{,}99 \pm 0{,}01$
	2000	$66{,}42 \pm 0{,}56$	$72{,}72 \pm 0{,}12$	$100{,}00 \pm 0{,}00$
	4000	$65{,}01 \pm 1{,}42$	$72{,}90 \pm 0{,}20$	$100{,}00 \pm 0{,}00$
	8000	$64{,}20 \pm 1{,}68$	$76{,}09 \pm 0{,}49$	$100{,}00 \pm 0{,}00$
	11812	$63{,}23 \pm 1{,}64$	$79{,}14 \pm 1{,}35$	$99{,}98 \pm 0{,}01$
Validação	*100*	$45{,}83 \pm 1{,}64$	$71{,}09 \pm 0{,}71$	$68{,}80 \pm 0{,}95$
	250	$48{,}50 \pm 2{,}12$	$75{,}13 \pm 0{,}80$	$80{,}10 \pm 0{,}90$
	500	$52{,}17 \pm 1{,}73$	$75{,}97 \pm 0{,}70$	$86{,}03 \pm 0{,}70$
	1000	$53{,}74 \pm 2{,}16$	$74{,}90 \pm 1{,}24$	$83{,}07 \pm 1{,}88$
	2000	$54{,}60 \pm 2{,}70$	$72{,}37 \pm 2{,}34$	$83{,}45 \pm 1{,}85$
	4000	$53{,}19 \pm 3{,}46$	$71{,}92 \pm 2{,}56$	$81{,}50 \pm 2{,}08$
	8000	$50{,}82 \pm 4{,}41$	$69{,}47 \pm 3{,}51$	$78{,}41 \pm 2{,}83$
	11812	$48{,}64 \pm 4{,}70$	$63{,}76 \pm 5{,}53$	$77{,}66 \pm 3{,}10$

APÊNDICE D - Código R reproduzível correspondente à metodologia de torneios com grupos aleatórios

```
#============================================================
# Carregando a biblioteca multicore e configurando o número de cores que
# serão utilizados em paralelo
#============================================================
library(multicore)
ncores <- multicore:::detectCores()
options(cores=ncores)
getOption('cores')

#============================================================
# Simulando uma matriz de incidência de genótipos, um vetor de efeitos de
# marcadores e um vetor de fenótipos
#============================================================
set.seed(1000)
X <- matrix(sample(0:2,120000,replace=T),ncol=1200)

beta <- rep(0,ncol(X))
set.seed(1000)
beta[sample(1:ncol(X),10)] <- rep(1,10)

set.seed(1000)
y <- X%*%beta+rnorm(100)

#============================================================
# Número de marcadores selecionados ao final do torneio
#============================================================
nfinal <- 100

#============================================================
# Tamanho dos grupos
#============================================================
ng <- 25

#============================================================
# Função para executar as análises em cada grupo do torneio: determina o
# marcador com maior valor-p do grupo i ou o que não pôde ser estimado
#============================================================
torneio <- function(i){
  lmi <- lm(y~-1+X[,grupos[[i]]])
  nas <- as.numeric(which(is.na(lmi$coef)))
  if(length(nas)==0){
    pval <- as.numeric(summary(lmi)$"coefficients"[,4])
    coef <- as.numeric(summary(lmi)$"coefficients"[,1])
```

```
  }
  if(length(nas)>0){
    semnas <- c(1:length(lmi$coef))[-nas]
    pval1 <- as.numeric(summary(lmi)$"coefficients"[,4])
    pval <- rep(1,length(grupos[[i]]))
    k <- 1
    for(j in semnas) {
      pval[j] <- pval1[k]
      k <- k+1
    }
    coef1 <- as.numeric(summary(lmi)$"coefficients"[,1])
    coef <- rep(0,length(grupos[[i]]))
    k <- 1
    for(j in semnas) {
      coef[j] <- coef1[k]
      k <- k+1
    }
  }
  names(pval) <- grupos[[i]]
  names(coef) <- colnames(X[,as.numeric(names(pval))])
  max <- as.numeric(which(pval==max(pval)))
  if(length(max)>1) max <- sample(max)[1]
  mpval <- pval[max]
  mcoef <- coef[max]
  c(mcoef,mpval)
}

#============================================================
# Função para gerar grupos aleatórios
#============================================================
geragrupos <- function(pos){
  if((p/ng-trunc(p/ng))==0) ngrupos <- round(p/ng,0)
  if((p/ng-trunc(p/ng))>0.5) ngrupos <- round(p/ng,0)
  if((p/ng-trunc(p/ng))<=0.5 & (p/ng-trunc(p/ng))!=0) ngrupos <- round(p/ng,0)+1
  snps <- sample(pos)
  if(ngrupos>1) {indgrupos <- rep(1:ngrupos,each=ng)[1:length(snps)]}
  if(ngrupos==1) {indgrupos <- rep(1,length(snps))}
  grupos <- tapply(snps,indgrupos,sort)

