

**UNIVERSIDADE ESTADUAL PAULISTA “JULIO DE MESQUITA FILHO”
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS**

CAMPUS DE JABOTICABAL

**Acurácia de predição do valor genético genômico em bovinos da raça Gir para
produção de leite até os 305 dias de lactação.**

Thomaz Marques Sena

Zootecnista

**JABOTICABAL-SP
2023
UNIVERSIDADE ESTADUAL PAULISTA “JULIO DE MESQUITA FILHO”
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CAMPUS DE JABOTICABAL**

**Acurácia de predição do valor genético genômico em bovinos da raça Gir
para produção de leite até os 305 dias de lactação.**

ALUNO: Thomaz Marques Sena

ORIENTADOR: Prof. Dr. Danísio Prado Munari

COORIENTADOR: Dr. Marcos Vinícius Gualberto Barbosa da Silva

COORIENTADOR: Dra. Tatiane Cristina Seleguim Chud

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias - Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Genética e Melhoramento Animal.

**JABOTICABAL-SP
2023**

S474a	Sena, Thomaz Marques Acurácia de predição do valor genético genômico em bovinos da raça Gir para produção de leite até os 305 dias de lactação. / Thomaz Marques Sena. -- Jaboticabal, 2023 69 p. Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal Orientador: Danísio Prado Munari Coorientador: Marcos Vinícius Gualberto Barbosa da Silva 1. Genética. 2. Animais Melhoramento genetico. 3. Inteligência artificial. 4. Machine learning. 5. Inteligência artificial. I. Título.
-------	---

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ACURÁCIA DE PREDIÇÃO DO VALOR GENÉTICO GENÔMICO EM BOVINOS DA RAÇA GIR PARA PRODUÇÃO DE LEITE ATÉ OS 305 DIAS DE LACTAÇÃO

AUTOR: THOMAZ MARQUES SENA

ORIENTADOR: DANISIO PRADO MUNARI

COORDENADORA: TATIANE CRISTINA SELEGUIM CHUD

COORDENADOR: MARCOS VINÍCIUS GUALBERTO BARBOSA DA SILVA

Aprovado como parte das exigências para obtenção do Título de Doutor em Genética e Melhoramento Animal, pela Comissão Examinadora:

Documento assinado digitalmente
 **DANISIO PRADO MUNARI**
Data: 03/01/2024 18:24:22 -0300
Verifique em <https://validar.it.gov.br>

Prof. Dr. DANISIO PRADO MUNARI (Participação Virtual)
Departamento de Engenharia e Ciências Exatas DECEX / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **SALVADOR BOCCALETTI RAMOS**
Data: 20/12/2023 09:06:40 -0300
Verifique em <https://validar.it.gov.br>

Prof. Dr. SALVADOR BOCCALETTI RAMOS (Participação Virtual)
Departamento de Engenharia e Ciências Exatas DECEX / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **SABRINA LUZIA CAETANO**
Data: 19/12/2023 10:26:49 -0300
Verifique em <https://validar.it.gov.br>

Profa. Dra. SABRINA LUZIA CAETANO (Participação Virtual)
Departamento de Engenharia e Ciências Exatas DECEX / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **MARCOS ELI BUZANSKAS**
Data: 19/12/2023 10:01:47 -0300
Verifique em <https://validar.it.gov.br>

Prof. Dr. MARCOS ELI BUZANSKAS (Participação Virtual)
Departamento de Zootecnia / Universidade Federal da Paraíba - Areia/PB

Dra. ANA PAULA SBARDELLA (Participação Virtual)
Criaris Agropecuaria Inteligente Ltda / Goiânia/GO
 **ANA PAULA SBARDELLA**
Data: 19/12/2023 17:37:37 -0300
Verifique em <https://validar.it.gov.br>

Jaboticabal, 27 de outubro de 2023

DADOS CURRICULARES DO AUTOR

Thomaz Marques Sena — Nascido em 03 de fevereiro de 1992 na cidade de Tabatinga – SP. Em março de 2012, ingressou no curso de Zootecnia da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de Jaboticabal. Durante a graduação, foi bolsista da FAPESP de iniciação científica sob orientação do Professor Dr. Ricardo Andrade Reis, por duas vezes consecutivas na área de nutrição de ruminantes e conservação de plantas forrageiras, respectivamente, e foi membro voluntário e bolsista do Programa de Educação Tutorial (PET Zootecnia) trabalhando com ensino, pesquisa e extensão universitária. Desenvolveu seu Trabalho de Conclusão de Curso no ano de 2017 sob orientação do Prof. Dr. Humberto Tonhati e concluiu a graduação no mesmo ano. Em agosto de 2017, ingressou curso de Mestrado em Genética e Melhoramento Animal sob orientação do Prof. Dr. Humberto Tonhati na área de melhoramento genético de bubalinos leiteiros, durante esse período foi bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), encerrando o mestrado em julho de 2019. Em agosto de 2019 ingressou no curso de Doutorado em Genética e Melhoramento Animal sob orientação do Prof. Dr. Danísio Prado Munari e sob coorientação do Dr. Marcos Vinícius Gualberto Barbosa da Silva e da Dra. Tatiane Cristina Seleguim Chud. Durante o curso de doutorado, foi bolsista CAPES

Tomei a decisão de fingir que todas as coisas que até então haviam entrado na minha mente não eram mais verdadeiras do que as ilusões dos meus sonhos.

René Descartes

Dedico

*Aos meus pais, Ivanilde Machado
Marques Sena e Irso Marques Sena, por
todo o amor e carinho, à minha avó Onofra
Peixoto (in memoriam) por suas incríveis
histórias e amor e à minha esposa, Caroline
Alves Garcia, por todo amor, apoio e carinho.*

AGRADECIMENTOS

Primeiramente à Deus, por todas as oportunidades, por ter guiado meus pensamentos e passos, e me abençoado em cada decisão. Aos meus pais, Ivanilde e Irso por tudo que fizeram e ainda fazem por mim. O que me tornei se deve por todas as lições de vida e sacrifícios que fizeram por mim. Passamos por vários perrengues, mas vocês sempre me ajudaram a superar. Obrigado por seus ensinamentos, exceto os que se referiam aos de cura de umbigo, esses tive que reaprender.

À minha esposa Caroline Alves Garcia por todos os momentos incríveis que passamos juntos. Obrigado por me apoiar durante toda minha trajetória na UNESP FCAV, iniciada em 2012. Eternas boas lembranças cultivamos nas salas e laboratórios dessa universidade que se tornou nossa segunda casa durante esse tempo. Obrigado por sempre estar ao meu lado e acreditar em mim. Obrigado por ser minha linda e incrível esposa que eu tanto amo e admiro.

Agradeço aos meus queridos peludos, embora eles não possam ver o que escrevo. Agradeço ao Thor, Pina, Moana e Maui, que me fizeram companhia e me “ouvirem” nos momentos de trabalho remoto, e que consegue alegrar meu dia sempre. Aos meus sogros Vera Lúcia e Ricardo, por terem colocado no mundo a mulher mais linda e incrível que eu já conheci na vida, por estarem sempre nos ajudando e por aquelas horas de conversa com a minha mãe em lei (sogra).

Agradeço ao Prof. Dr. Danísio Prado Munari pela oportunidade de poder trabalhar e aprender sobre o fazer científico e o ser professor. Agradeço, também, ao Prof. Dr. Nelson Peruzzi, pelos ensinamentos e exemplo nesses 10 anos que nos conhecemos; o senhor fez com que eu visse a beleza/arte da Matemática.

Agradeço a Embrapa Gado de Leite pela disponibilidade dos dados e confiança em poder desenvolver esse trabalho em conjunto, em especial ao Dr. Marcos Vinicius Gualberto Barbosa da Silva pela orientação, e a Dra. Tatiane Cristina Seleguim Chud pela orientação e ensinamentos passados. “O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de financiamento 001.”

ACURÁCIA DE PREDIÇÃO DO VALOR GENÉTICO GENÔMICO EM BOVINOS DA RAÇA GIR PARA PRODUÇÃO DE LEITE ATÉ OS 305 DIAS DE LACTAÇÃO.

RESUMO — Estudos para a incorporação de dados moleculares em conjunto com a avaliação genética animal têm sido propostos com os objetivos de aumento da acurácia na predição de valores genéticos genômicos (GEBV), aumento da intensidade de seleção e diminuição dos intervalos de gerações, visando aumentar o ganho genético pela seleção ao longo de sucessivas gerações de acasalamento. O uso de painéis customizados de nucleotídeos de polimorfismo único (SNP) de baixa densidade se mostra interessante devido ao menor custo em relação aos painéis de SNP de alta densidade. Dessa forma, um maior número de animais candidatos a seleção podem ser genotipados. Para a seleção dos SNP que compõem os painéis customizados o algoritmo de aprendizado de máquina tem demonstrado bons resultados em pesquisa. No entanto, com painéis de baixa densidade de SNP, dificuldades de convergência podem ocorrer nas análises de predição do GEBV. Desse modo, o objetivo desse estudo foi verificar a habilidade de predição do GEBV em bovinos da raça Gir a partir de subconjunto de SNP, classificado segundo sua importância para a produção de leite até os 305 dias de lactação. Os SNP foram classificados por regressão não paramétrica por meio do algoritmo Random Forest (RF) e foram testados 20 cenários, os quais variaram quanto à proporção da matriz A_{22} dentro da G e A_{22}^{-1} dentro da matriz H^{-1} , buscando-se verificar o impacto do subconjunto de SNP na acurácia e no viés de predição. A presença de genes em uma janela de 1Mb foi verificada usando os pacotes biomaRt, com a versão 110 de Ensembl Genomes, clusterProfiler do software R. Dentre os cenários de RF, o melhor foi aquele em que os grupos de treinamento e validação foram separados de acordo com as gerações, cujo $Ntree$ e $Mtry$ foram 8192 e 0,3 vezes no número de SNP presentes no painel de 400 mil marcadores, respectivamente. Os SNP com maior relevância apresentaram oito genes relacionados a importantes vias metabólicas e funções biológicas, presentes em várias espécies de animais. Para os diferentes cenários testados observou-se diferenças significativas (p -valor < 0,05) entre estimativas da variância e da média do GEBV, do fenótipo predito ($YHat$) e do valor dos resíduos que representam o valor fenotípico descontado dos efeitos incluídos no modelo) para os diferentes cenários de ponderação da matriz G e A_{22}^{-1} testados e para os diferentes painéis de customizados. Esse resultado mostrou-se promissor, uma vez que é possível criar painéis de SNP com menores densidades sem perder o poder de predição do GEBV.

Palavras-chave: *Bos taurus indicus*, Machine Learning, ssGBLUP, ssGWAS

Sumário

1. Introdução	1
2. Objetivos	3
3. Revisão de Literatura	5
3.1. Produção de leite	5
3.2. Estimação de valores genéticos genômicos	6
3.3. Método Random Forest (RF) de aprendizado de máquina	12
4. Material e métodos	15
4.1. Registros de pedigree e fenotípicos	15
4.2. Genótipos e controle de qualidade	16
4.3. Classificação dos SNP pelo método “Random Forest” (RF)	17
4.4. Habilidade de predição dos valores genéticos genômicos via ssGBLUP e ssGWAS	18
4.5. Cenários	21
5. Resultados e Discussão	22
6. Conclusão	32
7. Referências bibliográficas	33
8. APENDICÊS	42
8.1. APENDICE A.....	43
8.2. APÊNDICE B.....	63

1. Introdução

Atualmente o setor de produção de leite passa por um momento de crise no Brasil, devido à alta no preço dos insumos em anos anteriores, aliado a baixa competitividade do leite brasileiro frente ao mercado externo. Isso favoreceu a importação de leite em pó e promoveu redução de oferta de leite inspecionado no mercado interno no ano de 2021. Porém, houve aumento na oferta do produto na região Sul, que respondia por 33% do leite brasileiro, passando para 39% em 2021 (Anuário do leite, 2022). Na região Nordeste, que embora no período entre 2021 e 2022 teve aumento pouco expressivo na produção de leite, o aumento no período de 2013 a 2022 foi de 63% (Anuário do leite, 2022; Anuário do leite, 2023).

Há aproximadamente 800 raças de gado doméstico que são divididas em *Bos taurus tauruse* *Bos taurus indicus* (Feliuss1985). Presente em quase todo o Brasil, a raça Gir (*Bos taurus indicus*) representa aproximadamente 80% de rebanhos leiteiros, somando os animais puros da raça com os cruzados com animais da raça Holandesa, uma vez que esses animais podem ser criados em pastagem, possuem resistência aos endo e ectoparasitas e à alta temperatura, com produção máxima de leite de 14 kg de leite por dia (Pereira et al., 2012), representando mais de 80% do leite produzido no Brasil (Embrapa, 2015).

O Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL) da ABCGIL/EMBRAPA foi implantado em 1985, por meio da identificação e seleção de touros geneticamente superiores para as características de produção (leite, gordura, proteína, lactose e sólidos totais), conformação e manejo (Martinez et al., 2005), sendo a primeira raça leiteira do Brasil a participar de programas de melhoramento genético e ter touros provados para leite pelo teste de progênie (Verneque et al. 2000). Avaliações genéticas de características produtivas em bovinos Gir leiteiros no Brasil são conduzidos utilizando modelo animal para estimação de parâmetros genéticos de produção de leite até os 305 dias de lactação (Panetto et al., 2018).

A integração de informações genômicas nas avaliações genéticas, através do emprego do modelo animal, proporcionou uma melhoria significativa na precisão das

estimativas dos valores genéticos dos animais. Nesse contexto, a utilização da matriz de relacionamento genômico (Matriz G) desempenhou um papel crucial ao contribuir para a redução do intervalo de gerações., uma das razões fundamentais para sua implementação na bovinocultura de leite (Schaeffer, 2006). Isso possibilitou maiores ganhos genéticos. Com dados genômicos, também foi possível realizar análises de associação genômica ampla (“Genome-wide association study” - GWAS), para determinar regiões do genoma responsáveis pela variação genética aditiva de características de interesse econômico (Hayes & Goddard, 2010).

Um fator limitante para os painéis de genotipagem é a necessidade do desequilíbrio de ligação (DL) entre os marcadores genéticos, como por exemplo SNP, e o QTL (Quantitative Trait Locus), o que diminui a capacidade de identificação de variações causais no genoma (Rivas et al., 2011). O mesmo ocorre com alelos em baixa frequência, pouco presentes em painéis de marcadores “single nucleotide polymorphisms” (SNP) de baixa densidade e responsáveis por parte da variância genética aditiva em características quantitativas ou até mesmo causadores de doenças que atingem os rebanhos (Larmer, 2016).

Existem vários painéis de genotipagem disponíveis no mercado, variando desde painéis de baixa densidade, com cerca de 3 mil marcadores do tipo SNP, até painéis de alta densidade, como os HD com aproximadamente 777 mil marcadores. No entanto, uma grande quantidade de SNP presentes nos painéis de alta densidade apresenta pouco ou nenhum efeito sobre os fenótipos (MacLeod et al., 2016). Pensando nisso, há a possibilidade de se trabalhar com painéis customizados de baixa densidade de marcadores para avaliação genética dos animais, com pouco efeito na acurácia de predição do GEBV (Habier et al., 2009; Ogawa et al., 2014, Li et al., 2018; Xu et al., 2021).

A fim de promover a redução de custos, a imputação é uma ferramenta que possibilita a incorporação de dados de genotipagem obtidos em marcadores SNP de menores para maiores densidades. A redução de custos ocorre porque a maioria dos animais do rebanho pode ser genotipada com painéis de SNP com menor densidade, a custo menor. Com o genótipo dos animais imputados para painéis de SNP em alta

densidade, foi possível implementar a seleção genômica de forma mais eficiente e aumentar o progresso genético, especialmente na bovinocultura de leite (Hayes et al., 2009; Panetto et al. 2018). No entanto, como em qualquer método de predição, o uso da informação genômica imputada está sujeita aos erros de predição, uma vez que os métodos de imputação são influenciados pela estrutura da população, pelo tamanho efetivo da população referência, número de SNP no painel com baixa densidade, frequência alélica dos marcadores, nível de parentesco entre as populações referência e aquela que será imputada e pelas metodologias empregadas no procedimento de imputação (Hayes et al., 2012; Pausch et al., 2013), influenciando na matriz de SNP utilizada no cálculo da matriz de parentesco genômico.

Os modelos estatísticos empregados na avaliação genética sofrem influências de inúmeros fatores inerentes ao próprio genoma e inerentes aos métodos que utilizamos para sua obtenção. Assim, no intuito de minimizar os erros e encontrar um modelo que nos seja útil, vários métodos de predição de valores genéticos genômicos (GEBV) foram propostos (Meuwissen et al., 2001; Gianola et al., 2009). No entanto, o método “single-step” GBLUP (ssGBLUP), proposto por Aguilar et al. (2010) e Misztal et al. (2009), tem sido utilizado por combinar as informações dos animais genotipados e não genotipados em uma só matriz, denominada de matriz H.

As matrizes A e G, que compõem a matriz H no método ssGBLUP possuem divergências que se agravam com a redução do tamanho do painel de genotipagem, como os empregados nos painéis customizados e ou de baixa densidade, e devido possíveis erros no pedigree. Alguns parâmetros foram propostos para corrigir tais incompatibilidades, na tentativa de reduzir seus efeitos negativos na acurácia de predição e no coeficiente de determinação do modelo de ajuste (Chen et al., 2011; Vitezica et al., 2011; Tsuruta et al, 2011; Misztal et al., 2017; Hollifield et al., 2022; McWhorter et al, 2023), sendo ajustados de acordo com a população.

2. Objetivos

Objetivou-se verificar como diferentes subconjuntos de SNP mais relevantes para produção de leite até 305 dias de lactação de bovinos da raça Gir, combinados

com diferentes proporções do efeito poligênico residual dentro da matriz genômica ou combinados com diferentes proporções da inversa da matriz de parentesco genético aditivo dentro da matriz H, influenciam na acurácia e viés de predição do valor genético genômico.

3. Revisão de Literatura

3.1. Produção de leite

A quantidade de leite produzido e sua qualidade são influenciadas por fatores ambientais como a disponibilidade de alimento ao longo das estações do ano e por fatores genéticos que influenciam na fisiologia do animal. A escolha da raça é um dos determinantes quando se produz leite em um clima tropical, devido as altas temperaturas e períodos de secas. A raça Gir, originária da Índia (península de Kathiawar), localidade cujo clima é predominantemente tropical de monção, adaptou-se muito bem ao Brasil desde sua importação, que compreendeu os anos de 1870 a 1962. Nesse período, foram trazidos cerca de 6.000 zebuínos, sendo menos de 700 bovinos Gir. Assim, o gado zebuínio e o cruzamento desse com animais criolos se difundiu rapidamente no país (O'Brien et al., 2015).

A vaca Gir tem boa produtividade de leite, com média de produção de leite em 305 dias de lactação (P305) de $2907,47 \pm 1112,80$ kg segundo Lagrota et al., (2010). De acordo com Balieiro et al. (2003), a média para P305 na raça Gir foi de 2.653,58 kg. Ribeiro et al. (2009), trabalhando com dados de animais Gir oriundos do estado do Rio Grande do Norte, encontraram médias para P305 de $2114,128 \pm 625,88$ kg na primeira lactação e, nas vacas Gir com mais de uma lactação, a média foi de 2.575,1 kg. Conforme Panetto et. al. (2023), foi observado média e respectivo desvio padrão da produção de leite até 305 dias de lactação de 3719 ± 2102 kg, da produção de gordura de 128 ± 77 kg, da produção de proteína de 119 ± 72 kg e dos sólidos totais equivalentes a 480 ± 280 kg, com duração média da lactação foi de 273 ± 83 dias. Esses valores divulgados em literatura são superiores ao valor de produção nacional expressos em litros/vaca/ano que é próximo a 2200 Kg (Anuário do leite, 2023). A

busca por maior produtividade e a melhoria da eficiência técnica e econômica da atividade leiteira impulsionou pesquisas nas áreas de manejo nutricional das pastagens, novas cultivares de capim e a seleção de animais superiores, sob a era da seleção genômica.

Além da boa produtividade do Gir, outro fator importante é a menor exigência de energia metabolizável da raça em relação aos animais cruzados com a raça Holandesa (F1) e em relação ao sugerido pelas tabelas internacionais (Borges, 2016), que em geral baseia-se em animais Taurinos. A menor exigência nutricional permite que esses animais possam ser utilizados em sistemas produtivos a pasto ou pouco intensificados, o que justifica a alta proporção de leite oriunda desses animais no Brasil.

3.2. Estimação de valores genéticos genômicos

A metodologia dos modelos mistos que Henderson (1975) desenvolveu possibilitou a seleção de animais com base na estimação acurada de seus valores genéticos, permitindo a seleção para características quantitativas de importância econômica em plantas e animais. O modelo proposto para o computo do valor genético dos animais, antes do surgimento da informação genômica, utilizava a matriz A. Calculada como descrito por Henderson (1976), a matriz A contém os coeficientes de parentesco genético aditivo entre os animais, levando em consideração as anotações de pedigree.

Quando o conceito de seleção genômica proposto por Meuwissen et al. (2001) foi definido, ferramentas para obtenção da informação genômica a baixo custo ainda não havia sido desenvolvidas. Com os avanços das tecnologias genômicas e dos métodos estatísticos para estimar o efeito dos marcadores moleculares tornou-se possível a implantação da seleção genômica (Jiang et al., 2013; Feuk et al., 2006). A implementação da seleção genômica em outros países tornou-se uma importante ferramenta na cadeia produtiva de bovinos de leite, principalmente pela considerável diminuição no intervalo de geração (Schaeffer, 2006), além dos ganhos na acurácia de seleção quando comparado a seleção tradicional (Preisinger, 2012; Tribout et al., 2013).

Vários métodos de predição de valores genéticos genômicos foram propostos, podendo ser citados os métodos GBLUP, Bayes A, Bayes B e outros métodos bayesianos (Meuwissen et al., 2001; Gianola et al., 2009). A diferença entre essas abordagens está na suposição da distribuição *a priori* considerada para estimar os efeitos dos SNP. O método GBLUP assume o modelo infinitesimal com muitos *loci* de pequenos efeitos. O método Bayes A assume que poucos genes têm grandes efeitos e muitos genes apresentam pequenos efeitos. No método Bayes B, muitos efeitos de SNP são assumidos como zero *à priori*, reduzindo o tamanho do genoma devido a concentração nas partes em que existem QTL (“Quantitative trait loci”). De acordo com Resende et al. (2008), o melhor método predição de valores genéticos genômicos pode variar de acordo com a característica estudada, pois este deve levar em consideração a natureza biológica do caráter.

O método “single-step” GBLUP (ssGBLUP) proposto por Aguilar et al. (2010) e Misztal et al. (2009) leva em consideração as informações de parentesco genômico (Matriz G) e o parentesco genético aditivo (matriz A), combinados na matriz H e o desempenho produtivo dos animais genotipados e não genotipados da população (Chistensen e Lund., 2010).

A superioridade da seleção genômica em relação a tradicional, de acordo com Resende et al. (2012), pode ser atribuída aos seguintes fatores: uso da matriz de parentesco genômico (depende do número de SNP presente no painel) e própria de cada caráter; uso maior de informações, combinando três tipos de informação (fenotípica, genotípica e de parentesco genético aditivo), que pode aumentar a acurácia e a seleção precoce dos animais, o que aumenta o ganho genético por tempo. No entanto, os métodos de seleção genômica pode ser influenciados pela densidade dos marcadores no painel, número de animais com informações de genômicas, o grau de parentesco entre os animais da população de treinamento e a população de predição e a metodologia de predição dos valores genéticos genômicos (Akanno et al., 2014).

No tocante ao uso painéis destinados a seleção de um único caráter, o uso de painéis customizados associados a um fenótipo pode ser uma possível solução econômica para a seleção genômica (Habier et al., 2009; Ogawa et al., 2014). Na

seleção genômica, o uso de painéis de alta densidade é promissor para melhorar a resposta à seleção em populações sujeitas à seleção artificial. No entanto, a genotipagem de painéis de alta densidade de todos os candidatos a seleção em cada geração possui custo elevado. Painéis de SNP customizados, contendo SNP que mostram forte associação com o fenótipo podem ser usados, mas para isto são necessários painéis separados para cada característica ou conjunto de características, e cada população.

Considerando os elevados custos com a obtenção de genoma e o fato de grande quantidade de SNP presentes nos painéis de SNP de alta densidade apresentarem pouco ou nenhum efeito sobre os fenótipos (MacLeod et al., 2016), algumas metodologias de Machine Learning (ML) vêm sendo exploradas para a classificação de SNP segundo sua relevância em características de interesse econômico, visando elaborar painéis customizados de baixa densidade de marcadores.

Em bovinos da raça Brahman, os 3.000 SNP mais relevantes para peso corporal ou aqueles com valores positivos de importância, selecionados por método Random Forest (RF), Gradient Boosting Machine (GBM) e Extreme Gradient Boosting Method (XgBoost), estavam envolvidos principalmente no desenvolvimento do sistema nervoso, na percepção visual e na atividade celular; constatou-se ainda que a maioria dos SNPs tinha um pequeno efeito ou nenhum os valores de importância em RF, seguindo o mesmo padrão dos demais. Embora esses SNP com maiores efeitos estejam em menor frequência alélica e distribuídos de maneira desuniforme no genoma, estes estão relacionados com importantes vias metabólicas, portanto podem estar fixados em diferentes raças. Além disso, a acurácia de predição genômica usando 3.000 SNP classificados pelos métodos RF, GBM e XgBoost foi 0,41, 0,46 e 0,33, respectivamente, sendo próximos ao valor da acurácia quando utilizado o painel completo (0,425), enquanto para os 3.000 SNP uniformemente espaçados foi encontrada acurácia de 0,29 (Li et al., 2018).

As matrizes A e G, que compõem a matriz H no método ssGBLUP, apresentam divergências, uma vez que a primeira é obtida como duas vezes a média esperada de genes idênticos por descendência (IBD) compartilhada entre dois indivíduos parentes

(Malécot, 1948). A segunda é idêntica por estado (IBS), e as frequências alélicas da população base nas duas matrizes são diferentes (VanRaden, 2008). Essas divergências se agravam com a redução do tamanho do painel de genotipagem, como os empregados nos painéis customizados, e devido possíveis erros no pedigree. Alguns parâmetros foram propostos para corrigir tais incompatibilidades, na tentativa de reduzir seus efeitos negativos na acurácia de predição e no coeficiente de determinação do modelo de ajuste (Chen et al., 2011; Vitezica et al., 2011; Tsuruta et al, 2011; Misztal et al., 2017; Hollifield et al., 2022; McWhorter et al, 2023), sendo parâmetros ajustados de acordo com a população.

A incompatibilidade entre as matrizes A e G resulta do fato de que uma parte dos efeitos poligênicos tem componentes que não são explicados pelos SNP, a frequência alélica da população base ser diferente e pela combinação de populações de animais genotipados e não genotipados. Por exemplo: consideremos dois animais i e j sem relações de parentesco. Em sua matriz A, o valor atribuído é 0, no entanto algum valor diferente de zero será atribuído na matriz G, pois esses animais possuem SNP devido a um ancestral em comum.

Para melhor entendimento de como esses fatores estão relacionados, tomaremos como base o seguinte modelo estatístico:

$$y = Xb + Wp + Za + Zg + e$$

em que em y é o vetor de observações fenotípicas; X é a matriz de incidência dos efeitos fixos (b), Z é a matriz de incidência dos efeitos aleatórios genéticos aditivos explicados pelos SNP (a) e dos efeitos poligênicos residuais não explicados pelos SNP (g); W é a matriz de incidência do efeito aleatório de ambiente permanente (p); e é o vetor de efeitos residuais. Vamos assumir que α é a proporção da variância explicada pelos SNP, então, a variância genética aditiva total (u_g) é $u_g = a + g$. De modo que: $\text{Var}(u_g) = \alpha G\sigma_g^2 + (1-\alpha)A_{22}$, então $G_{\text{blending}} = \alpha G + \beta A_{22}$, sendo $\beta = 1 - \alpha$, o efeito poligênico residual considerado na matriz G (Legarra et al., 2018; VanRaden, 2008).

Vamos definir A como sendo a matriz de parentesco genético aditivo, obtido pelas anotações de pedigree e calculado de acordo com Henderson (1976), e suas

partições A_{ij} , onde $i, j \in \{1, 2\}$, são também matrizes que possuem as mesmas estruturas que a matriz A , ou seja, esparsas: $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, cuja inversa é $A^{-1} = \begin{bmatrix} A_{11}^{-1} & A_{12}^{-1} \\ A_{21}^{-1} & A_{22}^{-1} \end{bmatrix}$.

Se considerarmos que a matriz A é a distribuição de parentesco *a priori* e G os valores observados, temos que $P[A|G]$. Como não temos todos os animais da população genotipados, apenas uma parcela da população total, então temos que $P[A_{22}|G]$. Considerando a distribuição conjunta de probabilidade dos valores genéticos (u_g) dos animais não genotipados (u_{g1}) e genotipados (u_{g2}) (Legarra et al., 2009; Christensen et al., 2010) dadas por $p(u_{g1}, u_{g2}) = p(u_{g2})p(u_{g1}|u_{g2})$ e $p(u_{g2}) = N(0, G)$, e que a variância é $\text{Var}(u) = A\sigma_u^2$, temos, então, que a distribuição de probabilidades condicional dos valores genéticos para u_{g1} e u_{g2} é:

$$p(u_{g1}|u_{g2}) = N(A_{21}A_{11}^{-1}u_{g2}, A_{11} - A_{12}A_{22}^{-1}A_{21}).$$

Devido ao fato da variância dos valores genéticos dos animais não genotipados depender da relação de pedigree com os animais genotipados, temos que:

$$\text{Var}(u_{1g}) = \text{Var}(A_{12}A_{22}^{-1}u_{g2} + e) = \text{Var}(A_{12}A_{22}^{-1}u_{g2}) + \text{Var}(e)$$

$$\text{Var}(e) = A_{11} - A_{12}A_{22}^{-1}A_{21} = A_{11},$$

em que $\text{Var}(e)$ é a variância do erro. Substituindo as expressões acima, obtemos que:

$$\text{Var}(u_{g1}) = A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} + A_{11} - A_{12}A_{22}^{-1}A_{21}$$

Rearranjando os termos podemos simplificar a expressão acima, então teremos que as variâncias são:

$$\text{Var}(u_1) = A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}$$

$$\text{Var}(u_2) = \text{Var}(Za) = G$$

e suas covariâncias são $\text{cov}(u_{g1}, u_{g2}) = \text{cov}(u_{g2}, u_{g1}) = A_{12}A_{22}^{-1}\text{Var}(u_2) = A_{12}A_{22}^{-1}G$.

Finalmente podemos escrever o parentesco dos animais genotipados e não genotipados em uma matriz H , expressa na forma de matriz de covariância dos valores genéticos, como:

$$H = \begin{bmatrix} \text{Var}(u_1) & \text{cov}(u_{g1}, u_{g2}) \\ \text{cov}(u_{g2}, u_{g1}) & \text{Var}(u_2) \end{bmatrix}.$$

Colocando H em termos das matrizes A e G, e rearranjando as matrizes temos que:

$$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix},$$

As covariâncias dos valores genéticos entre os animais genotipados e não genotipados são invariantes em relação à ordem da multiplicação das matrizes, pois não é uma regra que a ordem da multiplicação das matrizes resulte em valores em valores distintos, como em $GA_{22}^{-1}A_{21} = A_{12}A_{22}^{-1}G$. Neste caso em particular das covariâncias, como em outros possíveis de verificar, são iguais. A matriz H ainda pode ser escrita da seguinte forma:

$$H = A + \begin{bmatrix} A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}(G - A_{22}) \\ (G - A_{22})A_{22}^{-1}A_{21} & (G - A_{22}) \end{bmatrix},$$

cujas inversa da matriz de covariância H é (Aguilar et al. 2010; Christensen et al., 2010):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

Como A^{-1} e G^{-1} faz referência ao parentesco dos animais, se faz necessário subtrair o parentesco dos animais genotipados (A_{22}^{-1}). Agora podemos simplificar o modelo supracitado como sendo:

$$y = Xb + Wp + Zu_g + e$$

em que Z é a matriz de incidência do valor genético aditivo (u_g), com probabilidade $u_g \sim N(0, H\sigma_u^2)$. Substituindo a matriz H nas equações de modelos misto, temos as equações do modelo acima para o método ssGBLUP:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & w'w + \frac{\sigma_e^2}{\sigma_p^2} & W'Z \\ W'X & Z'W & Z'Z + H^{-1} \frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} b \\ p \\ u_g \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix}$$

em que σ_e^2 , σ_u^2 e σ_p^2 são, respectivamente, as variâncias residual, genética e fenotípica.

A subtração entre as matrizes entre G e A_{22} para evitar dupla contagem pode envolver matrizes que não são equiparáveis, devido aos animais possuírem SNP em comum devido a ancestralidade; e sua compatibilidade depender de suas bases genéticas, bem como das frequências do SNP e dos genótipos na construção de G (VanRaden, 2008; Vitezica et al., 2011). A parte da variância $\text{Var}(u_g)$ que não é explicada pelos SNP pode ser corrigida, incluindo uma parte do efeito poligênico residual ($\beta = 1 - \alpha$) na matriz G .

3.3. Método Random Forest (RF) de aprendizado de máquina

O método RF é um método de aprendizagem supervisionado de máquina nos quais os dados com grande número de preditores, por exemplo SNP, é utilizado para prever uma variável dependente, como por exemplo a variável contínua fenótipo. O método é baseado em árvores de decisão, sendo utilizado para classificação ou regressão de múltiplas variáveis (Chen e Ishwaran, 2012; Alarcon et al., 2015 ; Li et al., 2016), por meio de diferentes medidas de importância, o que o torna uma ferramenta importante para análise de dados genômicos (Breiman, 2001; Chen & Ishwaran, 2012; Touw et al., 2012).

Para entender melhor o método RF, faremos uma breve explicação dos principais conceitos que o envolve, sendo eles: árvores de decisão, amostragem Bootstrap e Bagging. Árvores de classificação ou regressão é um método estatístico introduzido por Breiman et al. (1984). Esse método divide os preditores em J regiões distintas $R_1, \dots, R_J$. Para a construção das regiões, $R_1, R_2, \dots, R_{j-1}, R_j$, que minimizem de a soma dos quadrados dos resíduos (MSE) $\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$, onde $\hat{y}_{R_j}$ é a médias dos dados de treinamento para a J -ésima região. No entanto isso é inviável computacionalmente, então são selecionamos o preditor X_j e um ponto de corte s , tomado aleatoriamente, dividindo o espaço dos preditores nas regiões $\{X|X_j < s\}$ e $\{X|X_j \geq s\}$, minimizando MSE. Considerando apenas duas regiões R_1 e R_2 , e um ponto de corte s , temos que as regiões são definidas por $R_1(j, s) = \{X|X_j < s\}$ e $R_2(j, s) =$

$\{X|X_j \geq s\}$, nas quais buscamos encontrar j e s , minimizando a equação abaixo, até que um critério de parada pré-estabelecido seja alcançado:

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \widehat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \widehat{y}_{R_2})^2$$

Vamos definir como amostras de aprendizado (learning samples) $L_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$, compostas de n amostras independentes e identicamente distribuídas (i.i.d.), cuja a distribuição de probabilidade dos valores é desconhecida na maioria dos casos. Quando se modela problemas de regressão, $X \supset \{x_1, \dots, x_n\}$, $Y \supset \{y_1, \dots, y_n\}$ e $y \in \mathbb{R}$, e usando a amostra L_n , podemos construir o preditor $\hat{h}: x \rightarrow y$, no qual associa a previsão de y a uma observação de entrada $x \in X$. Logo o valor esperado do quadrado do erro é $E[(Y - \hat{h}(x))^2]$ e a média do erro quadrático (MSE) $\frac{1}{m} \sum_{i=1}^m (Y_i - \hat{h}(X_i))^2$. No MSE utiliza-se a amostra de teste $T_m = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_m, Y_m)\}$.

Bootstrap é uma técnica estatística introduzidas por Efron e Tibshirani (1983) são métodos computacionais que usam amostragem aleatória com reposição para calcular medidas de incerteza dos estimadores e obter informações da distribuição de alguma variável aleatória de interesse. O método consiste em: (1) Amostragem aleatória com reposição para formar uma amostra de tamanho m ; (2) calcula-se o estimador de interesse para cada uma das amostras e (3) repete-se os passos (1) e (2) vezes até para gerar B amostras *bootstrap*, nas quais converge para a distribuição verdadeira quando $B \rightarrow \infty$, para melhor entendimento de como a convergência ocorre, os leitores podem ler a respeito da lei fraca dos grandes números dos grandes.

A definição de RF, fornecida por Breiman (2001), é a seguinte: seja $\hat{h}(\cdot, \theta_1), \dots, \hat{h}(\cdot, \theta_q)$ um conjunto de árvores de decisão, com q variáveis i.i.d. $\theta_1, \dots, \theta_q$ obtidas de L_n . O preditor RF $\hat{h}_{rf}$ é obtido por agrupamento (*aggregating*) dessa coleção de árvores de decisão. Uma amostra bootstrap de L_n de tamanho n é obtida aleatoriamente de n observações de L_n com reposição, cada observação (X_n, Y_n) tem probabilidade $\frac{1}{n}$ de ser selecionada em cada sorteio. Cada uma dessas amostras gera uma informação e previsões diferentes, o que causa instabilidade na árvore de decisão.

Introduzido por Breiman (1996) o acrônimo Bagging deriva de bootstrap aggregating é obtido por um conjunto de árvores preditoras, cada uma construída por uma amostra bootstrap diferente, no qual o parâmetro *mtry* controla o número de variáveis que são aleatoriamente selecionadas em cada um dos nós que formarão as árvores, que por *default* é $\frac{p}{3}$ nos problemas de regressão; *ntree* é o número de árvores na floresta, denotado com q , que por *default* é igual a $q = 500$.

Outro conceito importante é o Out-Of-Bag (OOB) que são as observações que ficaram de fora da amostra *bootstrap*, são utilizadas para calcular o erro OOB como sendo uma população de teste. Portanto, para prever a i -ésima observação de X_i , apenas agregamos preditores construídos em amostras de bootstrap que não contêm (X_i, Y_i) , resultando na predição de $\hat{Y}_i$ como saída da i -ésima observação. O erro OOB para regressão é obtido pela expressão $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.

Então, resumidamente, o algoritmo RF funciona da seguinte forma: (1) uma amostra selecionada aleatoriamente um subconjunto de observações, denominadas como amostras de aprendizado (L_n); (2) dentro da amostra L_n seleciona-se um subconjunto de tamanho *mtry* de preditores (marcadores SNP, no contexto em que estamos aplicando); (3) desenvolver uma árvore única por meio de divisões recursivas do subconjunto de SNPs nas amostras, com o propósito de criar nós na árvore, visando separar o conjunto de amostras de observação em dois grupos distintos. Ao dividir um nó na árvore, é selecionado o SNP que possui a maior capacidade de reduzir o Erro Quadrático Médio (MSE) dos nós filhos; (4) usar OOB para determinar o MSE da árvore de decisão formada, no qual para cada variável (SNP) na árvore de decisão (modelo) será realizado uma permutação aleatório nas variáveis e calculada a diferença entre o MSE da nova árvore formada com a variável permutada em relação ao MSE anterior a permutação; (5) gerar uma floresta de árvores aleatórias repetindo as quatro etapas anteriores; (6) obter os valores finais de importância da variável independente, calculando a média dos valores do MSE em todas as árvores da floresta que contém aquela variável que foi permutada.

A importância das variáveis (VI), anteriormente mencionada na quarta etapa do funcionamento do algoritmo RF, é obtida fixando $j \in \{1, \dots, p\}$ e calculando $VI(X^j)$, o índice de importância da variável X^j , considerando o procedimento abaixo: (1) considere uma amostra bootstrap $L_n^{\theta l}$ e a amostra OOB_l associado; (2) calcula-se o erro OOB_l para a árvore construída para a amostra $L_n^{\theta l}$; (3) permuta-se aleatoriamente os valores da variável X^j na amostra OOB_l , obtendo um novo $O\ddot{O}B_l$; (4) calcula-se o novo erro $O\ddot{O}B_l$ da árvore sobre a amostra $L_n^{\theta l}$ e (5) repetir as 4 operações anteriores para todas as amostras bootstrap. Finalmente a importância da variável X^j é:

$$VI(X^j) = \frac{1}{q} \sum_{i=1}^q (eO\ddot{O}B_l - eOOB_l)$$

em que $eO\ddot{O}B_l$ é o erro calculado na etapa 4 e $eOOB_l$ é o erro calculado na etapa 2.

4. Material e métodos

4.1. Registros de pedigree e fenotípicos

O conjunto de dados referentes a produção de leite até 305 dias de lactação pertence ao Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL), disponibilizado pela Embrapa Gado de Leite, contando com pedigree com quinze gerações de animais com fenótipo e/ou com genótipo (Figura 1A), e múltiplas ordens de lactação (Figura 2A), totalizando 120.269 lactações, com média e desvio padrão de 3618 e 2069,7 quilogramas de leite. A produção máxima observada foi de 15.000 Kg e o primeiro e terceiro quartil foram 2114 e 4798 Kg, respectivamente. Os dados foram colhidos pelos técnicos da Associação Brasileira dos Criadores de Gir Leiteiro (ABCGIL) e da Associação Brasileira de Criadores de Zebu (ABCZ), a partir de rebanhos distribuídos pela região sudeste do país.

A normalidade dos resíduos padronizados do modelo animal foram verificados por meio do ajuste de um modelo linear geral considerando os efeitos fixos de grupo de contemporâneos ao parto (rebanho e ano), composição genética do rebanho,

tratamento preferencial, que consiste nas vacas que receberam ocitocina para os partos ou que participaram de torneios leiteiros, e os efeitos linear e quadrático de idade da vaca ao parto, e posteriormente submetidos ao um teste F de Snedecor (Tabela 1) e foram removidos os resíduos padronizados superiores a 3 e inferiores a -3.

Tabela 1 – Análise de variância da característica produção de leite até 305 dias de lactação

Fontes de Variação	GL	SQ	QM	F	p-valor
Grupo de Contemporâneos	6976	3,04E+15	4,35E+11	26.07	<2.2e-16
Composição Genética	7	3,32E+13	4,74E+12	284.10	<2.2e-16
Época	1	2,15E+12	2,15E+12	128.88	<2.2e-16
Tratamento Preferencial	1	1,42E+14	1,42E+14	8518.84	<2.2e-16
Idade (Linear)	1	2,09E+13	2,09E+13	1253.29	<2.2e-16
Idade (Quadrático)	1	2,75E+13	2,75E+13	1648.41	<2.2e-16
Resíduos	113281	1,89E+15	1,67E+10		

GL – “Graus de liberdade”; SQ – “Soma dos quadrados”, QM – “quadrado médio”, F – “valor de calculado”

4.2. Genótipos e controle de qualidade

As informações dos genótipos utilizadas na execução do presente estudo foram obtidas a partir de quatro diferentes painéis nos quais 609 animais foram genotipados com BovineHD BeadChip (Illumina, Inc., San Diego, CA) com 777.962 marcadores SNP, 1.464 animais com BovineSNP50 BeadChip v2 (Illumina, Inc., San Diego, CA) com 53.714 marcadores, 10390 animais com GGP Indicus (Neogen Genomics Corporation, Lincoln, OR) com 35.090 marcadores, e 25.440 animais com Z-Chip (Neogen Genomics Corporation, Lincoln, OR) com 27.533 marcadores distribuídos ao longo do genoma. Os animais genotipados com painéis de baixa densidade (BovineSNP50, GGP Indicus e Z-Chip) foram imputados para o painel BovineHD utilizando o programa FImpute (Sargolzaei et al., 2014).

No controle de qualidade realizado após a imputação dos painéis de SNP, somente os cromossomos autossômicos foram mantidos, sendo excluídos SNP com MAF menor ou igual a 0,015, SNP com desvio de Equilíbrio de Hardy-Weinberg ($p < 10^{-6}$), “GenCall Score” menor ou igual a 0,70 e SNP com correlação entre si maior que 0,995. Para o controle de qualidade das amostras, foram excluídas as amostras com

“call rate” menor ou igual a 90%. Após o controle de qualidade inicial restaram 71436 animais com fenótipo, 21.656 animais com genótipo, 85.258 animais com fenótipo ou genótipo, 13.822 animais com genótipo e sem informação de fenótipo e 61.849 pais sem registro de fenótipo e genótipo, totalizando 147.107 animais no rebanho.

4.3. Classificação dos SNP pelo método “Random Forest” (RF)

Para a classificação dos SNP foi realizada análise de regressão não paramétrica por meio do algoritmo RF. Para a análise RF a PL305 foi utilizada como variável resposta e os SNP foram considerados os preditores da análise. As análises foram realizadas utilizando a função “ranger” (Wright et al. 2022) aplicada em R. Para a execução da função foram necessários dois parâmetros: (1) o número de árvores na floresta, denominado *Ntree*, e (2) o número de marcadores presentes em cada conjunto de variáveis usados para formar a árvore, conhecido como *Mtry*.

Para determinar os valores dos parâmetros utilizados na classificação dos SNP, foi avaliado o impacto dos valores *Ntree* e *Mtry* no quadrado médio do erro Out-of-bag – OOB (MSE) e no R^2 OOB. Os valores *Ntree*, escolhidos arbitrariamente, testados foram 512, 1024, 2048, 4086, 8192. Para o parâmetro *Mtry* foram testadas quatro opções: (i) $\sqrt{M}$; (ii) $2\sqrt{M}$; (iii) 0,1M; e (iv) 0,3M, em que M é igual a número de marcadores SNP no painel originalmente.

Cada *Ntree* e *Mtry* testado utilizou duas populações diferentes de treinamento/validação. A primeira população de treinamento e validação, 70% e 30% respectivamente, foi formada com os animais escolhidos aleatoriamente; a segunda população foi definido a partir do número de gerações, de modo que os animais fundadores até a décima geração foram considerados a população de treinamento, e os animais da validação da décima primeira até a décima quinta geração. Para expressar a importância dos SNP ao longo dos cromossomos autossômicos, o SNP que ocupa o primeiro lugar no ranking foi atribuído o valor de cem por cento de importância, e a importância dos demais SNP foi expressa em relação a este.

4.4. Gene ontology enrichment analysis e Basic local alignment search tool

Após selecionar os melhores parâmetros e realizadas análises de classificação, por meio da regressão não paramétrica, foram selecionados os 15 primeiros SNP de maior relevância, utilizando uma janela de 500 Kpb acima e abaixo da posição do SNP de modo a verificar a existência de genes. As posições desses SNP estavam na versão UMD 3.1 e foram atualizadas para a versão do mapa de SNP ARS-UCD1.2. (Rosen et al. 2020)

A existência de genes foi verificada por meio o pacote biomaRt do software R (Durinck et al., 2005) e a versão 110 de Ensembl Genomes. O pacote clusterProfiler (Yu et al., 2012 e Wu et al., 2021) foi usado para realizar análises de enriquecimento funcional de genes têm que possuem funções biológicas significativamente enriquecidas em relação a um conjunto de genes de interesse, juntamente com o pacote org.Bt.eg.db (Carlson, 2019) com o banco de dados de anotação para o organismo *Bos taurus taurus*.

A análise BLAST (Basic Local Alignment Search Tool) (BLAST, 2023 e ALTSCHUL et al., 1990) foi realizada com o intuito de buscar maiores informações a respeito dos genes em outras espécies, a partir do alinhamento e comparação entre sequências de nucleotídeos. A ferramenta BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) e as sequências de nucleotídeo estão disponíveis no *National Library of Medicine* (NCBI - <https://www.ncbi.nlm.nih.gov/>).