  # Se o número de marcadores do último grupo for menor que ng divide os
  # marcadores dele entre osoutros grupos
  grupok <- grupos[[dim(grupos)]]
  k2 <- 1
  if(length(grupok)<ng){
    for(i2 in 1:length(grupok)){
      grupos[[k2]] <- sort(c(grupos[[k2]],grupok[i2]))
```

```
      if(k2<length(grupos)) k2 <- k2+1
      if(k2==length(grupos)) k2 <- 1
    }
    grupos <- grupos[-dim(grupos)]
  }

  # Quando sobrar apenas 1 grupo e o numero de elementos restantes for maior
  # que ng divide este grupo em dois
  if(length(grupos)==1 & length(grupos[[1]])>ng){
    g <- grupos[[1]]
    grupos[[1]] <- g[1:round((length(g)/2),0)]
    grupos[[2]] <- g[(round(length(g)/2,0)+1):length(g)]
  }
  grupos
}

#============================================================
# Criando objetos auxiliares
#============================================================
p <- ncol(X)
posicao <- 1:p
pos <- posicao
verifica <- list("Iteracao"=0,"ngrupos"=0)
excluir <- 0
pval <- 0
coef <- 0
contagem <- 1

#============================================================
# Gerando os grupos para iniciar o torneio
#============================================================
grupos <- geragrupos(pos)
ngrupos <- length(grupos)

#============================================================
# Torneio (elimina marcadores até que restem "nfinal" marcadores)
#============================================================
while(length(pos)>nfinal){

  # Rodando a função torneio em paralelo: analiza cada grupo em um core
  result <- mclapply(1:ngrupos,torneio)

  # OBS: Caso não tenha a biblioteca multicore instalada ou esteja utilizando
  # sistema operacional Windows pode executar o torneio sem paralelização
  # substituindo a função "mclapply()" pela função "lapply()" no comando anterior
```

```
  #Atualiza pvalcoef e excluir a partir de result
  pval <- c(pval,unlist(result)[seq(2,length(unlist(result)),2)])
  coef <- c(coef,unlist(result)[seq(1,length(unlist(result)),2)])
  excluir <- c(excluir,sort(as.numeric(names(unlist(result)[seq(2,length(unlist(
          result)),2)]))))

  # Verifica o número de iterações e o número de grupos em cada iteração
  verifica$Iteracao <- contagem
  verifica$ngrupos <- c(verifica$ngrupos,ngrupos)

  # Prepara para a próxima etapa do loop (atualiza "pos")
  pos <- posicao[-excluir]

  # Seleciona novos grupos para a próxima iteração do torneio
  snps <- sample(pos)
  p <- length(pos)
  grupos <- geragrupos(pos)
  ngrupos <- length(grupos)

  # Atualiza a contagem (numero de iterações)
  contagem <- contagem+1
}

#============================================================
# Ajustando objeto verifica$ngrupos
#============================================================
if(verifica$ngrupos[1]==0) verifica$ngrupos <- verifica$ngrupos[-1]

#============================================================
# SNPs selecionados
#============================================================
selecionados <- sort(as.numeric(unlist(grupos)))

#============================================================
# Classificação dos SNPs
#============================================================
classificacao <- c(selecionados,rev(excluir[-1]))

#============================================================
# Criando um objeto contendo os resultados do torneio
#============================================================
TS <- list()
TS[[1]] <- nfinal
TS[[2]] <- ng
TS[[3]] <- verifica
TS[[4]] <- selecionados
```

```
TS[[5]] <- excluir[-1]
TS[[6]] <- classificacao
names(TS) <- c("nfinal","ng","verifica","selecionados","excluidos","classificacao")
TS

#============================================================
# Verificando os marcadores selecionados
#============================================================
plot(beta,pch=19,axes=F);axis(2)
abline(v=TS$selecionados)
```