4.5. ssGBLUP, habilidade de predição dos valores genéticos genômicos e ssGWAS.

Os valores genéticos genômicos foram preditos pelo método de “Single-step” GBLUP (ssGBLUP). O seguinte modelo estatístico unicaracterística foi considerado para PL305:

$$y = Xb + Wp + Zu_g + e,$$

em que y é o vetor de observações fenotípicas; X é a matriz de incidência dos efeitos fixos; b é o vetor de efeitos fixos, Z é a matriz de incidência dos efeitos aleatórios

genéticos diretos; a é o vetor de efeitos aleatórios para o efeito genético aditivo; W é a matriz de incidência do efeito aleatório de ambiente permanente; p é o vetor do efeito aleatório de ambiente permanente; e e é o vetor de efeitos residuais.

O ssGBLUP foi implementado substituindo a matriz A utilizada no BLUP tradicional pela matriz H , que inclui a informação genômica. A inversa da matriz H será determinada de acordo com Aguilar et al. (2010) e Legarra et al. (2009):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

A matriz G (matriz de parentesco genômico) foi construída de acordo com VanRaden (2008) e as análises foram realizadas nos programas da família BLUPF90 (Misztal, 2012).

A Matriz G foi obtida por meio da matriz $M_{n \times m}$, em que n é o número de animais genotipados e m é o número de SNP, que conta o número de alelos B na população, codificada como, 0, 1 e 2. A operação matricial MM' tem em sua diagonal o número de *loci* homozigotos para cada animal e fora da diagonal o número de alelos compartilhados entre os indivíduos da população. Assumiu-se P como o vetor que contém as frequências alélicas da população e p_i como sendo a frequência do alelo B para cada *loci*.

Temos que $Z = M - P$, em que $P = (2p_i - 0,5)$. Subtrair P de M , favorece mais os alelos raros da população do que alelos comuns quando estamos calculando a matriz de parentesco genômico. O coeficiente de endogamia genômico, nesse caso, é maior em animais que são homozigotos para alelos raros da população. Por fim, temos, pelo método de VanRaden (2008), que $G = \frac{ZZ'}{2 \sum p_i(1 - p_i)}$.

O denominador da razão se faz necessário para tornar a matriz de parentesco genômico análoga a matriz de parentesco com base no pedigree. Assim como na matriz MM' , na diagonal da matriz G temos o número de *loci* homozigotos para cada indivíduo e o número de alelos compartilhados entre os indivíduos, mas agora centrada na frequência alélica da população.

Como os marcadores SNP são apenas um tipo de polimorfismo genético presente no DNA, e a proposta do estudo é descobrir qual o menor número necessário de SNP pode ser utilizado para o computo do GEBV sem perda de acurácia de predição, foi avaliado o *blending* da matriz G com a matriz A_{22} , como apresentado na equação abaixo:

$$G_{\text{blending}} = \alpha G + \beta A_{22}$$

Em que G é a matriz de parentesco genômico original, e β reflete a proporção de variância poligênica residual não contabilizada por G (Habier et al., 2007; Liu et al., 2016; Mäntysaari et al., 2017), e por fim $\alpha = 1 - \beta$.

As diferentes ponderações das matrizes G^{-1} e A_{22}^{-1} foram consideradas para melhorar a convergência das análises e reduzir o viés da estimação do valor do GEBV de animais jovens da população. Assim, diferentes proporções de ω foram introduzidas no cálculo da matriz H^{-1} , que misturam o inverso de uma parte de A com o inverso de G, como descrito na equação abaixo, e $\tau = 1$, que segue o default dos programas da família blupf90:

$$H^{-1} = A^{-1} + \begin{vmatrix} 0 & 0 \\ 0 & \tau G_{\text{blending}}^{-1} - \omega A_{22}^{-1} \end{vmatrix}$$

Para testar o desempenho de cada cenário utilizado na predição genômica para PL305, a capacidade de predição do GEBV foi avaliada pela correlação de Pearson entre o fenótipo ajustado e o GEBV, do quadrado médio do erro e do coeficiente de regressão. Para determinar o coeficiente de regressão foi utilizado o seguinte modelo:

$$y = b_0 + b_1 x + e$$

em que y são os fenótipos ajustados, b_0 o intercepto, b_1 o coeficiente de regressão linear indicando viés de predições, x são os GEBVs e e o resíduo. As análises para habilidade de predição e o modelo de regressão foram realizadas utilizando os softwares predictf90 e a linguagem R.

As médias e variâncias do GEBV de cada cenário estudado foram testadas duas a duas por um teste t de Student, e um teste de razão de variâncias (Teste F), em um nível de significância de 5%. Os valores do fenótipo predito, que é o somatório dos efeitos incluídos no modelo e o resíduo, que resulta do fenótipo descontado os efeitos

incluídos no modelo, também foram submetidos aos testes de comparação de médias e variâncias.

Para servir de comparação com os resultados da classificação de importância dos SNP distribuídos nos cromossomos, realizada pelas análises de machine learning, foi utilizado o software postGSf90 (GWAS by ssGBLUP) para expressar o $\log_{10}$ (p-valor) dos SNP, a estimativa dos efeitos dos SNP, e a proporção da variância explicada por cinco SNP adjacentes distribuídos nos cromossomos autossômicos. Este software utilizou-se do valor do GEBV do cenário 400k_1w (default), a matriz G^{-1} (OPTION saveGInverse), a matriz A_{22}^{-1} (OPTION saveA22Inverse) e o p-valor a partir de elementos da inversa das equações de modelos mistos obtidos pelo blupf90 (OPTION snp_p_value), como descrito em Lourenco et al. (2020). Terminada a análise no postGSf90, os valores dos resultados de efeito de SNP, proporção da variância explicada por 5 SNP adjacentes (OPTION windows_variance 5) e $-\log_{10}$ (p-valor), foram expressos ao longo dos cromossomos autossômicos.

4.6. Cenários

O painel de genótipo completo, após o controle de qualidade, possuía 416.155 SNP, denominado como 400K, que no presente estudo foi considerado como controle, para comparar o desempenho dos demais cenários com as menores densidades propostas. Foram testados arbitrariamente os painéis nas densidades 40.000, 10.000 e 4.000 SNP, denominados como 40k, 10k e 4k respectivamente.

Em cada um dos painéis foram testados três diferentes parâmetros (ω , α e β) nos quais foram utilizados com a finalidade de tornar a informação genômica compatível com as informações da matriz de parentesco genético aditivo (matriz A_{22}), para proporcionar a inversão das matrizes e controlar o viés das estimativas de GEBV dos animais jovens da população, como supracitados.

Os *defaults* estabelecidos pelo programa blupf90 são $\tau = 1$, $\omega = 1$, $\alpha = 0,95$ e $\beta = 0,05$. A partir desses padrões foram elaborados três cenários para a parametrização de ω : {1; 0,85; 0,70}, fixando $\alpha = 0,95$, $\beta = 0,05$ e $\tau = 1$; e três outros cenários para a

parametrização do α : {0,95; 0,70; 0,55} e β : {0,05; 0,30; 0,45}, com os valores fixos de τ e ω iguais a 1.

Dessa forma, ao combinar os 4 painéis, com as três parametrizações de ω e juntamente com as parametrizações de α e β , tem-se os cenários testados, e descritos na Tabela 2. Ressalta-se que os cenários $nK_1\omega$, em que n representa o tamanho do painel, são aqueles em que as parametrizações seguem o default do programa blupf90.

Tabela 2 – Cenários testados para os diferentes valores de α , β e ω

	$\alpha = 0,95, \beta = 0,05$ e $\tau = 1$			$\tau = 1, \omega = 1$	
	$\omega=1$	$\omega=0,85$	$\omega=0,70$	$\alpha = 0,70 \beta = 0,30$	$\alpha = 0,55 \beta = 0,45$
400K	400k_1 ω	400k_085 ω	400k_070 ω	400k_070 α _030 β	400k_055 α _045 β
40K	40k_1 ω	40k_085 ω	40k_070 ω	40k_070 α _030 β	40k_055 α _045 β
10k	10k_1 ω	10k_085 ω	10k_070 ω	10k_070 α _030 β	10k_055 α _045 β
4k	4k_1 ω	4k_085 ω	4k_070 ω	4k_070 α _030 β	4k_055 α _045 β

5. Resultados e Discussão

Após a edição dos dados fenotípicos e remoção dos resíduos, com relação ao modelo de efeitos fixos, restaram 116.521 registros de produção de leite até os 305 dias de lactação, com média e desvio-padrão igual a 3.539 Kg e 1.924,30 Kg, respectivamente.

Nos diferentes valores testados dos parâmetros de *Ntree* e *Mtry*, resultaram em um valor médio de RMSE, OOB (MSE) e R^2 OOB, respectivamente, iguais à 1.841,37 2.635.722,25 e 0,31 para as análises em que o grupo de treinamento e validação foi formado a partir das gerações dos animais, e de 1.658,65, 2.800.000,80 e 0,28, para as análises em que os dados foram separados aleatoriamente (Figura 3A).

Os valores de *Ntree* e *Mtry* para a classificação dos SNP utilizados no modelo de análise do RF foram iguais a 8.192 e 124.846, respectivamente, uma vez que resultaram no menor valor de OOB (MSE) e no maior R^2 OOB observado. O valor de *Mtry* escolhido representa 0,3 vezes no número de SNP presentes no painel 400k.

A importância dos SNP ao longo dos cromossomos autossômicos expressa em porcentagem (Figura 4A), apresenta poucos SNP com grande efeito e muitos com

pouco ou nenhum efeito, evidenciando o efeito poligênico da característica. Na análise de RF, o cromossomo 14 se destacou por conter o SNP de maior importância para PL305. Na literatura, a região do cromossomo 14 tem recebido atenção considerável, já que esta região abriga um grande número de QTL que afetam a produção de leite. O gene Diacilglicerol O-aciltransferase 1 (DGAT1) localizado em 1 Mb em BTA14 é geralmente aceito como um gene importante para a produção de leite (Jiang et al., 2015; Nayeri et al., 2016), no entanto está distante dos SNP que ocupa a primeira e décima quarta posição no ranking.

Foram encontrados 131 genes (Tabela 12B), distribuídos nos cromossomos 3, 6, 12, 14, 15, 21, 24 e 28. Na região da janela de 1Mb do SNP mais relevante, presente no cromossomo 14, foram encontrados 8 genes que estão relacionados com importantes vias metabólicas. Por exemplo, genes relacionados com: *i*) a produção de ATP (ATPase family AAA domain containing 2); *ii*) a permissividade da atividade do fator de transcrição de ligação ao DNA, específica para RNA polimerase II; *iii*) permissividade da ligação de íons metálicos; *iv*) permissividade da ligação às proteínas, que permite atividade de heterodimerização de proteínas (zinc fingers and homeoboxes 1); *v*) com o transporte de substrato da membrana do retículo endoplasmático para o citosol; *vi*) com o controle de ciliogênese primária, uma organela sensorial baseada em microtúbulos que liga as vias de sinalização à diferenciação, crescimento e desenvolvimento celular permite a atividade do adaptador molecular e a ligação às proteínas (TBC1 domain family member 31); *vii*) envolvido na proliferação da população célula e na via de sinalização do receptor do fator de crescimento epidérmico, que permite ligação idêntica a proteínas, a ligação da subunidade reguladora da fosfatidilinositol 3-quinase (family with sequence similarity 83 member A) (NCBI, 2023).

O cromossomo 12, o segundo SNP de maior importância, possui o gene ENSBTAG00000043985 - Fator de transcrição 1 da família Dachshund (DACH1), que codifica uma proteína associada à cromatina que se associa a outros fatores de transcrição de ligação ao DNA para regular a expressão gênica e a determinação do destino celular durante o desenvolvimento (NCBI, 2023). Devido a sobreposição das

janelas de alguns SNP dos cromossomos 1, 12 e 14, foram encontrados os mesmos genes dentro da janela de 1Mb, assim temos mais de um SNP próximo ao gene, e possivelmente em desequilíbrio de ligação.

O gene ENSBTAG00000007383, que é responsável pela síntese da proteína zinc finger 521, é conhecido por sua função como fator de transcrição, principalmente na ligação ao DNA. Rakhra e Rakhra (2021), têm destacado a associação desse gene com o controle e a resposta imunológica. Além disso, evidências indicam que o gene desempenha papel crucial em características relacionadas à fertilidade em fêmeas de gado vermelho Nórdico (Nordic Red Cattle), como o número de inseminações por concepção e o intervalo entre o parto e a primeira inseminação. Essas características são componentes essenciais do índice de fertilidade das novilhas.

Características ligadas à fertilidade tais como idade ao primeiro parto, cuja variabilidade fenotípica é limitada por práticas de manejo frequentemente possuem baixa herdabilidade e são correlacionadas negativamente com a PL305 (PANETTO et al., 2022). Shao et al., (2021) resumiram um conjunto de estudos de herdabilidade de GWAS e observaram a presença de 15 QTL associados à idade ao primeiro parto no cromossomo 14, e estimativas de herdabilidade das características reprodutivas variando de 0 a 0,57, com grandes diferenças entre as populações de bovinos de corte, leite e búfalos. Esses estudos elucidam a possibilidade de melhorar as características de fertilidade sem suprimir a produção de leite.

Ao buscar as sequências previamente depositadas no NCBI (2023), utilizando-se da ferramenta Basic Local Alignment Search Tool (BLAST, 2023 e ALTSCHUL et al., 1990) em outras espécies que possuem os genes ENSBTAG00000002002 e ENSBTAG00000007383, na coleção de nucleotídeos presente no banco de dados (nr/nt), esses genes foram encontrados em outras espécies de ruminantes de interesse econômico, diferentes de *Bos taurus* e *Bos indicus*, como por exemplo *Bubalus bubalis* e *Bubalus bubalis carabanesis*, *Capra hircus* e *Ovis aries*, e outros tipos de ruminantes como o *Bos mutus*, *Rangifer tarandus platyrhincus* e *Capricornis sumatraensis*. Dada a associação do gene ENSBTAG00000007383 com respostas imunológicas, que contribuem para a rusticidade do animal, bem como sua relação com PL305, a

pesquisa desse gene se torna de grande interesse em estudos que buscam aprimorar a idade ao primeiro parto no contexto do programa nacional de melhoramento do Gir leiteiro.

Destaca-se ainda a presença de 21 genes relacionados a receptores olfatórios (RO) de diferentes tipos no cromossomo 15, no qual fazem parte da janela de 1Mb do SNP de nona posição no ranking. Em bovinos, estima-se que tenham 272 subfamílias de genes de RO, sugerindo que a diversidade de sequências desses genes é menor que espécies como cães e humanos, e que foram influenciados pela seleção natural (Lee et al., 2013). Um desses genes é o ENSBTAG00000030908 (olfactory receptor family 9 subfamily Q member 6), que transcreva receptores acoplados à proteína G (G-protein-coupled receptors, GPCR), e que compartilham uma estrutura comum de sete segmentos transmembranares, semelhante à de muitos outros receptores hormonais e de neurotransmissores, sendo essenciais para o reconhecimento e a transmissão dos sinais ligados aos odores. Este gene está presente em diferentes espécies de herbívoros, diferentes das supracitadas, como por exemplo *Hippopotamus amphibius kiboko*, *Camelus dromedarius*, *Bison bison bison*, *Moschus berezovskii*, *Odocoileus virginianus texanus*, *Oryx dammah*, e *Budorcas taxicolor*.

A análise gene ontology enrichment (GO) considerando janela de 1Mb detectou 41 genes (Tabela 13) com p-valor menor que 0,05. No entanto, o p-valor ajustado (~0,27) e q-valor (~0,25), que são respectivamente, a correção estatística aplicada ao p-valor para controlar o erro do tipo I quando se realizam múltiplos testes de hipótese e a correção do p-valor para testes de múltiplas hipóteses, indicou baixa probabilidade desses termos GO estarem associados aos genes de interesse.

Na figura 5^a, gráfico A, foi expressa a importância dos SNP, em porcentagem, em função do ranking dos SNP, e na Figura 5A gráfico B tem-se a porcentagem de importância dos cem primeiros SNP. Podemos observar que a porcentagem de importância dos SNP decresce rapidamente, com os valores tendendo a zero logo nos cem primeiros SNP. Desse modo, mesmo no menor painel proposto, há uma grande quantidade de SNP de pouca importância para PL305.

A média e a variâncias das frequências alélicas nas densidades de painéis de SNP foram diferentes (Tabela 3). As maiores variâncias da frequência alélica foram observadas nos painéis 400k e 4K, respectivamente, seguida dos painéis 10k e 40k, resultando em distribuições das frequências alélicas diferentes (Figura 6A).

A distribuição dos valores da diagonal da matriz de parentesco genômico (Matriz G) apresentaram maior desvio em relação à média com a redução do número de SNP. Os valores fora da diagonal de G foram mais próximos da média (Figura 7A). Os coeficientes de assimetria e curtose da diagonal para os cenários 400k, 40k, 10k e 4k, foram respectivamente iguais a 0,471; 0,996; 1,403 e 1,684, e 3,328; 4,257; 5,542 e 6,818, e os coeficientes fora da diagonal de G, foram iguais a 1,164; 1,169; 0,896 e 0,711, e 5,337; 5,338; 4,479 e 4,311, o que justifica a variação da distribuição dos dados da matriz G em torno da média.

Tabela 3 – Estatísticas descritivas das frequências alélicas para as diferentes densidades de painel

	400k	40K	10K	4K
Número de SNP	416155	40000	10000	4000
Média	0,539	0,517	0,523	0,521
Mínimo	0,015	0,015	0,015	0,015
Máximo	0,999	0,985	0,985	0,985
Variância	0,084	0,049	0,067	0,082

A distribuição de valores da matriz A_{22} (Figura 8A) possui grande quantidade de registros de animais que não possuem nenhuma relação de parentesco, diferente da matriz G em que algum parentesco entre os animais não relacionados é observado. Muitos genes foram atribuídos a ancestrais em comum, um dos fatores que promovem as diferenças observadas nas matrizes A_{22} e G.

Os coeficientes de assimetria e curtose da diagonal para matriz A_{22} são 2,496 e 14,141, e fora da diagonal 1,210 e 4,398. Com base nesses coeficientes, podemos observar as diferenças entre as matrizes no que diz respeito a distribuição de valores em torno da média. As médias dos valores da diagonal e fora da diagonal da matriz G foram iguais ao da matriz A_{22} . No entanto, a variância dos elementos da G é menor em

densidades maiores (Tabela 4). Essas médias foram iguais, pois escalamos a matriz G para ser compatível com a matriz A_{22} , seguindo o default do programa blupf90.

Tabela 4 – média e desvio-padrão da matriz de parentesco genômico (G_{Original}) e da matriz de parentesco genético aditivo (A_{22}).

Matriz	Médias		Desvio Padrão	
	Diagonal	Off Diagonal	Diagonal	Off Diagonal
G_{400k}	1,02	0,09	0,0671	0,0630
G_{40k}	1,02	0,09	0,0681	0,0613
G_{10k}	1,02	0,09	0,0997	0,0722
G_{4k}	1,02	0,09	0,1361	0,0859
A_{22}	1,02	0,09	0,0266	0,0753

Examinando separadamente as matrizes consideradas no cálculo da G pelo método de VanRaden (2008), podemos inferir que as diferenças observadas nas distribuições dos valores na matriz G são em decorrência do número de SNP e da frequência alélica na população de cada painel proposto. A matriz A_{22} calculada com base nas relações de pedigree é obtida como duas vezes a média esperada de genes idênticos por descendência (IBD) compartilhada entre dois indivíduos parentes (Malécot, 1948), a matriz G , derivada de marcadores moleculares, é idêntica por estado (IBS), portanto, apresentam maior variação entre animais geneticamente próximos (VanRaden, 2008; Hayes et al., 2009).

A compatibilidade entre G e A_{22} depende de suas bases genéticas, bem como das frequências do SNP e dos genótipos na construção de G (VanRaden, 2008; Vitezica et al., 2011). As diferenças na base genética entre G e A_{22} ocorreram porque G não é construído com as frequências alélicas da população base (VanRaden, 2008) e o pedigree pode ser incompatível com a matriz G (Misztal et al., 2013).

A correlação entre o coeficiente de endogamia genômico e genético aditivo, a correlação dos elementos diagonal, entre todos os elementos, e a correlação dos elementos fora da diagonal da matriz G_{blending} com a matriz de parentesco genético aditivo, matriz A_{22} (Figura 9A), não se alteraram com os diferentes valores de ω testados, uma vez que esse parâmetro altera a proporção da inversa da matriz A_{22} na

inversa da matriz H , sendo a G_{blending} igual para todos os cenários de ω testados. No entanto, as correlações ficaram menores quando o número de informações presentes no painel de SNP decresce. Os diferentes cenários testados para α e β resultaram em correlações maiores conforme se aumentou a proporção da matriz A_{22} na matriz G_{blending} , considerando uma mesma densidade do painel; porém, painéis menores resultaram em correlações menores, para todos os cenários de α e β avaliados.

Os diferentes cenários testados para α e β proporcionaram grande amplitude de valores de correlação dos coeficientes de endogamia, possivelmente influenciados pelas frequências alélicas da população e as demais matrizes que compõem a G , observadas neste estudo. Corroborando com nossas observações, segundo Nishio et al (2023), a grande variação da correlação entre o coeficiente de endogamia genômico e genético aditivo (0,33 a 0,86) observada em seu estudo com gado japonês, poderia ser consequência das metodologias utilizadas para o cálculo do coeficiente de endogamia genômico depender das frequências alélicas, e estas serem diferentes da população base.

Os valores das médias e das variâncias da diagonal da matriz G_{blending}^{-1} menos a matriz A_{22}^{-1} aumentaram com a redução do número de marcadores presentes no painel de SNP, e com a redução de ω (Figura 10A). Os valores das médias fora da diagonal apresentaram diferenças pequenas com relação aos diferentes cenários de ω e painéis propostos, no entanto a variância de fora da diagonal teve maior variação (Figura 11A).

Os cenários nos quais foram testados diferentes valores de α e β resultaram em maiores médias e variâncias da diagonal da matriz $G_{\text{blending}}^{-1} - A_{22}^{-1}$ com a redução do número de SNP no painel (Figura 12A). No entanto, o acréscimo da proporção da matriz A_{22} dentro da matriz G_{blending} , ou seja, quando o valor de α cresceu, resultou em decréscimo nos valores das médias e das variâncias da diferença. Nota-se também que a influência da redução do SNP na diferença $G_{\text{blending}}^{-1} - A_{22}^{-1}$ são menores quando se considera maiores valores de α . As médias e variâncias fora da diagonal tiveram pouca variação com os diferentes parâmetros de α , β e diferentes números de SNP (Figura 13A). Os valores da média e desvio-padrão da estimativa do GEBV da geração zero

(animais fundadores da população) até a décima geração obtidos nos diferentes cenários tiveram influência da quantidade de SNP considerado no cálculo da matriz G , e dos parâmetros ω , α e β testados. Os valores do ganho genético por gerações acompanharam o mesmo comportamento da média do valor genético em toda a população (Tabela 5).

Tabela 5 – GEBV e ganho genético por geração, da primeira a décima geração (população de treinamento)

	μ	σ_p	b_0	σ_{ep}	R^2
400k 070w	230,9	365,2	48,28	300,9	0,32
10k 070w	232,3	368,3	48,82	303,0	0,32
40k 070w	232,3	369,4	48,80	304,4	0,32
4k 070w	232,5	364,6	48,65	299,0	0,33
400k 085w	258,3	390,0	52,80	317,6	0,34
40k 085w	261,3	397,0	53,69	323,4	0,34
10k 085w	261,9	397,4	53,88	323,4	0,34
4k 085w	262,0	394,0	53,70	319,7	0,34
4k 045 α 055 β	320,3	465,0	64,43	374,0	0,35
400k 1w	320,9	453,4	63,22	363,4	0,36
40k 1w	320,9	453,4	63,22	363,4	0,36
400k 045 α 055 β	321,2	445,8	63,27	353,7	0,37
10k 045 α 055 β	321,3	456,6	64,09	364,6	0,36
40k 045 α 055 β	321,3	456,6	64,09	364,6	0,36
400k 070 α 030 β	322,7	450,9	63,55	359,3	0,37
40k 070 α 030 β	328,3	467,8	65,34	374,6	0,36
4k 070 α 030 β	329,6	477,1	65,92	384,3	0,35
10k 070 α 030 β	330,9	477,8	66,22	384,2	0,35
4k 1w	335,2	485,1	66,75	391,7	0,35
10k 1w	337,0	487,6	67,25	393,1	0,35

μ e σ_p , são respectivamente, a média e o desvio padrão do GEVB; b_0 , σ_{ep} , R^2 , são respectivamente, o ganho genético por geração, o erro padrão residual do modelo e o coeficiente de determinação do modelo.

As demais estatísticas descritivas de cada cenário mantiveram-se próximas umas das outras nos cenários em que α e β foram alternados, diferindo nos cenários de ω , e é possível observar que a dispersão e a quantidade de observações consideradas *outliers* são maiores em alguns cenários (Figuras 14A até 17A). As médias e variâncias do GEBV de cada cenário, comparadas duas a duas por um teste t de Student, e um teste de razão de variâncias (Teste F), em um nível de significância de 5% e de acordo com o número de SNP e a proporção da variância poligênica residual considerada nas análises (α e β), provocaram diferenças significativas em alguns dos cenários testados, bem como o parâmetro ω , que controla a matriz A_{22}^{-1} na H^{-1} (Tabelas 6B e 7B).

As diferenças observadas nas médias e variâncias do GEBV podem ser em decorrência das bases genéticas das matrizes G e matriz A_{22} serem diferentes, embora tenhamos ajustado G a A_{22} antes de misturar com um efeito poligênico residual para evitar inconsistências, conforme descrito por McWhorter et al. (2022). A base genética em A_{22} era diferente para fazer tanto o *blending* das matrizes como para ajustar A_{22}^{-1} na H^{-1} . Espera-se que essas diferenças sejam pequenas no GEBV quando o parâmetro de mistura β tem um valor pequeno e α grande, por exemplo, 0,05 como em VanRaden (2008).

Andersen et. al. (2021), trabalhando com características de tipo em bovinos e testando as proporções de efeito poligênico de 10% ou 30%, mostraram que desvios-padrão de GEBV eram semelhantes entre si. Outros autores também utilizaram valores de β maiores nas avaliações genéticas (Liu et al., 2014; Andersen et al., 2021; Interbull Centre, 2021; Alkhoder & Liu, 2021, Hollifield et al., 2022), maiores que o *default* do BLUPf90, de modo que $\beta = 0,9999$ tornaria nosso modelo próximo a um modelo de avaliação genética convencional, sem efeitos de marcadores. Mas os valores entre 0,0001 e 0,5 foram considerados, e nenhuma diferença nos critérios de convergência foi observada (Liu et al., 2014). Aguilar et al. (2010) observaram melhor acurácia de predição do GEBV com o aumento de β .

As mudanças provocadas na matriz G_{blending} e na H^{-1} não alteraram o valor da acurácia e do viés de predição do GEBV cuja média de valores foram de 0,61 e 0,44, respectivamente, nas populações da décima primeira geração até a décima quinta geração (p-valor < 0,001, para os coeficientes de correlação e de inclinação da reta, respectivamente). A média do R^2 dos modelos de ajuste que verificaram o viés de predição foi de 0,37. Hollifield et al (2022), estudaram o impacto do *blending* da matriz G com as matrizes A_{22} e I (matriz identidade) em proporções que variaram de 0,30 a 0,001, observando alta correlação do GEBV do cenário de $\beta = 0,05$ (default) com os demais cenários testados, e que não houve alteração na acurácia e no viés de predição do GEBV nas diferentes ponderações de β . Li et al (2018) apresentaram acurácias de predição muito próximas entre si, em estudo que testou conjuntos diferentes de SNP

selecionados por métodos de aprendizado de máquina, e diferentes densidades de painéis de SNP.

Como o tamanho efetivo da população (N_e) de bovinos da raça Gir é relativamente pequeno, $N_e \sim 100$ (Malhado et al., 2010; Reis Filho et al., 2010; Santana et al., 2014), os segmentos cromossômicos em desequilíbrio de ligação são extensos e em menor número e, portanto, menor número de SNP é necessário para explicar a variância genética das características (Goddart et al., 2010). Além disso, foram selecionados os SNP que explicam quase que a totalidade da variância genética da característica, o que também explicaria as estimativas de acurácia e viés de predição de GEBV observados nos cenários utilizados não apresentarem diferenças.

Valores próximos aos dos cenários testados foram observados por Su et al (2010), em um estudo com vacas da raça Holandesa usando validação cruzada. Os autores obtiveram acurácia de predição do GEBV para produção de leite entre 0,64 a 0,70. Luan et al (2010) relataram acurácia do GEBV para produção de leite de 0,54 a 0,56 em seu estudo com Norwegian Red Cattle. No entanto, Ding et al (2013), em estudo com animais da raça Holandesa observaram que a acurácia de predição usando os métodos Bayes-B e GBLUP para produção de leite variou de 0,317 a 0,380, sendo valores abaixo do que foi observado no presente estudo. A acurácia de predição do GEBV depende de vários parâmetros, incluindo o tamanho da população de referência e sua estrutura genética (Le et al., 2017), da relação entre os animais das populações de validação e treinamento, da hereditariedade da característica estudada, podendo variar entre diferentes raças e populações (Bolormaa et al. 2013).

A acurácia do viés de predição não foi afetada nos cenários testados, mas como consequência das diferenças significativas observadas no GEBV, os valores do fenótipo predito (Y_{Hat}) e o valor dos resíduos, ambos utilizados para o cálculo da acurácia e do viés de predição, apresentaram diferenças significativas nas médias e variâncias em parte dos cenários testados (Figuras 16 e 17, e Tabelas 8B a 11B).

Observando o logaritmo negativo na base 10 do p-value dos SNP (Figura 18A), a estimativa dos efeitos dos SNP (Figura 19A), e a proporção da variância explicada por cinco SNP adjacentes distribuídos nos cromossomos autossômicos (Figura 20A), notou-

se que os SNPs de maior relevância para PL305 estão em regiões cromossômicas diferentes. Ambas as metodologias para classificar os SNP de maior relevância identificaram 211, 946 e 7578 SNP em comum para os painéis 4K, 10K e 40K respectivamente. Uma provável explicação para essas diferenças pode ser atribuída ao fato de que a análise ssGWAS leva em consideração os efeitos fixos do modelo e que os resíduos assumem distribuição gaussiana, enquanto a análise de RF é não paramétrica.

Os SNP encontrados nas análises de RF, estão associados a genes que estão fortemente conectados a vias metabólicas que são indispensáveis a sobrevivência da espécie e a disseminação de descendentes. Assim esses genes estão no genoma devido a ancestralidade em comum e, portanto, presentes em todas as gerações da população influenciando nas estimativas do GEBV. A proximidade de mais de um SNP a um mesmo gene, sugere que o desequilíbrio de ligação pode estar afetando a importância de cada um dos SNP preditores no processo de análise RF e ssGWAS. Outros algoritmos poderiam ser verificados para compreender a complexidade da arquitetura genética da característica, como por exemplo os estudados por Li et al. (2018), que capturaram diferentes subconjuntos de SNP com os métodos de machine learning testados. Maturana et al (2014) testaram e mostraram que os métodos Bayes A e Bayesian LASSO em diferentes arquiteturas genéticas, evidenciaram a influência dessas variáveis nos SNP capturados, erro tipo 1 e acurácia de predição, de modo que marcadores com alta MAF são mais fáceis de serem detectados pelos métodos utilizados pelos autores, principalmente se os SNP têm grande efeito na característica.

6. Conclusão

A análise de RF mostrou-se promissora para selecionar um subconjunto de SNP preditores no modelo, de forma a não alterar a habilidade de predição, obtendo vários SNP em comum com a análise de ssGWAS. O subconjunto dos quinze primeiros SNP mais relevantes evidenciou muitos genes, presentes em ancestrais em comum, que estão relacionados com funções biológicas e vias metabólicas primordiais a

sobrevivência e disseminação dos animais, contribuindo para as divergências entre as matrizes que compõem a H .

Os cenários testados apresentaram diferenças nas estimativas da média e variância do GEBV, do Y_{Hat} e do *Residuals*, os dois últimos utilizados no cálculo da acurácia e viés de predição, para os diferentes cenários de ponderação do efeito poligênico residual considerado na matriz G e proporções de A_{22}^{-1} na inversa da matriz H , bem como para os diferentes painéis de customizados. Mesmo com as diferenças observadas no estudo, a acurácia e viés de predição não se alteraram diante dos diferentes cenários, de modo que um pequeno conjunto de SNP seria suficiente para explicar quase que a totalidade da variação da característica, sem perder a habilidade de predição do modelo.

7. Referências bibliográfica

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal Dairy Science** 93: 743–752.

Akanno EC, Schenkel FS, Sargolzaei M, Friendship RM, Robinson JA (2014) Persistency of accuracy of genomic breeding values for different simulated pig breeding programs in developing countries. **Journal of Animal Breeding and Genetics** 131: 367–378.

Alarcon, F., Perduca, V., and Nuel, G. (2015). Is it possible to detect $G \times E$ interactions in GWAS when causal exposure is unobserved? **Journal Epidemiologic Research**. 2, 109–117.

Alkhoder, H, Liu Z. (2021). Application of a single-step SNP BLUP model to conformation traits of German Holsteins. **Interbull Bulletin**, 56, 30–40.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. **Journal of Molecular Biology**. v. 215, p. 403-410.

Andersen, T., Nielsen, U. S., Aamand, G. P., Mäntysaari, E., Pösö, J., & Rius-Vilarrasa, E. (2021). Using single-step genetic evaluation for type traits in the Nordic countries. **Interbull Bulletin**, 56, 90–93.

Anuário Leite 2022: pecuária leiteira de precisão (2022). Empresa Brasileira de Pesquisa Agropecuária – Embrapa Gado de leite. Disponível em:

<<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1144110>> Acesso em: 11 Set. 2023

Anuário Leite 2023: leite baixo carbono (2023). Empresa Brasileira de Pesquisa Agropecuária – Embrapa Gado de leite. Disponível em: <<https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1154264/1/Anuario-Leite-023.pdf>> Acesso em: 11 Set. 2023

Balieiro ES, Pereira JCC, Verneque RS, Balieiro JCC, Valente J (2003) Estimativas de herdabilidade e correlações fenotípicas, genéticas e de ambiente entre algumas características reprodutivas e produção de leite na raça Gir. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, vol.55, n.1, p.85- 91.

Basic Local Alignment Search Tool disponível em: <<https://blast.ncbi.nlm.nih.gov/Blast.cgi>> Acesso em: 11 Out. 2023

Bolormaa S, Pryce JE, Kemper K, Savin K, Hayes BJ, Barendse W, (2013) Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle. **Journal Animal Science**. 91(7):3088-104

Borges ALCC (2016) Exigências nutricionais para manutenção e lactação e perfil metabólico de vacas F1 Holandês x Gir e Gir, em diferentes planos nutricionais. 189 f. **Tese (Doutorado em Zootecnia)** – Escola de Veterinária da Universidade Federal de Minas Gerais, Belo Horizonte.

Breiman, L (2001) Random forests. **Machine Learning** 45(1): 5–32.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2)

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). Classification and regression trees.

Carlson M (2019). *org.Bt.eg.db: Genome wide annotation for Bovine*. R package version 3.8.2.

Carvalho R, Boison, SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien AMP, Sölkner J, McEwan JC, Van Tassell CP, Sonstegard TS, Garcia JF (2014) Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution** 46: 69

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. **Genomics** 99: 323–329.

Chen CY, Misztal I, Aguilar I, Legarra A, Muir WM (2011). Effect of different genomic relationship matrices on accuracy and scale. **Journal Animal Science**. 89:2673–2679. doi:10.2527/jas.2010-3555

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. **Genomics**. 99, 323–329.

Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**, 42, 2.

Ding X, Zhang Z, Li X (2013) Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. **Journal Animal Science**. 96 :5315–5323.

Duchesne P, Rémillard B (2005) Statistical modeling and analysis for complex data problems. **Springer Science & Business Media**.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005). “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.” **Bioinformatics**, 21, 3439–3440.

Eck SH, Benet-Pagès A, Flisikowski K, Meitinger T, Fries R, Strom TM (2009) Whole genome sequencing of a single *bos taurus* animal for single nucleotide polymorphism discovery. **Genome Biology** 10(8): R82.

Efron B, Tibshirani R (1983). An Introduction to the bootstrap. Chapman and Hall, New York.

Empresa Brasileira De Pesquisa Agropecuária (2015). Pesquisadores brasileiros sequenciam o genoma do Gir Leiteiro. **Produção animal**.

Felius M (1985). Genus Bos: Cattle breeds of the world.

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. **Nature Review Genetics** 7: 85–97.

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. **Genetics** 136: 245–247.

Habier D, Fernando RL, Dekkers JC (2009) Genomic selection using low-density marker panels. **Genetics** 182: 343–353.

Hayes B, Goddard M (2010). Genome-wide association and genomic selection in animal breeding. **Genome** 53(11): 876-883.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science** 92: 433-43.

Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, Van der Werf JHJ (2012) Accuracy of genotype imputation in sheep breeds. **Animal Genetics** 80: 43-72.

Henderson CR (1975) Best Linear Unbiased estimation and prediction under a selection model. **Biometrics** 31: 423-439.

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, 32:69-83.

Hollifield MK, Bermann M, Lourenco D, Misztal I. Impact of blending the genomic relationship matrix with different levels of pedigree relationships or the identity matrix on genetic evaluations. **JDS Commun.** 2022 Jul 26;3(5):343-347

Interbull Centre. (2021). The national genomic evaluation forms provided by countries.

Jiang L, Jiang J, Yang J, Liu X, Wang H, Ding X, Liu J, Zhang Q (2013) Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. **BMC Genomics** 14: 131.

Jiang L, Liu J, Sun D. (2010) Genome wide association studies for milk production traits in Chinese Holstein population. **PloS one**. 27;5(10):e13661

Lagrotta MR, Euclides RF, Verneque RS, Santana Júnior ML, Pereira RJ, Torres RA (2010) Relação entre características morfológicas e produção de leite em vacas da raça Gir. **Pesquisa Agropecuária Brasileira**. v.45, n.4, p.423-429.

Larmer SG (2016) Next generation sequencing data in bovine: quality control, imputation, and application. 142 f. **Tese (Doctor of Philosophy) – University of Guelph, Canadá.**

Lee K, Nguyen DT, Choi M, Cha SY, Kim JH, Dadi H, Seo HG, Seo K, Chun T, Park C. (2013). Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. **BMC Genomics**.

Lee SH, Clark S, van der Werf JH. Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. **PloS one**. 2017;12:e0189775.

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science** 92: 4656–4663.

Legarra, A.; Lourenco, D.A.L.; Vitezica, Z. Bases for Genomic Predictions. 2018. Available online: <http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=gsip.pdf> (accessado dia 09/09/2023)

Li B, Zhang N, Wang Y-G, George AW, Reverter A and Li Y (2018) Genomic Prediction of Breeding Values Using a Subset of SNP Identified by Three Machine Learning Methods. **Frontiers in Genetics** 9: 237.

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., and Jason, H. M. (2016). Detecting gene-gene interactions using a permutation-based random forest method. **BioData Mining**. 9:14.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. **R News** 2: 18–22.

Liu Z, Goddard ME, Reinhardt F, Reents R (2014). A single-step genomic model with direct estimation of marker effects. **Journal of Dairy Science**, 97(9), 5833-5850.

Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen TH. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross validation. **Genetics**.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile MM, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. **BMC Genomics** 17(1): 1-21. doi: 10.1186/s12864-016-2443-6

Malécot G. (1948) Les Mathématiques de L'hérédité Masson et Cie, Paris, France (1948)

Malhado, C.H.M., Carneiro, P.L.S., Malhado, A.C.M., Martins, J.A.M., Martins Filho, R., Bozzi, R., 2010. History of registered Gir breed in Brazilian Northeast: population structure and genetic. **Ciência Rural**. v.40, n.6, p.1385-1391

Maturana EL, Ibáñez-Escriche N, González-Recio Ó, Marenne G, Mehrban H, Chanock SJ, Malats N (2014). Next generation modeling in GWAS: comparing different genetic architectures. **Human genetics**, 133, 1235-1253.

McWhorter TM, Bermann M, Garcia ALS, Legarra A, Aguilar I, Misztal I, Lourenco D. (2023). Implication of the order of blending and tuning when computing the genomic relationship matrix in single-step GBLUP. **Journal of Dairy Science**.

McWhorter, T. M., Bermann, M., Garcia, A. L., Legarra, A., Aguilar, I., Misztal, I., & Lourenco, D. (2023). Implication of the order of blending and tuning when computing the genomic relationship matrix in single-step GBLUP. **Journal of Animal Breeding and Genetics**, 140(1), 60-78.

Meuwissen THE, Goddard ME, Hayes BJ (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819-1829.

Misztal I (2012) BLUPF90 - A flexible mixed model program in Fortran 90. Athens, GA: University of Georgia. Disponível em: <http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90.pdf>

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2009) “BLUPF90 and related programs (BGF90)” **Proceedings of the 7th World Congress on Genetics Applied to Livestock Production**. Montpellier: INRA, 21–22p.

Misztal I, Bradford HL, Lourenco DAL, Tsuruta S, Masuda Y, Legarra A, Lawlor TJ (2017) Studies on inflation of GEBV in single-step GBLUP for type. **Interbull Bull**. 51:38–42.

Misztal, I. BLUPF90 - a flexible mixed model program in Fortran 90. Manual. University of Georgia. 2012

National Center for Biotechnology Information Disponível em: <
<https://www.ncbi.nlm.nih.gov/gene/>> Acesso em: 19 Set. 2023

Nayeri S, Sargolzaei M, Abo-Ismael MK. (2016) Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. **BMC Genetics**. 17,75.

Nishio M, Inoue K, Ogawa S, Ichinoseki K, Arakawa A, Fukuzawa Y, Kobayashi E, Taniguchi M, Oe M, Ishii K (2023). Comparing pedigree and genomic inbreeding coefficients, and inbreeding depression of reproductive traits in Japanese Black cattle. **BMC Genomics**, 24(1).

Ogawa S, Matsuda H, Taniguchi Y, Watanabe T, Nishimura S, Sugimoto Y, Iwaisaki H (2014) Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese black beef cattle. **BMC Genetics** 15: 1-13.

Panetto JCC, Silva, MVGB da, Verneque RS, Machado MA, Fernandes AR, Martins MF, Reis DRL, Arbex WA, Maciel RS, Ventura HT, Pereira MA (2018). Programa Nacional de Melhoramento do Gir Leiteiro: **Sumário Brasileiro de Touros: resultado do teste de progênie**: abril 2018. Juiz de Fora: Embrapa Gado de Leite, 96p. (Embrapa Gado de Leite. Documentos, 216).

Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R (2013) Imputation of high-density genotypes in the Fleckvieh cattle population. **Genetics Selection Evolution** 13: 30-45.

Pereira, R.J.; Verneque, R.S.; Lopes P.S.; Santana, M.L.; Lagrotta J.R.M.R.; Torres, R.A.; Vercesi Filho, A.E.; Machado, M.A. Milk yield persistency in Brazilian Gyr cattle based on a random regression model. **Genetics and Molecular Research** v.11, p.1599-1609, 2012.

Preisinger R (2012) Genome-wide selection in poultry. **Animal Production Science** 52: 121–125.

Reis Filho, J.C., Lopes, P.S., Verneque, R.S., Torres, R.A., Teodoro, R.L., Carneiro, P.L.S., 2010. Population structure of Brazilian Gir dairy cattle. **Revista Brasileira de Zootecnia**, 39, 2640–2645.

Resende MDV, Lopes OS, Silva RL, Pires IE (2008) Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira** 56: 63-77.

Resende MDV, Silva FF, Lopes OS, Azevedo CF. Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p.

Ribeiro AB, Tinoco AFF, Lima GFC, Guilhermino MM, Rangel AHN (2009) Produção e composição do leite de vacas gir e guzerá nas diferentes ordens de parto. **Revista Caatinga**. v.22, n3,p46-51.

Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang, CK (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. **Nature Genetics** 43(11): 1066-1073.

Rosen Benjamin D, Bickhart Derek M, Schnabel Robert D, Koren Sergey, Elisk Christine G, Tseng Elizabeth, Rowan Troy N, et al. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. **GigaScience**, 9(3).

Sahana G, Gulbrandsen B, Thomsen B, Holm L, Panitz F, Brøndum RF, Bendixen C, Lund MS (2014) Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. **Journal of Dairy Science** 97(11): 7258-7275.

Samorè AB, Fontanesi L (2016) Genomic selection in pigs: state of the art and perspectives. **Italian Journal of Animal Science** 15: 211-232.

Santana Jr ML, Pereira RJ, Bignardi AB, El Faro L, Tonhati H, Albuquerque LG (2014) History, structure, and genetic diversity of Brazilian Gir cattle. **Livestock Science**, 163, 26-33.

Sargolzaei M, Chesnais JP, Schenkel FS (2014). A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, 15:478

Schaeffer R (2006) Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding Genetics** 123: 218-223.

Shao B, Sun H, Ahmad MJ, Ghanem N, Abdel-Shafy H, Du C, Deng T, Mansoor S, Zhou Y, Yang Y, Zhang S, Yang L and Hua G (2021) Genetic Features of Reproductive

Traits in Bovine and Buffalo: Lessons From Bovine to Buffalo. *Frontiers in Genetics* 12:617128

Silva-Vignato B, Cesar ASM, Afonso J, Moreira GCM, Poleti MD, Petrini, J, Coutinho, L. L. (2022). Integrative Analysis Between Genome-Wide Association Study and Expression Quantitative Trait Loci Reveals Bovine Muscle Gene Expression Regulatory Polymorphisms Associated With Intramuscular Fat and Backfat Thickness. **Frontiers in Genetics**, 13.

Su G, Guldbbrandtsen B, Gregersen V, Lund M. (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. **Journal of Dairy Science**.

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT (2012) Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? **Briefings Bioinformatics** 14: 315-326.

Tribout T, Larzul C, Phocas F (2013) Economic aspects of implementing genomic evaluations in a pig sire line breeding scheme. **Genetics Selection Evolution** 45: 40-56.

Tsuruta S, Misztal I, Aguilar I, Lawlor TJ (2011). Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. **Journal of Dairy Science**. 94:4198– 4204.

Usui T, Suzuki K, Kaji Y, Amano S, Miyata K, Heldin P, Yamashita H. Hyaluronan synthase expression in bovine eyes. *Invest Ophthalmol Vis Sci*. 1999 Mar;40(3):563-7. PMID: 10067958.

VanRaden PM (2008) Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91, pp. 4414-4423

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. **Journal Dairy Science** 92: 16–24.

VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME. (2011) **Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss**

Verneque RS, Ledic IL, Martinez ML, Teodoro RL, Paula LR De O, Cruz M, Campos J De P (2000) Programa de Melhoramento do Gir Leiteiro, In: **SIMPÓSIO NACIONAL DE MELHORAMENTO ANIMAL**, 3, 2000. Anais... Belo Horizonte: SBMA, p. 212-218.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. **Genetics Research**. 93:357–366.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu x, Liu S, Bo X, Yu G (2021). "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data." **The Innovation**, 2(3), 100141.

Xu L, Niu Q, Chen Y, Wang Z, Xu L, Li H, Xu L, Gao X, Zhang L, Gao H, et al. Validation of the Prediction Accuracy for 13 Traits in Chinese Simmental Beef Cattle Using a Preselected Low-Density SNP Panel. **Animals**. 2021; 11(7):1890.

Yang L, Mei Y, Fang Q, Wang J, Yan Z, Song Q, Lin Z, Ye G (2017). Identification and characterization of serine protease inhibitors in a parasitic wasp, *Pteromalus puparum*. **Scientific Reports**, 7(1), 15755.

Yu G, Wang L, Han Y, He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." **OMICS: A Journal of Integrative Biology**, 16(5), 284-287.

APÊNDICES

Apêndice A

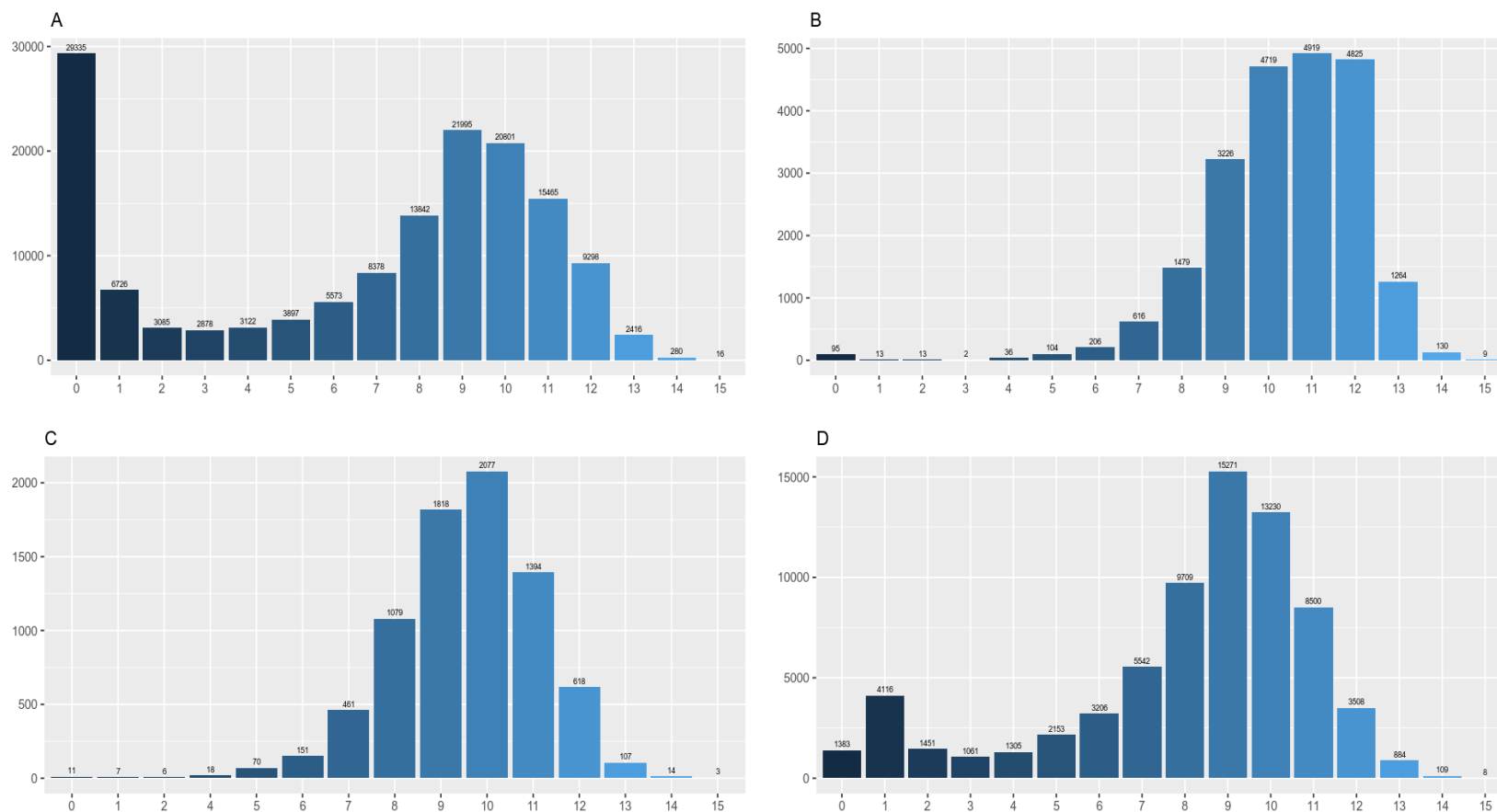


Figura 1 – Número de animais distribuídos ao longo das gerações nos dados de pedigree (A), com genótipo (B) e com genótipo e fenótipo (C) e apenas com fenótipo (D)

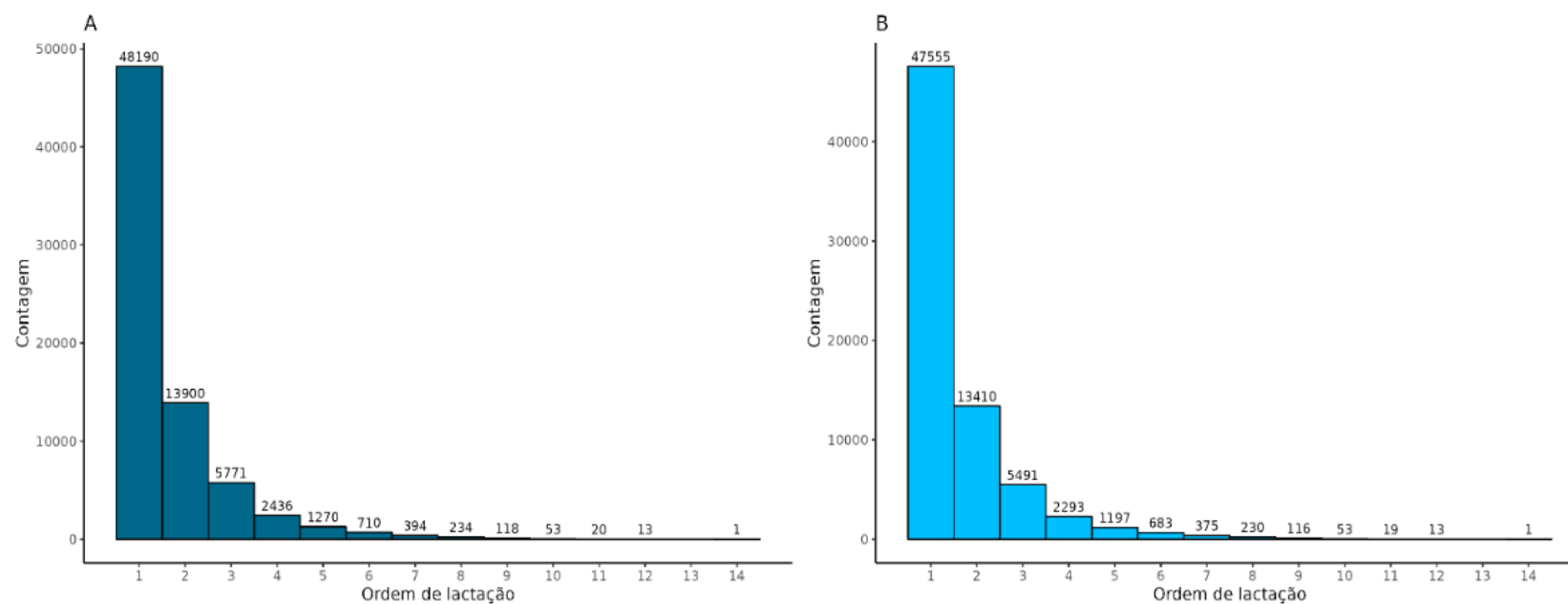


Figura 2 – Número de animais por ordem de lactação. Os gráficos A e B são os dados pré e pós controle de qualidade, respectivamente.

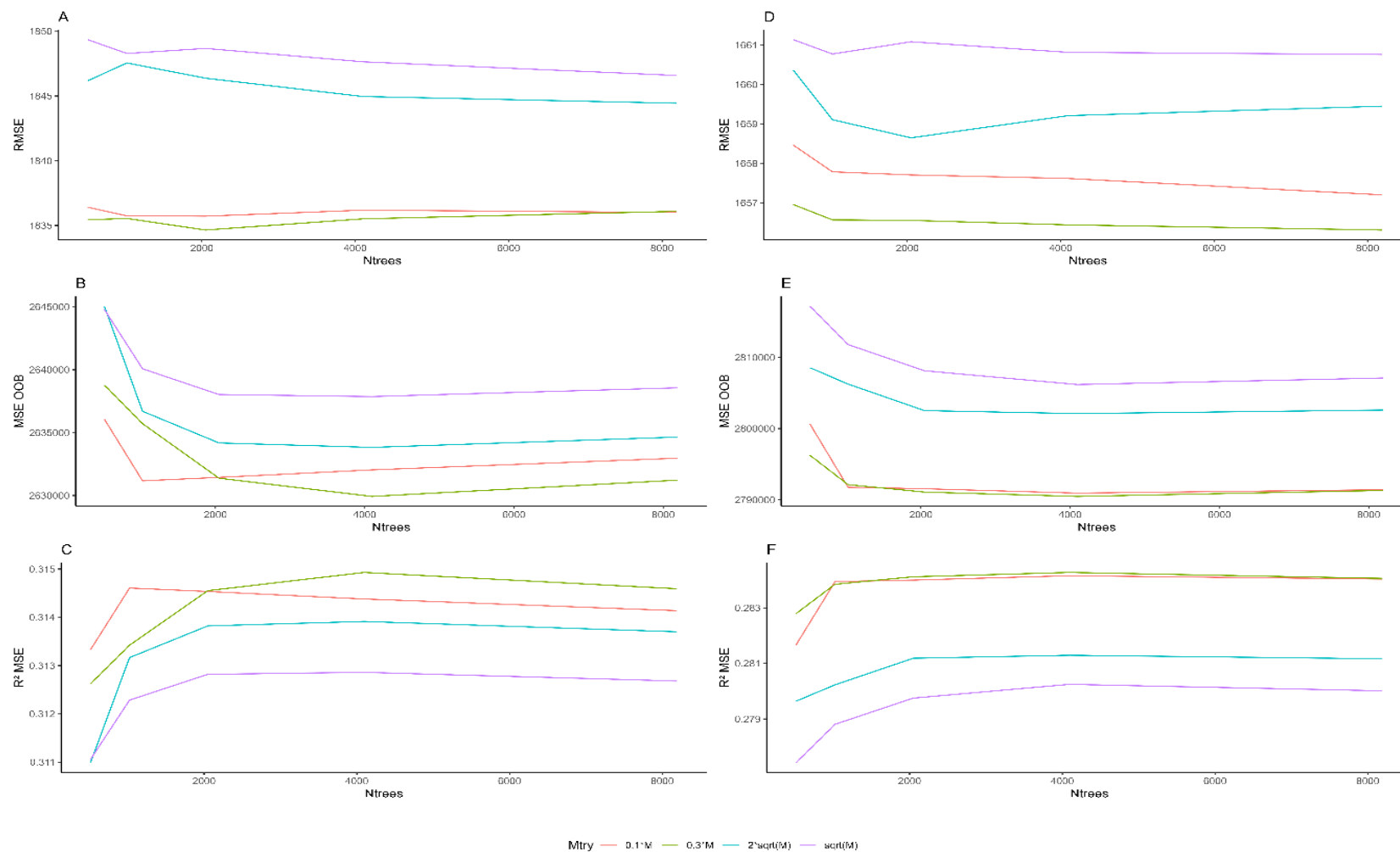


Figura 3 – Métricas de desempenho das análises de Radom Forest para os diferentes cenários testados para grupo de treinamento e validação divididos por geração em A, B e C, e dividido aleatoriamente em D, E, e F.

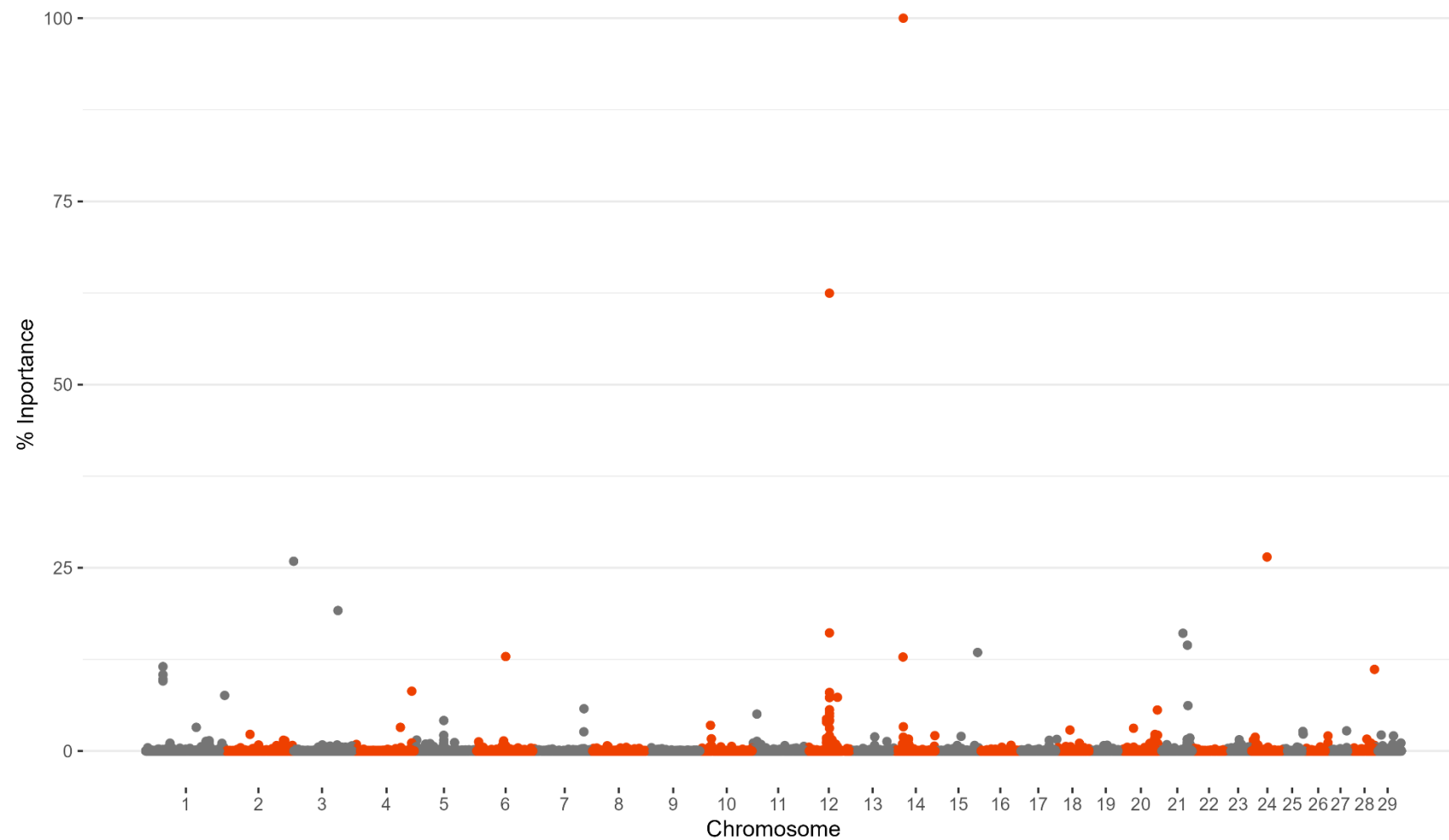


Figura 4 – Porcentagem relativa da importância dos SNP distribuídos nos cromossomos autossômicos.

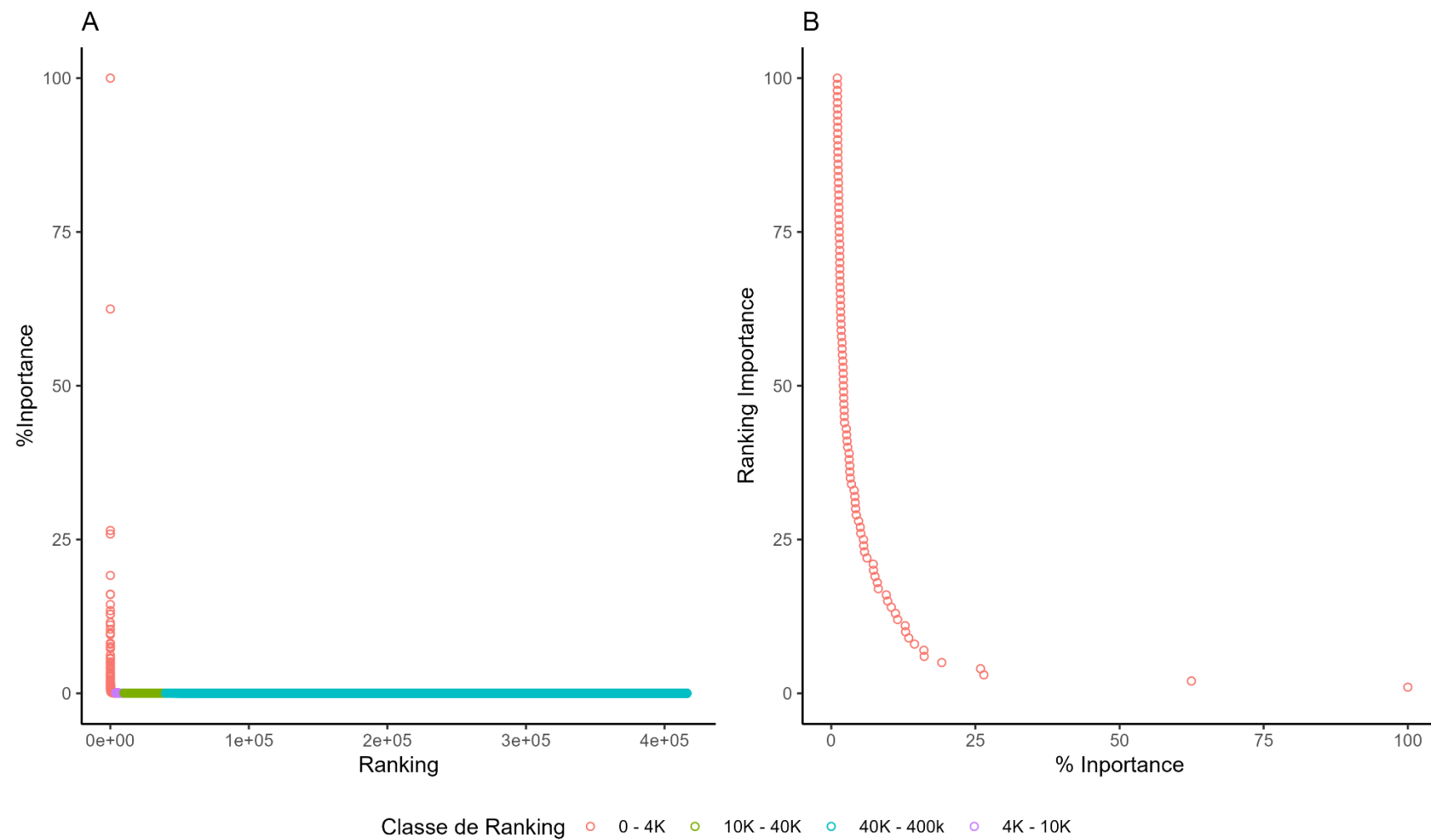


Figura 5 – Relação entre ranking e porcentagem de importância dos SNP do painel completo (A) e dos cem primeiros SNP (B)

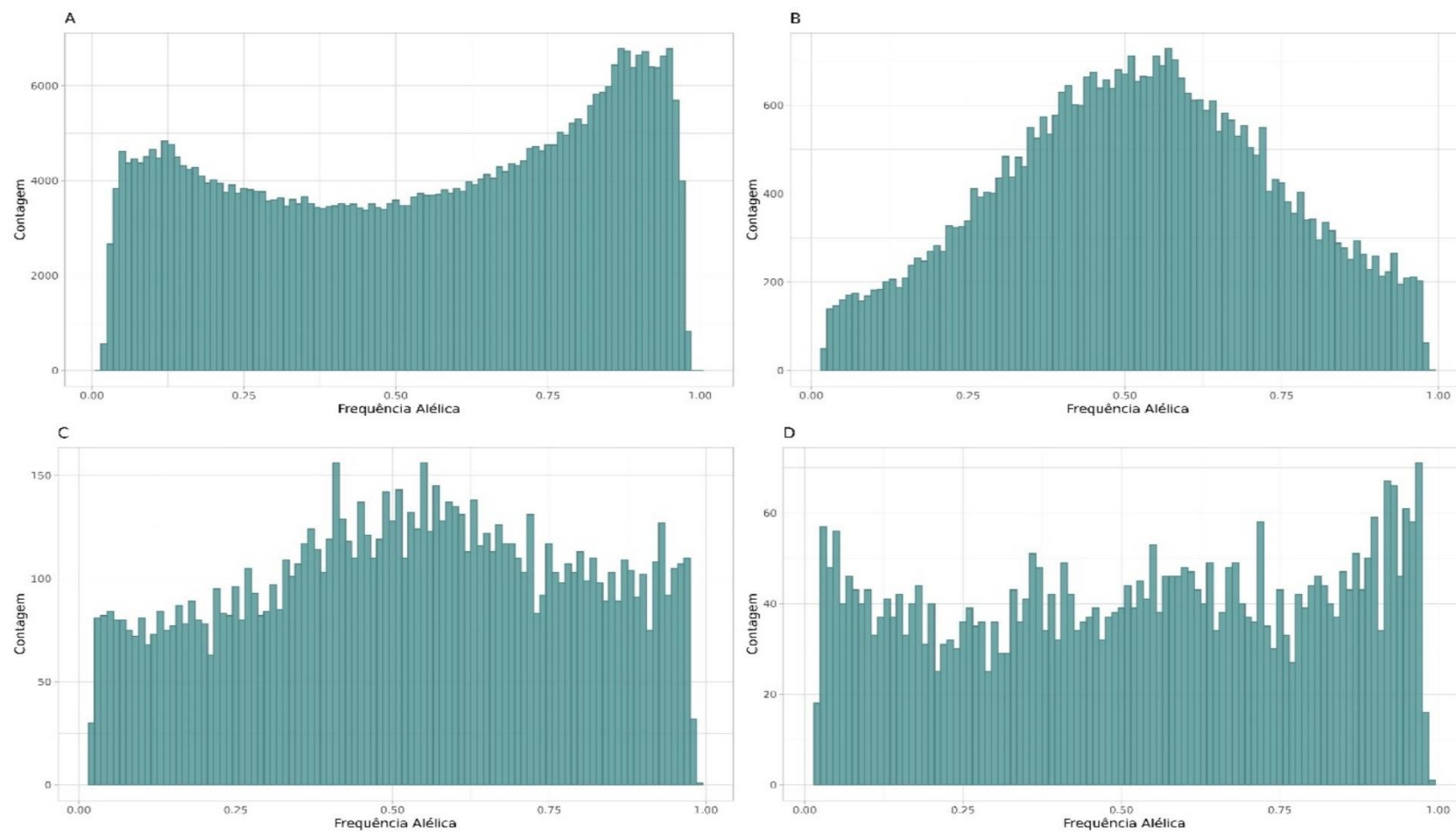


Figura 6 – Frequência alélica da população para os painéis de (A) 400k, (B) 40k, (C) 10k e (D) 4k

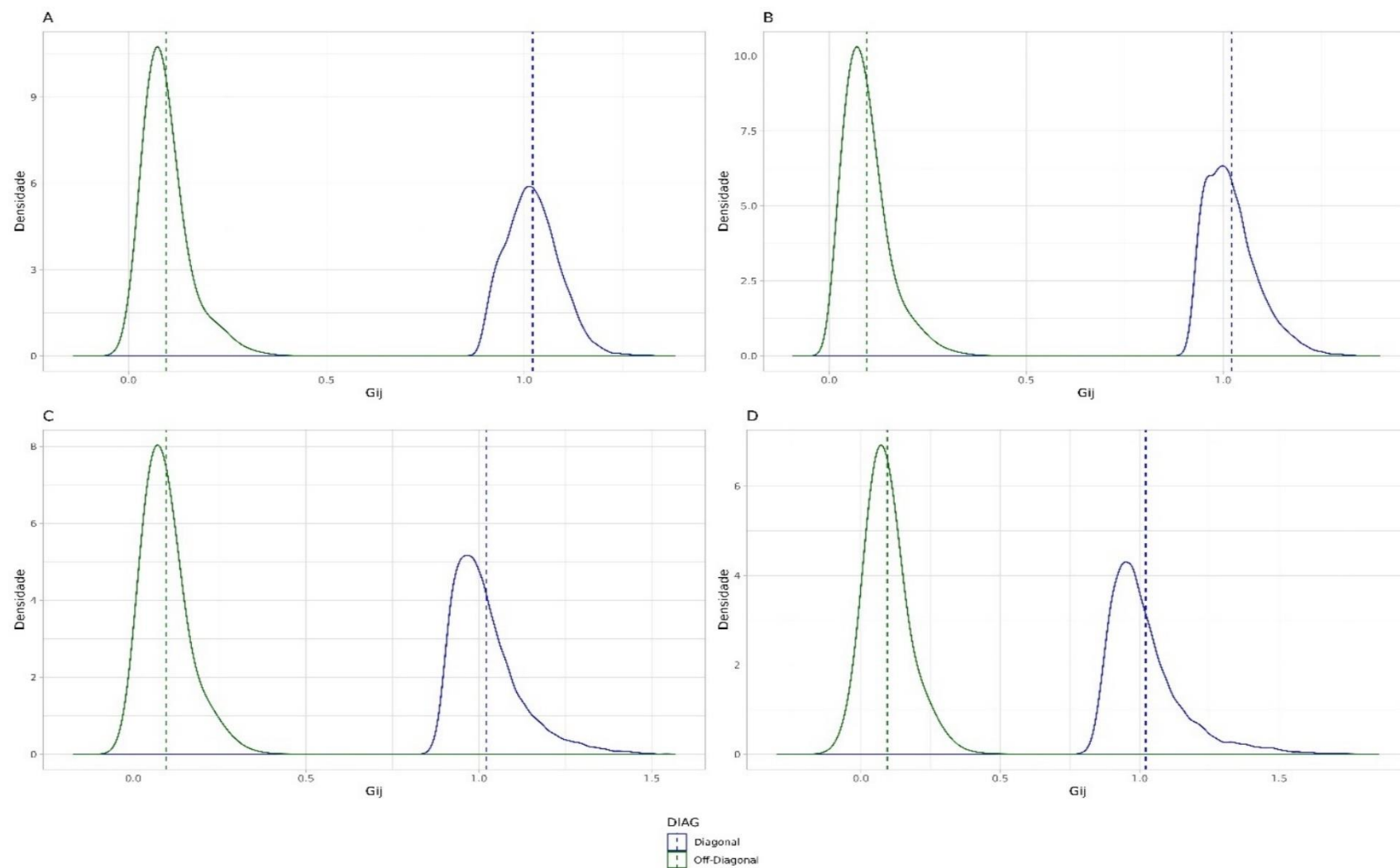


Figura 7 – Frequência dos valores observados da matriz G para os painéis (A) 400k, (B) 40K, (C) 10k e (D) 4k

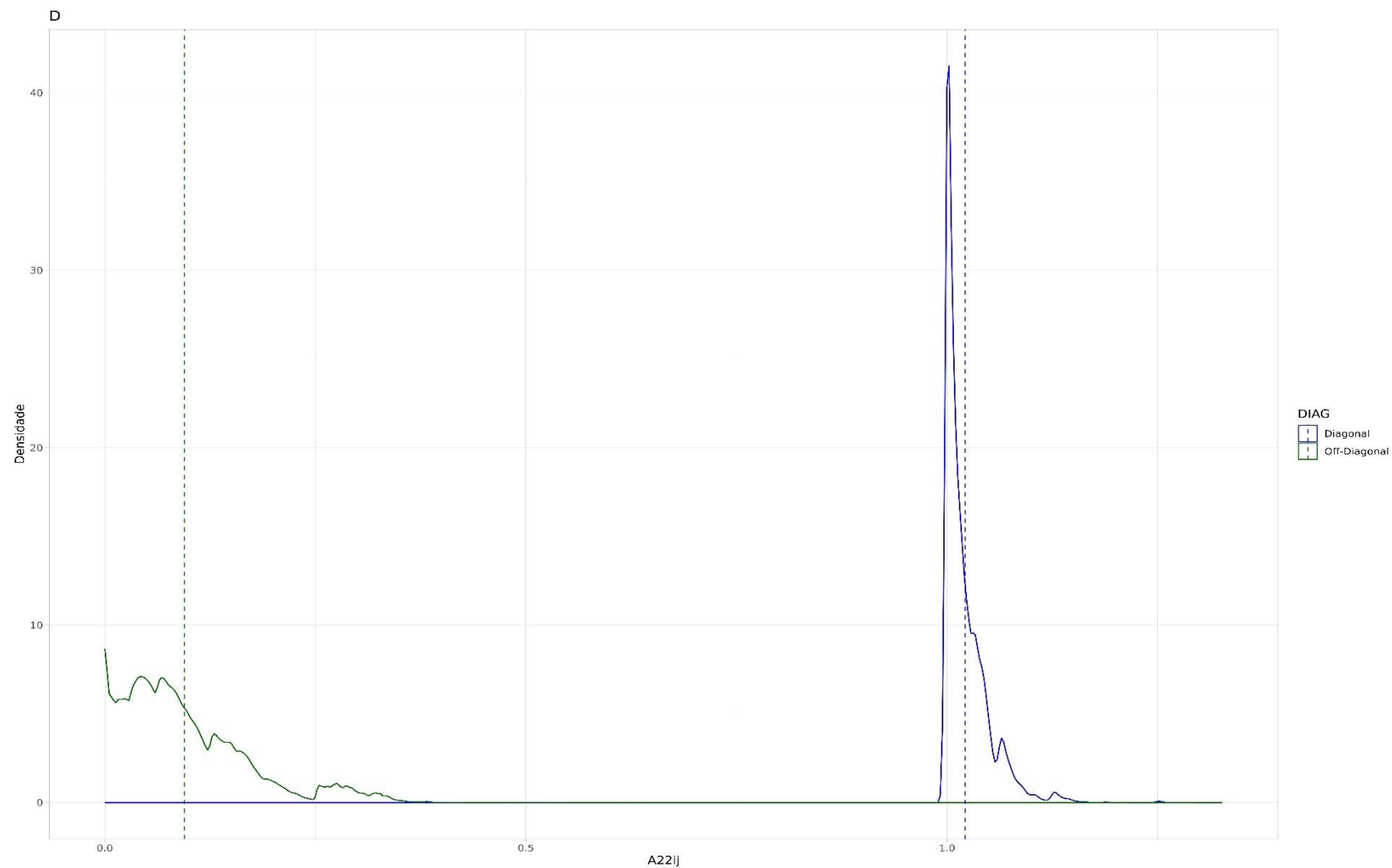


Figura 8 – Matriz de parentesco genético aditivo referente aos animais genotipados (Matriz A_{22})

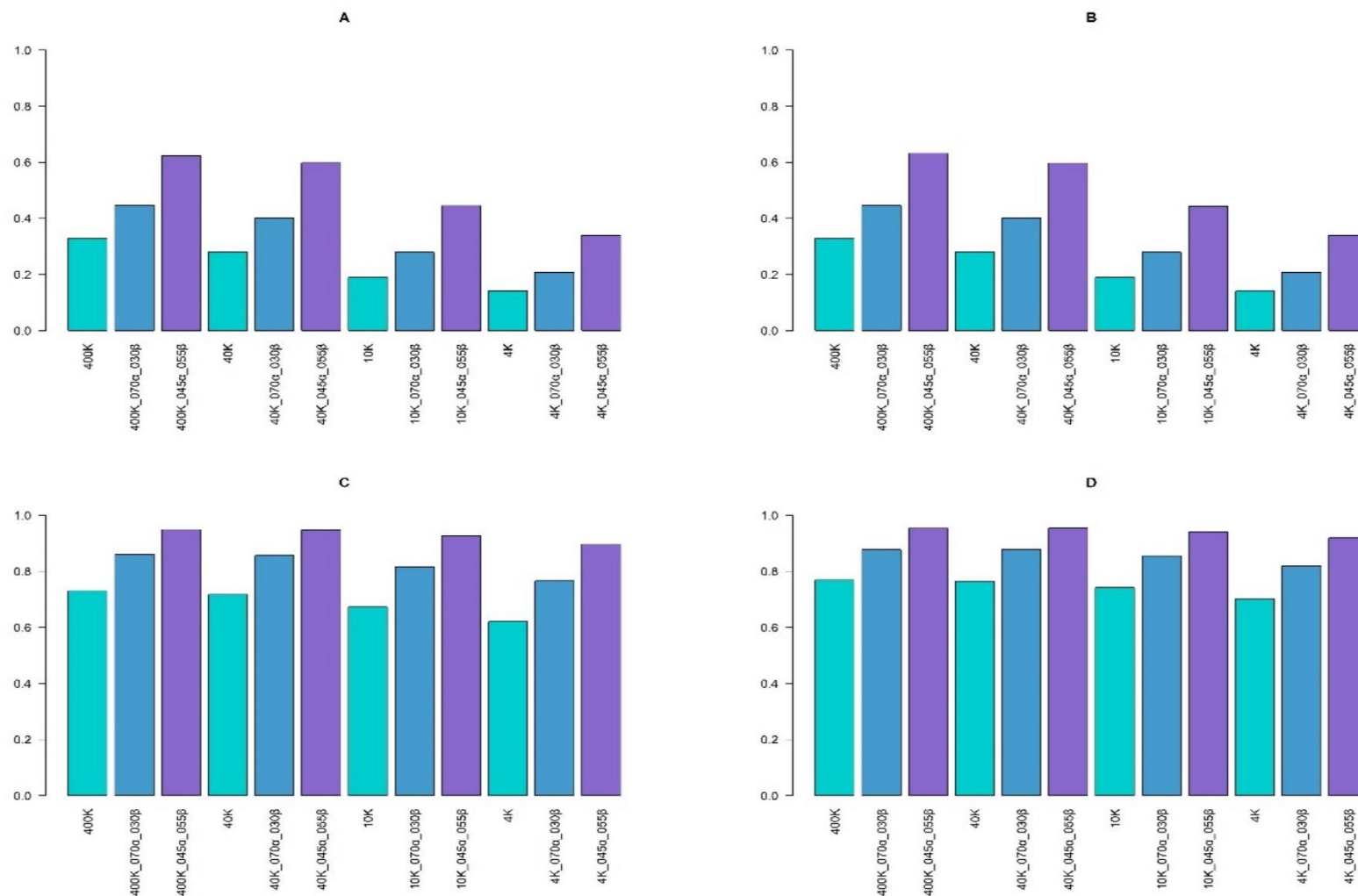


Figura 9 – (A) Correlação entre o coeficiente de endogamia genômico e genético aditivo, (B) correlação dos elementos diagonal, (C) entre todos os elementos e (D) correlação dos elementos fora da diagonal da matriz de G e A_{22}

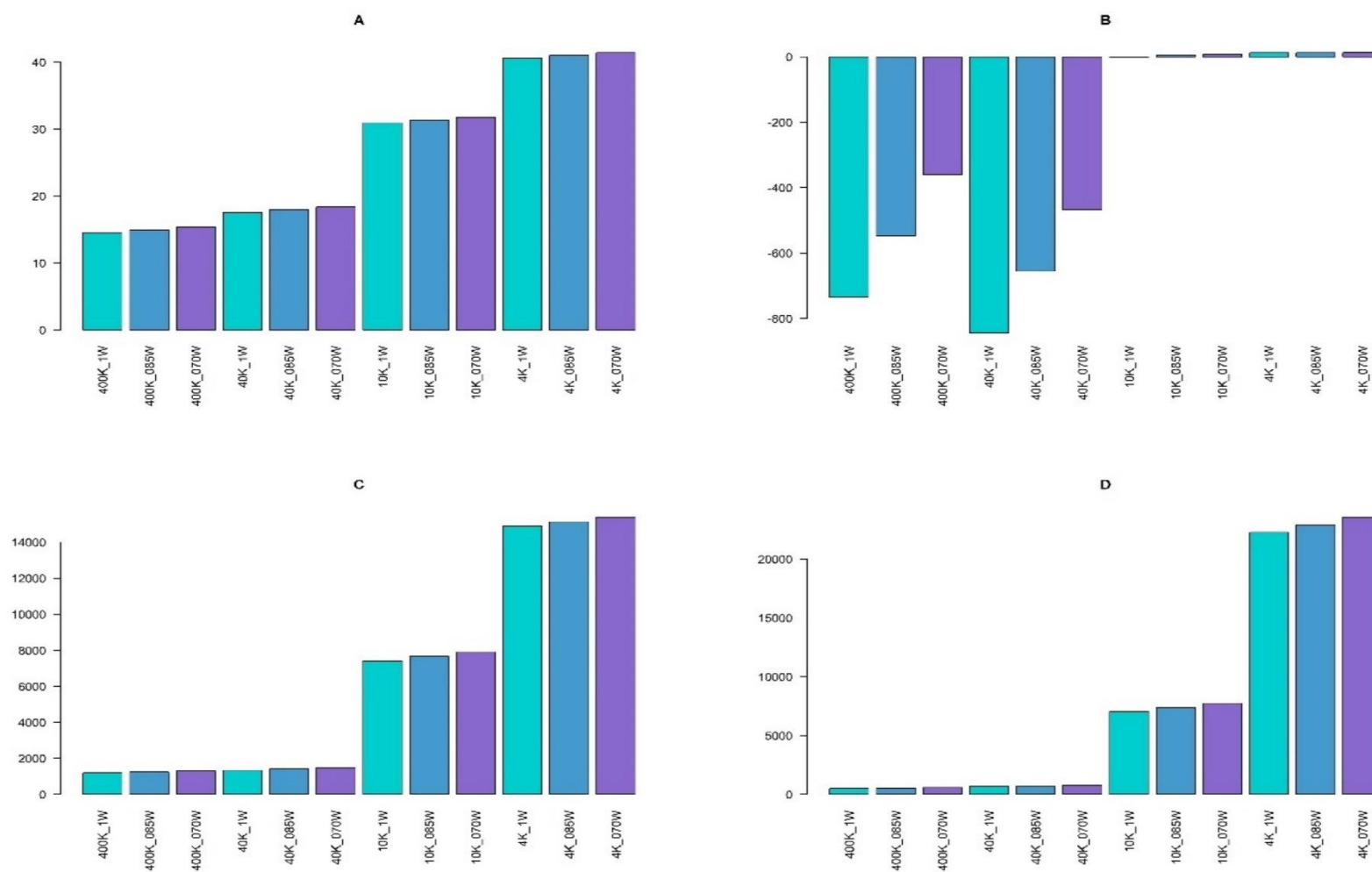


Figura 10 – Valores da (A) média, (B) mínimo, (C) máximo e (D) variância dos valores da diagonal da inversa da matriz G menos a matriz A_{22}^{-1} , para os diferentes valores de ω

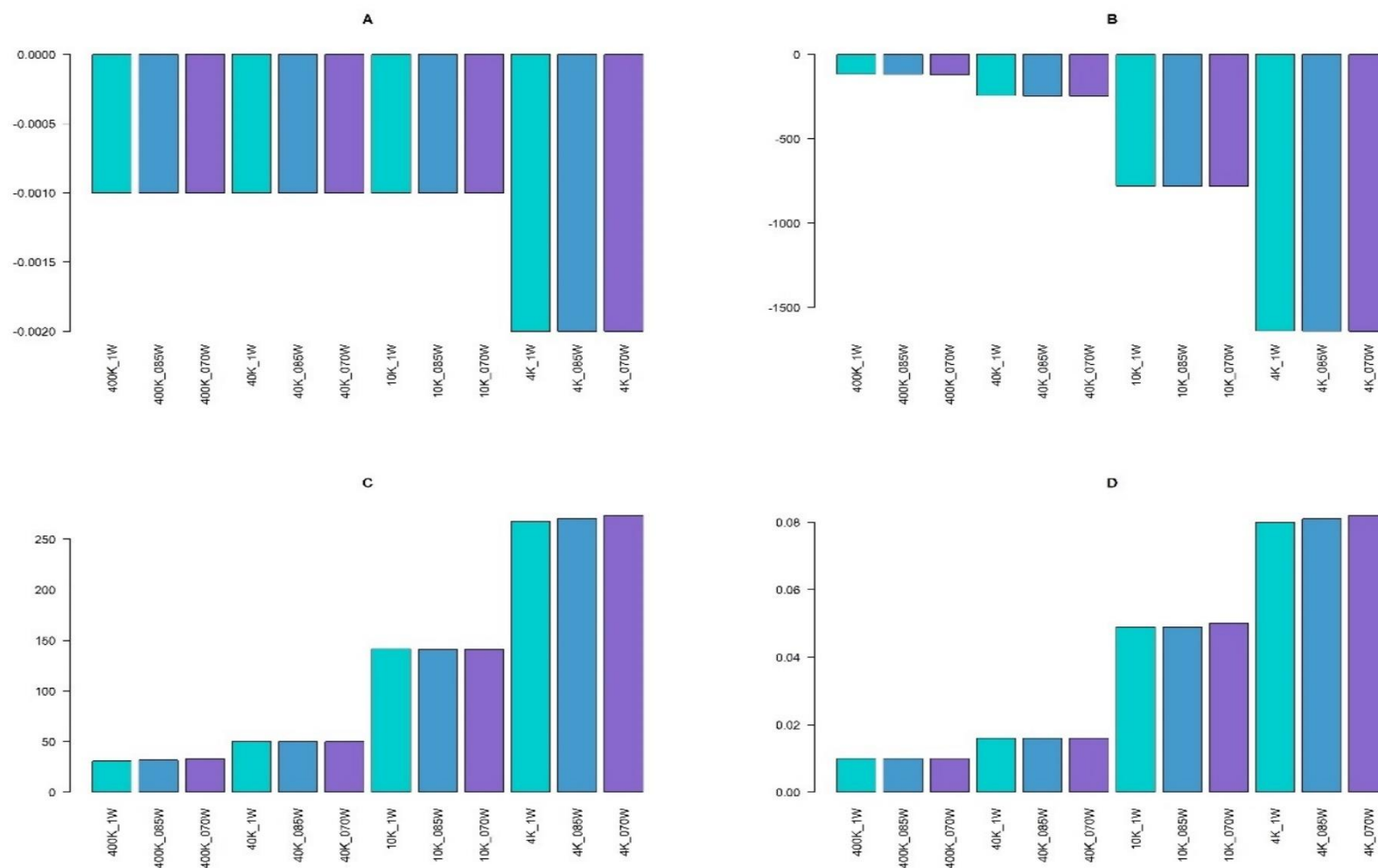


Figura 11 – Valores da (A) média, (B) mínimo, (C) máximo e (D) variância dos valores de fora da diagonal da inversa da matriz G menos a matriz A_{22}^{-1} , para os diferentes valores de ω

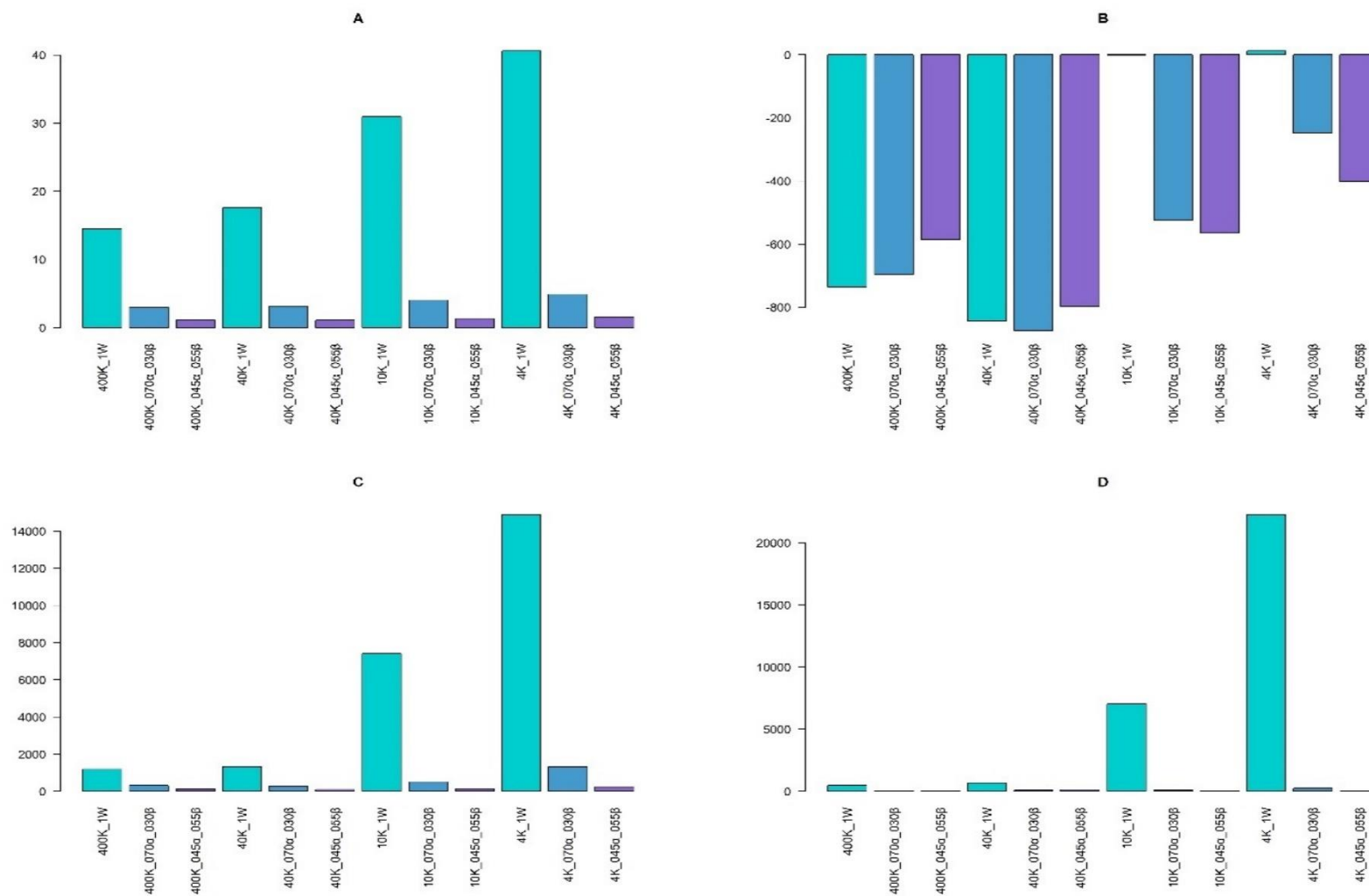


Figura 12 – Valores da (A) média, (B) mínimo, (C) máximo e (D) variância dos valores da diagonal da inversa da matriz G menos a inversa da matriz A_{22}^{-1} , para os diferentes valores de α e β

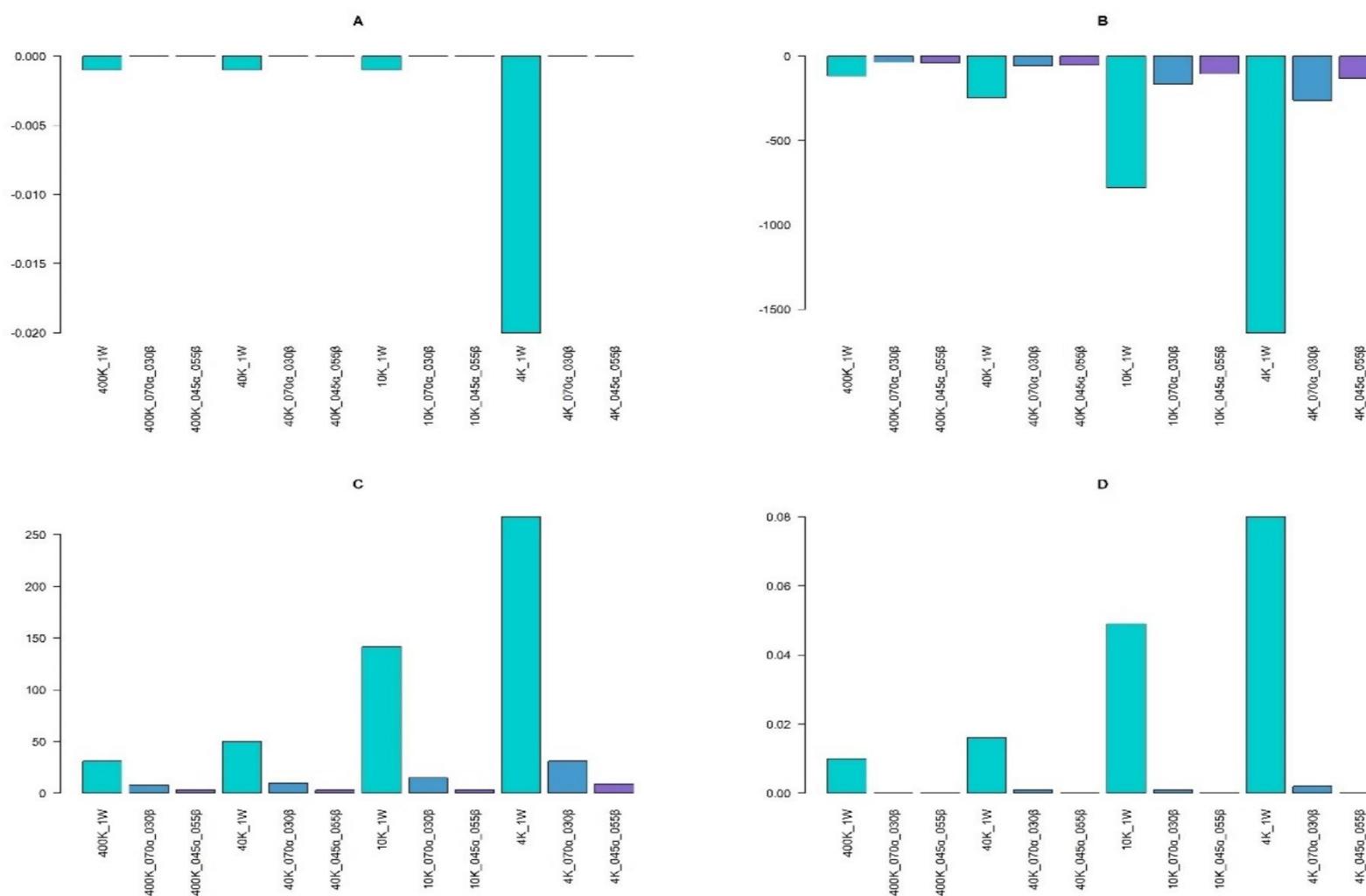


Figura 13 – Valores da (A) média, (B) mínimo, (C) máximo e (D) variância dos valores de fora da diagonal da inversa da matriz G menos a inversa da matriz A_{22}^{-1} , para os diferentes valores de α e β

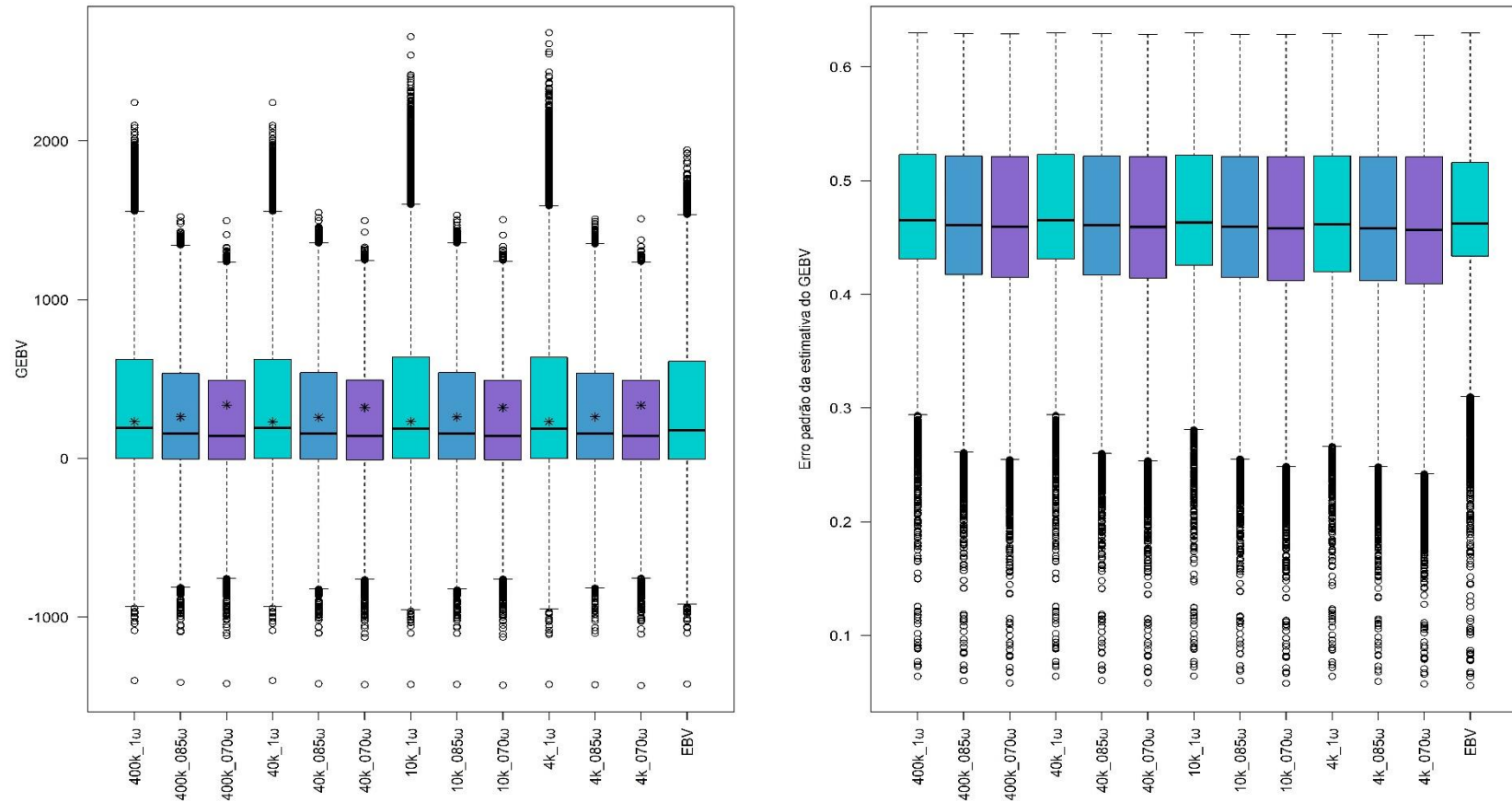


Figura 14 – Valor da estimativa do GEBV e erro padrão da estimativa nos diferentes painéis e para os diferentes valores de ω , fixados $\tau = 1$, $\alpha = 0,95$ e $\beta = 0,05$.

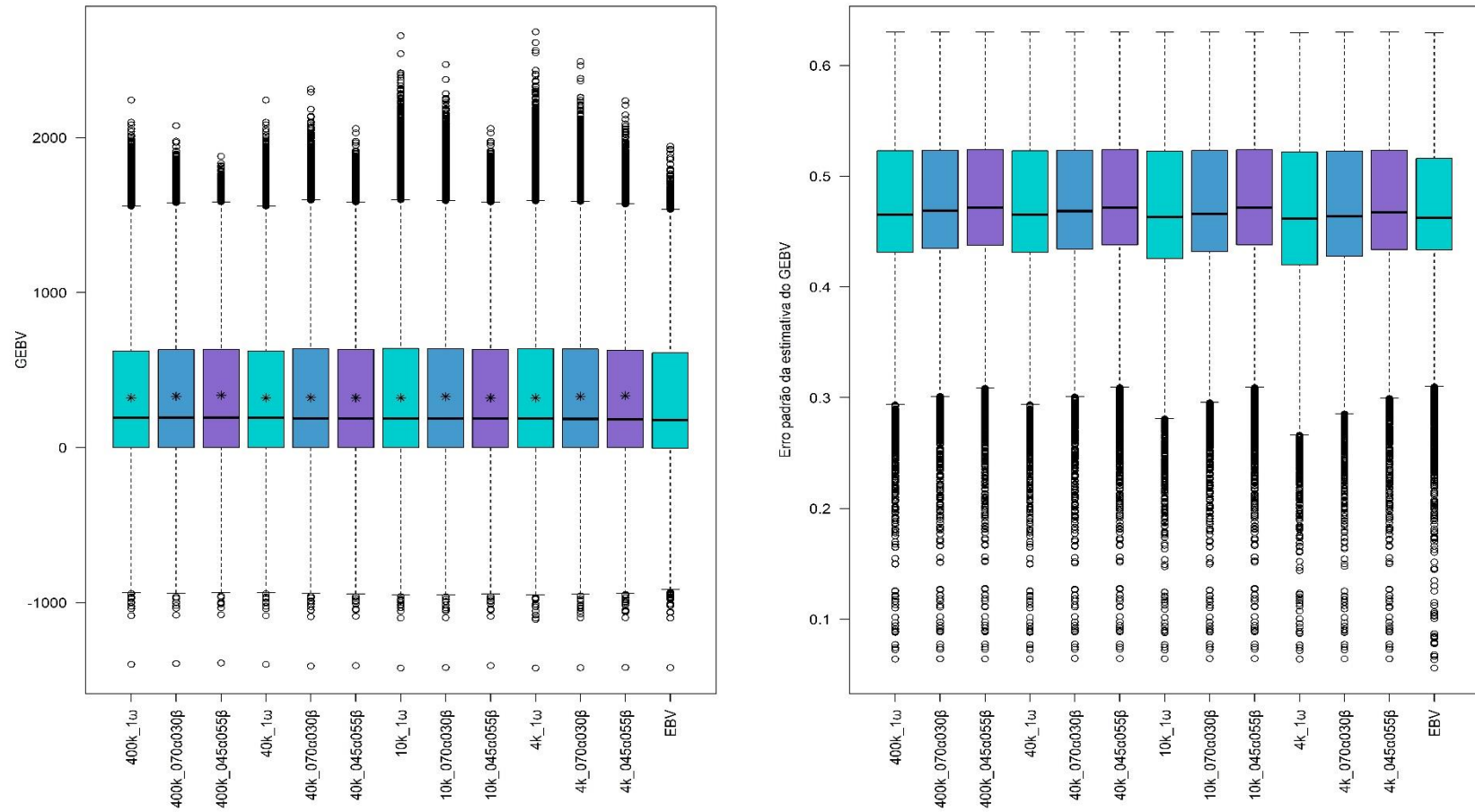


Figura 15 – Valor da estimativa do GEBV e erro padrão da estimativa nos diferentes painéis e para os diferentes valores de α e β , fixados $\tau = \omega = 1$.

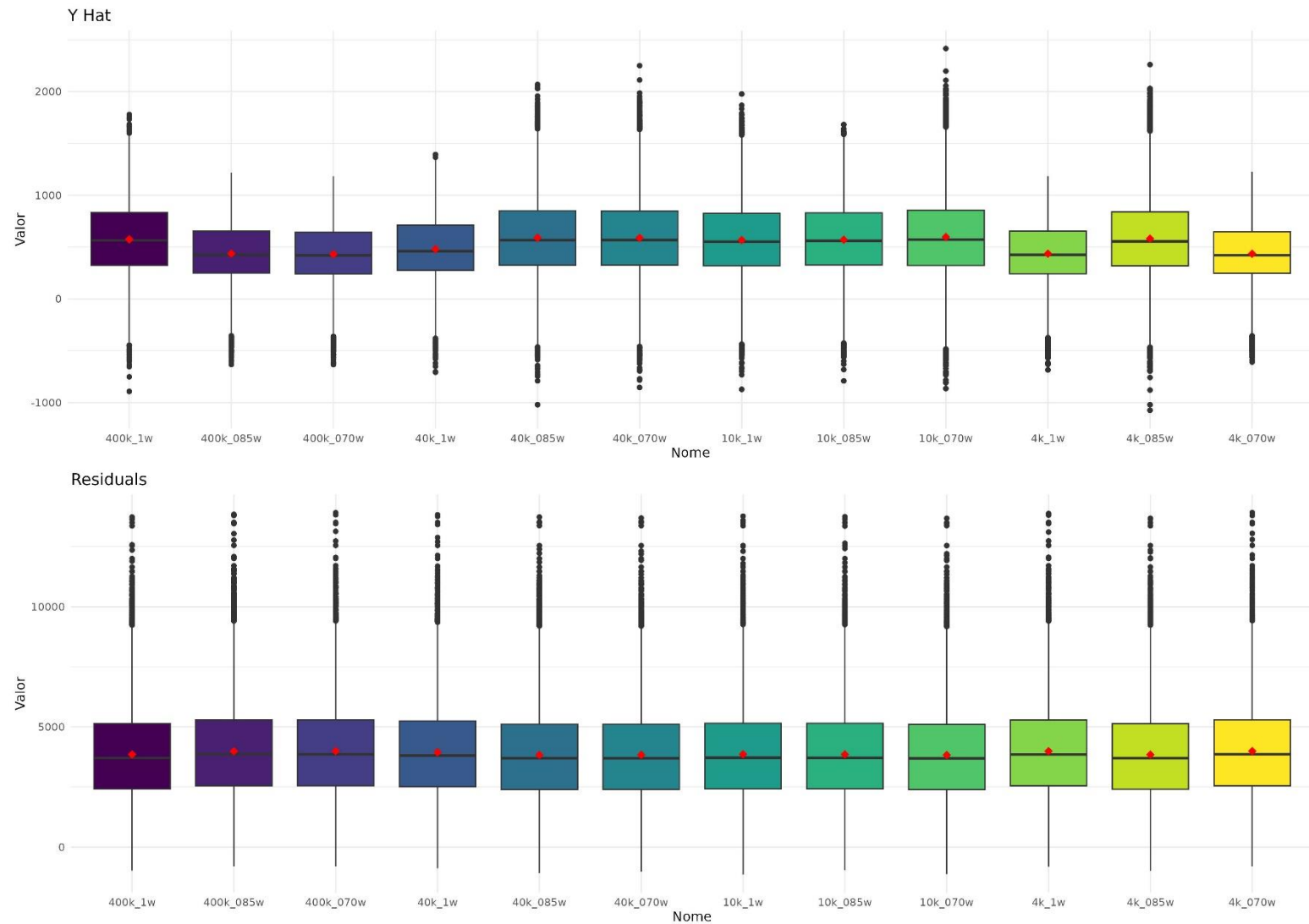


Figura 16 – Fenótipo predito (Y_{Hat}) e fenótipo descontados os efeitos fixos para os quais o modelo foi corrigido ($Residuals$) para os cenários de ω

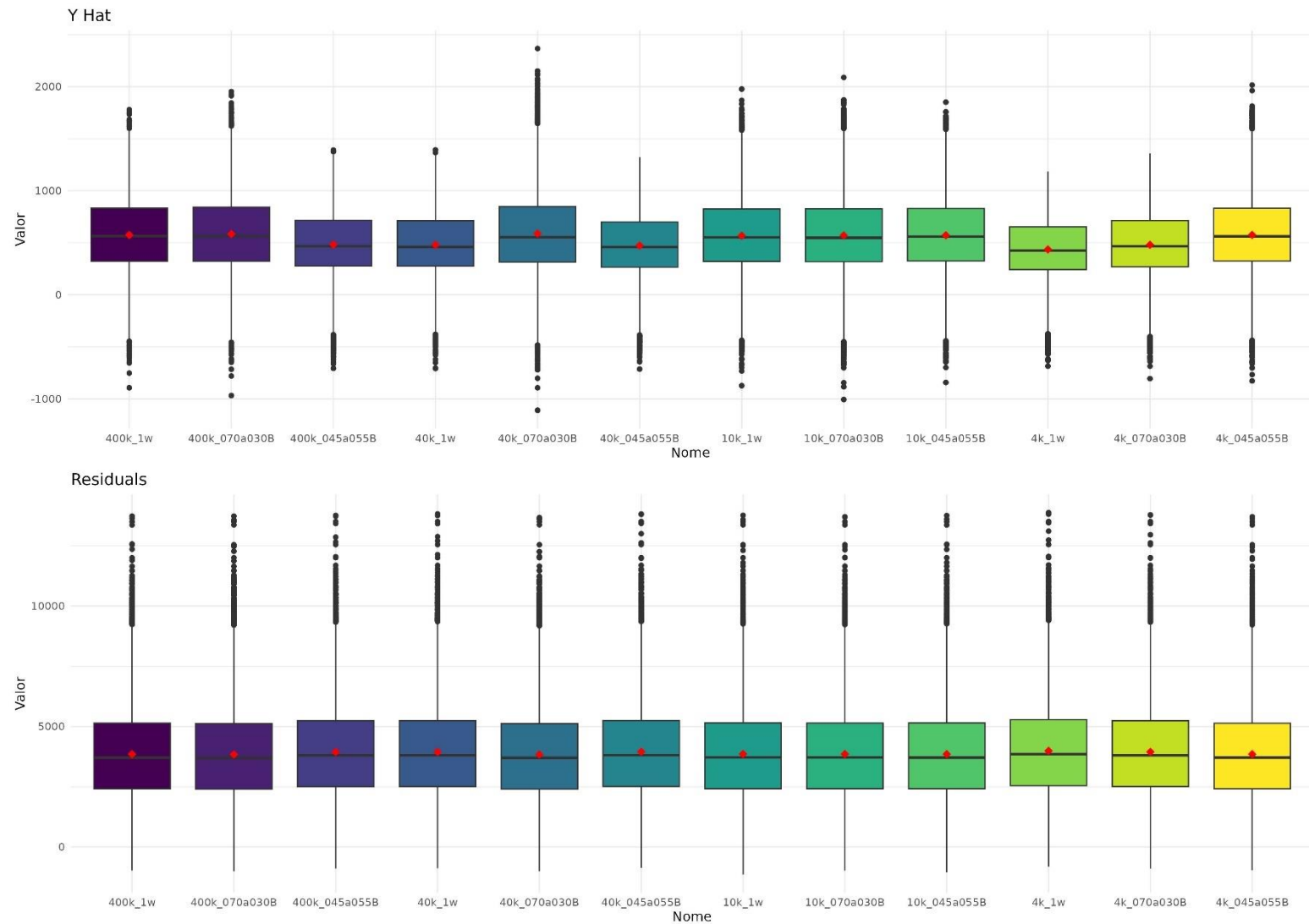


Figura 17 – Fenótipo predito ($\hat{Y}$) e fenótipo descontados os efeitos fixos para os quais o modelo foi corrigido ($Residuals$) para cenários de α e β .

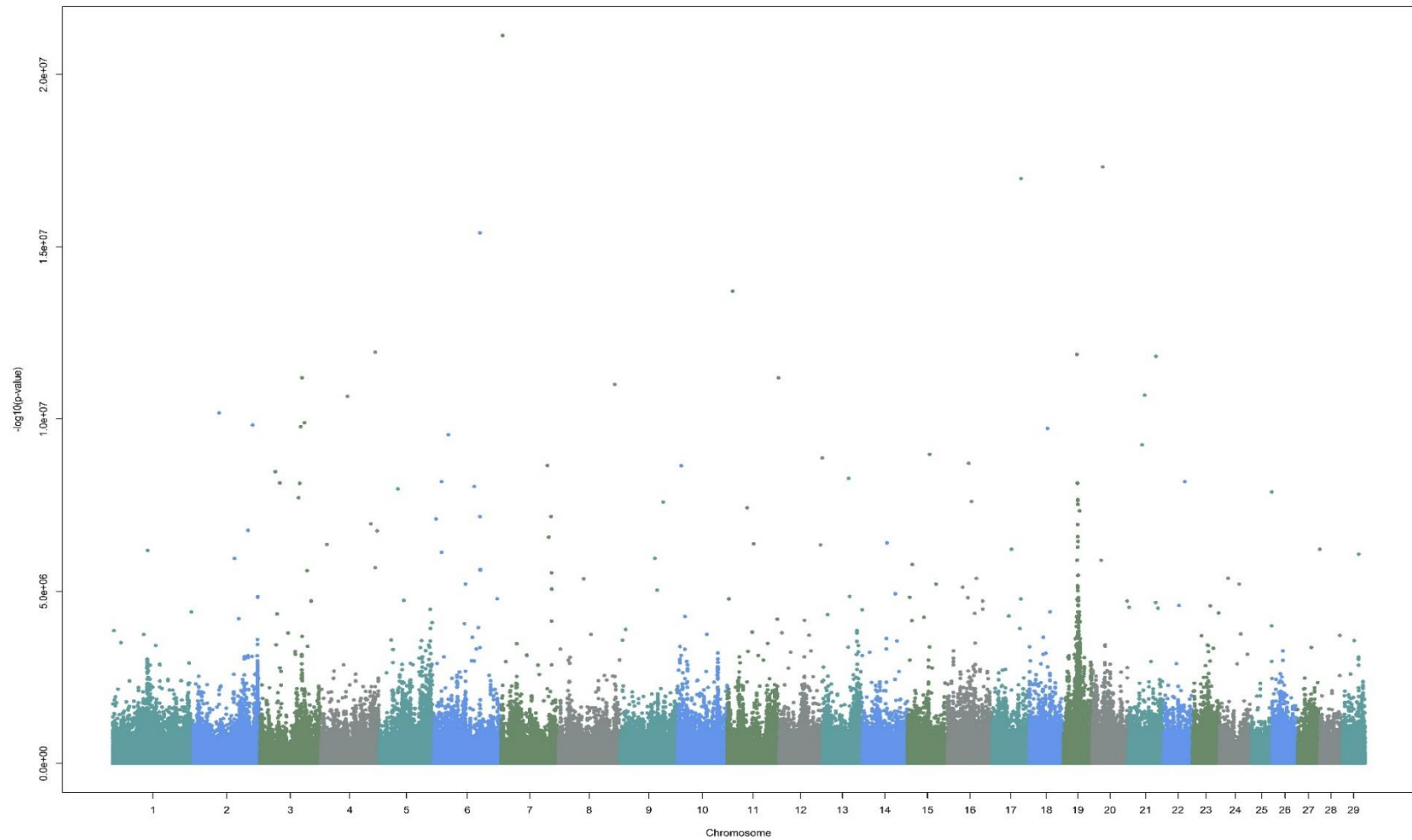


Figura 18 – Logaritmo negativo na base 10 do p-value dos SNP distribuídos nos cromossomos autossômicos.

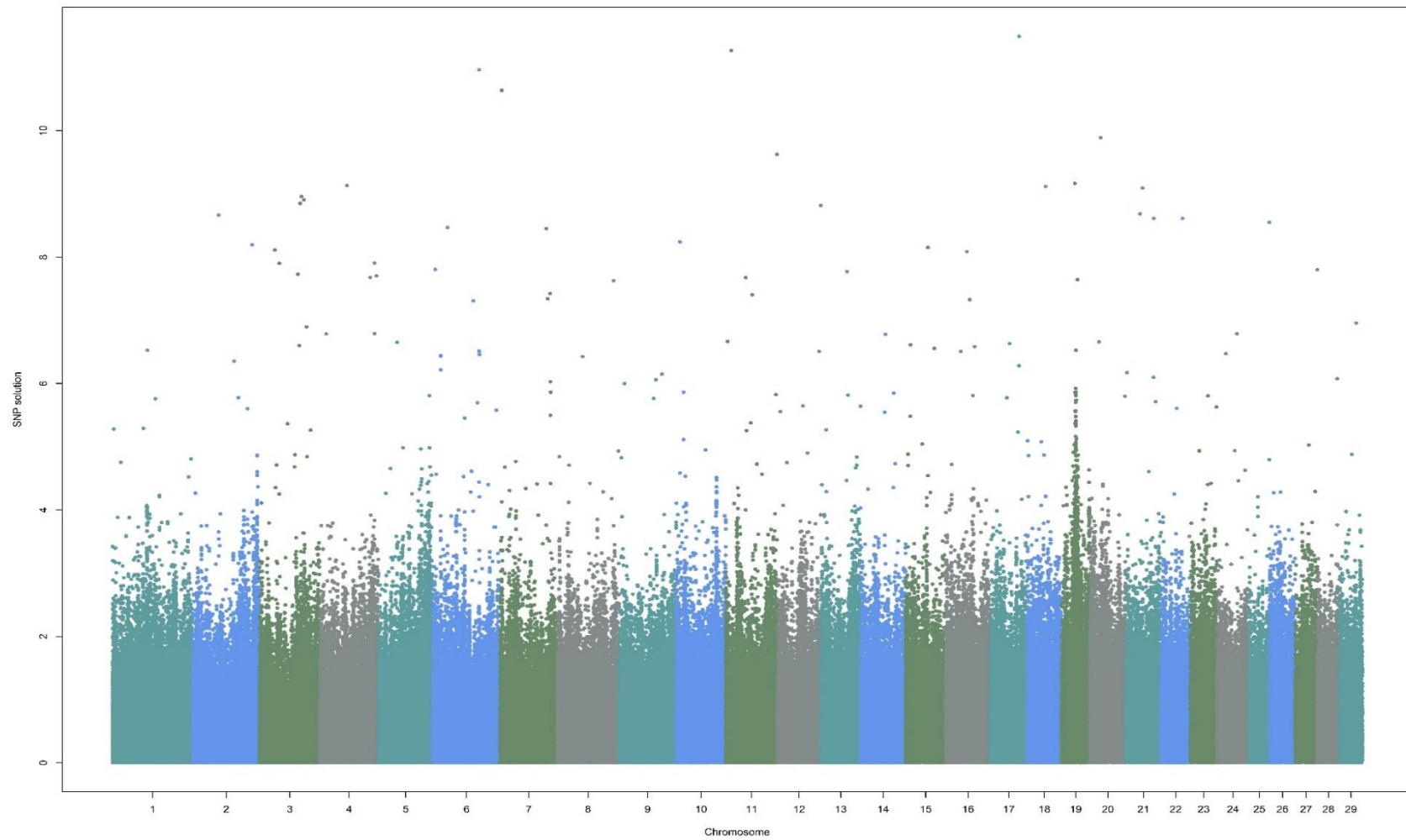


Figura 19 – Estimativa dos efeitos dos SNP distribuídos nos cromossomos autossômicos.

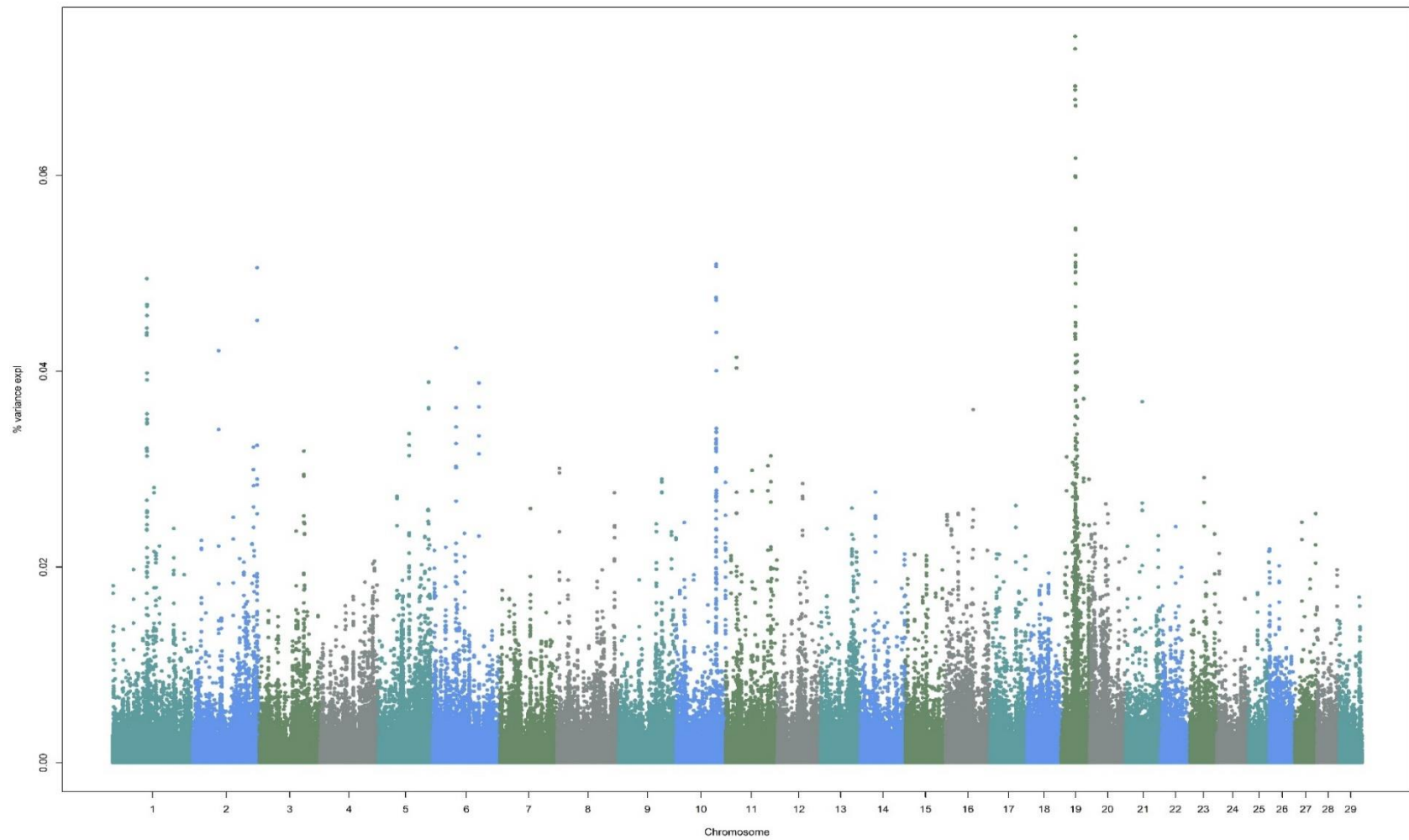


Figura 20 – Proporção da variância explicada por cinco SNP adjacentes distribuídos nos cromossomos autossômicos.

APÊNDICE B

Tabela 6 – Teste t de Student para as médias e teste de razão entre variâncias (Teste F) dos valores de GEBV para os diferentes valores de ω

	400k 1 ω	400k 085 ω	400k 070 ω	40k 1 ω	40k 085 ω	40k 070 ω	10k 1 ω	10k 085 ω	10k 070 ω	4k 1 ω	4k 085 ω	4k 070 ω
400k 1 ω	#	***	***	1	***	***	***	***	***	***	***	***
400k 085 ω	***	#	***	***	*	***	***	*	***	***	*	***
400k 070 ω	***	***	#	***	***	0,28	***	***	0,29	***	***	0,23
40k 1 ω	1	***	***	#	***	***	***	***	***	***	***	***
40k 085 ω	***	***	***	***	#	***	***	0,69	***	***	0,61	***
40k 070 ω	***	***	***	***	***	#	***	***	0,97	***	***	0,93
10k 1 ω	***	***	***	***	***	***	#	***	***	0,31	***	***
10k 085 ω	***	***	***	***	0,69	***	***	#	***	***	0,92	***
10k 070 ω	***	***	**	***	***	0,26	***	***	#	***	***	0,89
4k 1 ω	***	***	***	***	***	***	.	***	***	#	***	***
4k 085 ω	***	***	***	***	**	***	***	***	***	***	#	***
4k 070 ω	***	***	0,49	***	***	***	***	***	***	***	***	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$ e . $p < 0,1$.

Tabela 7 – Teste t de Student para as médias e teste de razão entre variâncias (Teste F) dos valores de GEBV para os diferentes valores de α e β

	400k 1 ω	400k 070 α 030 β	400k 045 α 055 β	40k 1 ω	40k 070 α 030 β	40k 045 α 055 β	10k 1 ω	10k 070 α 030 β	10k 045 α 055 β	4k 1 ω	4k 070 α 030 β	4k 045 α 055 β
400k 1 ω	#	0,30	0,89	1	***	0,84	***	***	0,84	***	***	0,70
400k 070 α 030 β	*	#	0,37	0,30	***	0,41	***	***	0,41	***	***	0,16
400k 045 α 055 β	***	***	#	0,89	***	0,95	***	***	0,95	***	***	0,60
40k 1 ω	1	*	***	#	***	0,84	***	***	0,84	***	***	0,70
40k 070 α 030 β	***	***	***	***	#	***	***	0,14	***	***	0,46	***
40k 045 α 055 β	**	***	***	**	***	#	***	***	1	***	***	0,56
10k 1 ω	***	***	***	***	***	***	#	***	***	0,31	***	***
10k 070 α 030 β	***	***	***	***	***	***	***	#	***	*	0,46	***
10k 045 α 055 β	**	***	***	**	***	1	***	***	#	***	***	0,56
4k 1 ω	***	***	***	***	***	***	,	***	***	#	**	***
4k 070 α 030 β	***	***	***	***	***	***	***	0,61	***	***	#	***
4k 045 α 055 β	***	***	***	***	*	***	***	***	***	***	***	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$ e . $p < 0,1$.

Tabela 8 – Teste t para as médias e teste de razão entre variâncias dos valores do $\hat{y}$ (predictf90) para os cenários de α e β

	400k 1 ω	400k 070 α 030 β	400k 045 α 055 β	40k 1 ω	40k 070 α 030 β	40k 045 α 055 β	10k 1 ω	10k 070 α 030 β	10k 045 α 055 β	4k 1 ω	4k 070 α 030 β	4k 045 α 055 β
400k 1 ω	#	*	***	***	**	***	.	0,30	0,50	***	***	0,79
400k 070 α 030 β	***	#	***	***	0,43	***	***	***	**	***	***	*
400k 045 α 055 β	***	***	#	0,43	***	**	***	***	***	***	0,50	***
40k 1 ω	***	***	0,27	#	***	*	***	***	***	***	0,91	***
40k 070 α 030 β	***	***	***	***	#	***	***	***	***	***	***	**
40k 045 α 055 β	***	***	**	*	***	#	***	***	***	***	*	***
10k 1 ω	0,88	***	***	***	***	***	#	0,55	0,32	***	***	.
10k 070 α 030 β	***	0,51	***	***	***	***	***	#	0,71	***	***	0,20
10k 045 α 055 β	.	***	***	***	***	***	.	***	#	***	***	0,35
4k 1 ω	***	***	***	***	***	***	***	***	***	#	***	***
4k 070 α 030 β	***	***	0,99	0,27	***	**	***	***	***	***	#	***
4k 045 α 055 β	**	0,18	***	***	***	***	**	0,49	***	***	***	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** p < 0,001, ** p < 0,01, * p , 0,05 e . p < 0,1.

Tabela 9 – Teste t para as médias e teste de razão entre variâncias dos valores $\hat{y}$ (predictf90) para os diferentes valores de ω (Teste F)

	400k 1 ω	400k 085 ω	400k 070 ω	40k 1 ω	40k 085 ω	40k 070 ω	10k 1 ω	10k 085 ω	10k 070 ω	4k 1 ω	4k 085 ω	4k 070 ω
400k 1 ω	#	***	***	***	***	***	.	0,29	***	***	.	***
400k 085 ω	***	#	.	***	***	***	***	***	***	0,72	***	0,52
400k 070 ω	***	0,13	#	***	***	***	***	***	***	0,15	***	0,24
40k 1 ω	***	***	***	#	***	***	***	***	***	***	***	***
40k 085 ω	***	***	***	***	#	0,76	***	***	0,11	***	.	***
40k 070 ω	***	***	***	***	0,85	#	***	***	.	***	0,10	***
10k 1 ω	0,88	***	***	***	***	***	#	0,54	***	***	***	***
10k 085 ω	***	***	***	***	***	***	***	#	***	***	**	***
10k 070 ω	***	***	***	***	***	***	***	***	#	***	***	***
4k 1 ω	***	0,53	*	***	***	***	***	***	***	#	***	0,78
4k 085 ω	***	***	***	***	0,48	0,60	***	***	***	***	#	***
4k 070 ω	***	0,12	0,98	***	***	***	***	***	***	*	***	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** p < 0,001, ** p < 0,01, * p , 0,05 e . p < 0,1.

Tabela 10 – teste t para as médias e teste de razão entre variâncias dos valores do resíduo (predictf90) para os cenários de α e β (Teste F)

	400k 1 ω	400k 070 α 030 β	400k 045 α 055 β	40k 1 ω	40k 070 α 030 β	40k 045 α 055 β	10k 1 ω	10k 070 α 030 β	10k 045 α 055 β	4k 1 ω	4k 070 α 030 β	4k 045 α 055 β
400k 1 ω	#	0,64	***	***	0,53	***	0,75	0,84	0,90	***	***	0,96
400k 070 α 030 β	0,94	#	***	***	0,87	***	0,43	0,50	0,55	***	***	0,67
400k 045 α 055 β	0,90	0,84	#	0,90	***	0,64	***	***	***	*	0,91	***
40k 1 ω	0,84	0,79	0,95	#	***	0,74	***	***	***	*	0,98	***
40k 070 α 030 β	0,88	0,94	0,78	0,73	#	***	0,34	0,40	0,45	***	***	0,56
40k 045 α 055 β	0,73	0,68	0,83	0,89	0,63	#	***	***	***	.	0,72	***
10k 1 ω	0,83	0,77	0,93	0,98	0,71	0,91	#	0,91	0,85	***	***	0,71
10k 070 α 030 β	0,91	0,96	0,81	0,75	0,97	0,65	0,74	#	0,94	***	***	0,80
10k 045 α 055 β	0,83	0,77	0,93	0,98	0,71	0,91	1,00	0,74	#	***	***	0,86
4k 1 ω	0,77	0,71	0,87	0,92	0,66	0,97	0,94	0,68	0,94	#	*	***
4k 070 α 030 β	0,88	0,82	0,98	0,97	0,76	0,86	0,95	0,79	0,95	0,89	#	***
4k 045 α 055 β	0,88	0,94	0,78	0,73	1,00	0,63	0,71	0,97	0,71	0,66	0,76	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** p < 0,001, ** p < 0,01, * p , 0,05 e . p < 0,1.

Tabela 11 – teste t para as médias e teste de razão entre variâncias dos valores do resíduo (predictf90) para os cenários de ω

	400k 1 ω	400k 085 ω	400k 070 ω	40k 1 ω	40k 085 ω	40k 070 ω	10k 1 ω	10k 085 ω	10k 070 ω	4k 1 ω	4k 085 ω	4k 070 ω
400k 1 ω	#	***	***	***	0,46	0,50	0,75	0,84	0,29	***	0,73	***
400k 085 ω	0,76	#	0,78	*	***	***	***	***	***	0,96	***	0,92
400k 070 ω	0,67	0,90	#	*	***	***	***	***	***	0,83	***	0,86
40k 1 ω	0,84	0,92	0,82	#	***	***	***	***	***	*	***	*
40k 085 ω	0,92	0,69	0,60	0,77	#	0,95	0,29	0,35	0,74	***	0,69	***
40k 070 ω	0,83	0,61	0,52	0,68	0,91	#	0,32	0,38	0,69	***	0,74	***
10k 1 ω	0,83	0,94	0,84	0,98	0,75	0,66	#	0,91	0,17	***	0,51	***
10k 085 ω	0,80	0,96	0,86	0,96	0,72	0,64	0,97	#	0,21	***	0,59	***
10k 070 ω	0,82	0,60	0,51	0,67	0,90	0,99	0,65	0,63	#	***	0,47	***
4k 1 ω	0,77	1,00	0,90	0,92	0,69	0,61	0,94	0,96	0,60	#	***	0,97
4k 085 ω	0,87	0,65	0,56	0,72	0,95	0,96	0,71	0,68	0,94	0,65	#	***
4k 070 ω	0,72	0,95	0,95	0,87	0,64	0,56	0,89	0,91	0,55	0,95	0,60	#

Os elementos abaixo da diagonal são referentes ao teste t e os acima da diagonal ao teste F, *** p < 0,001, ** p < 0,01, * p , 0,05 e . p < 0,1.

Tabela 12 – Descrição, cromossomo e ranking dos Genes presentes na janela de 1Mb
(Continua)

ensembl_gene_id	Chr	Len	R	Descrição do gene	Spos	Epos	Rpos
ENSBTAG00000002002	14	46,05	1	ATPase family AAA domain containing 2	16,473	16,519	-0,522
ENSBTAG00000005877	14	21,41	1	family with sequence similarity 83 member A	16,588	16,609	-0,407
ENSBTAG00000015319	14	17,62	1	chromosome 14 C8orf76 homolog	16,566	16,584	-0,429
ENSBTAG00000015334	14	22,80	1	zinc fingers and homeoboxes 1	16,534	16,557	-0,461
ENSBTAG00000020693	14	23,59	1	derlin 1	16,715	16,738	-0,281
ENSBTAG00000021156	14	66,94	1	TBC1 domain family member 31	16,631	16,698	-0,364
ENSBTAG00000026309	14	20,43	1	zinc fingers and homeoboxes 2	16,772	16,793	-0,223
ENSBTAG00000049059	14	84,21	1		16,869	16,953	-0,127
ENSBTAG00000043985	12	439,01	2	dachshund family transcription factor 1	46,109	46,548	0,212
ENSBTAG00000007383	24	312,26	3	zinc finger protein 521	31,381	31,693	-0,042
ENSBTAG000000048351	24	0,05	3	bta-mir-12051	31,513	31,513	0,091
ENSBTAG000000049616	24	0,07	3	bta-mir-378c	31,436	31,436	0,014
ENSBTAG000000050912	24	4,34	3		31,032	31,036	-0,391
ENSBTAG000000000584	3	5,37	4		0,181	0,186	-0,387
ENSBTAG000000000616	3	54,21	4	G protein-coupled receptor 161	0,507	0,561	-0,061
ENSBTAG00000001052	3	33,95	4	adenylate cyclase 10	0,809	0,843	0,241
ENSBTAG00000002823	3	81,86	4	myelin protein zero like 1	0,915	0,997	0,347
ENSBTAG00000002952	3	26,29	4	T-box transcription factor 19	0,235	0,261	-0,334
ENSBTAG000000005393	3	69,05	4	TOR signaling pathway regulator	0,386	0,455	-0,182
ENSBTAG000000014151	3	71,86	4	RCS domain containing 1	1,011	1,083	0,443
ENSBTAG000000020968	3	30,71	4	mitochondrial pyruvate carrier 2	0,752	0,783	0,184
ENSBTAG000000020969	3	179,67	4	DDB1 and CUL4 associated factor 6	0,573	0,752	0,004
ENSBTAG000000037819	3	19,70	4	SFT2 domain containing 2	0,310	0,330	-0,258
ENSBTAG000000045284	3	0,11	4		0,656	0,656	0,087
ENSBTAG000000013802	3	458,97	5	DAB adaptor protein 1	88,677	89,136	0,032
ENSBTAG000000043985	12	439,01	6	dachshund family transcription factor 1	46,109	46,548	0,191
ENSBTAG000000003708	21	64,69	7	SEC23 homolog A, COPII coat complex component	49,039	49,103	3,121
ENSBTAG000000006130	21	1,89	7	C-type lectin domain containing 14A	48,394	48,396	2,476
ENSBTAG000000006582	21	2,58	7	somatostatin receptor 1	48,348	48,351	2,430
ENSBTAG000000051859	21	4,61	7		49,074	49,079	0,458
ENSBTAG000000004310	21	16,88	8	ubiquitin protein ligase E3 component n-recognin 7	57,831	57,848	0,453
ENSBTAG000000004598	21	28,88	8	golgin A5	57,479	57,508	0,101
ENSBTAG000000006620	21	188,94	8	solute carrier family 24 member 4	57,008	57,197	-0,370
ENSBTAG000000009836	21	13,35	8	chromogranin A	57,572	57,585	0,193
ENSBTAG000000009845	21	163,61	8	inositol-tetrakisphosphate 1-kinase	57,589	57,753	0,211
ENSBTAG000000010416	21	135,89	8	Ras and Rab interactor 3	57,232	57,367	-0,147
ENSBTAG000000031885	21	2,30	8	lysosomal enzyme trafficking factor	57,811	57,813	0,433
ENSBTAG000000042548	21	0,11	8	U6 spliceosomal RNA	57,774	57,774	0,396
ENSBTAG000000044369	21	0,06	8	bta-mir-2284f	57,156	57,156	-0,222
ENSBTAG000000046185	21	89,13	8	BTB domain containing 7	57,858	57,947	0,480
ENSBTAG000000046979	21	35,79	8	legumain	57,379	57,415	0,001
ENSBTAG000000046990	21	7,84	8	GON7 subunit of KEOPS complex	57,823	57,831	0,445
ENSBTAG000000053304	21	0,12	8	5S ribosomal RNA	57,258	57,258	-0,120
ENSBTAG000000002381	15	18,28	9	zinc finger DHHC-type palmitoyltransferase 5	80,913	80,931	-0,349
ENSBTAG000000002389	15	8,59	9	mediator complex subunit 19	80,932	80,941	-0,329
ENSBTAG000000002404	15	8,67	9	thioredoxin related transmembrane protein 2	80,941	80,950	-0,320
ENSBTAG000000002411	15	45,36	9	catenin delta 1	80,967	81,013	-0,294
ENSBTAG000000006493	15	3,10	9	cleavage factor polyribonucleotide kinase subunit 1	80,901	80,904	-0,361
ENSBTAG000000007971	15	0,95	9	olfactory receptor family 6 subfamily Q member 1	81,198	81,199	-0,064
ENSBTAG000000012989	15	14,20	9	ubiquitin conjugating enzyme E2 L6	80,816	80,830	-0,446
ENSBTAG000000016267	15	12,70	9	serpin family G member 1	80,851	80,863	-0,411
ENSBTAG000000021375	15	118,93	9		81,279	81,398	0,018
ENSBTAG000000024701	15	2,60	9	translocase of inner mitochondrial membrane 10	80,788	80,791	-0,473
ENSBTAG000000027524	15	8,73	9	smoothelin like 1	80,805	80,814	-0,456
ENSBTAG000000030122	15	0,09	9	bta-mir-130a	80,888	80,888	-0,373
ENSBTAG000000030908	15	0,94	9	olfactory receptor family 9 subfamily Q member 6	81,372	81,373	0,110
ENSBTAG000000036042	15	0,94	9	olfactory receptor family 5 subfamily B member 148	81,681	81,682	0,419
ENSBTAG000000038267	15	0,98	9	olfactory receptor family 9 subfamily I member 2	81,304	81,305	0,043
ENSBTAG000000039592	15	0,95	9	olfactory receptor family 5 subfamily B member 24	81,615	81,616	0,353
ENSBTAG000000039893	15	21,23	9	olfactory receptor family 9 subfamily I member 17	81,259	81,280	-0,003

Chr – Cromossomo; Len – Comprimento do gene (Mb); R – Ranking do SNP; Spos – Posição inicial do gene (Gb); Epos – Posição final do gene (Gb); Dpos – Distancia do SNP até o Spos (Gb)

Tabela 12 – Descrição, cromossomo e ranking dos Genes presentes na janela de 1Mb
(Continua)

ENSBTAG00000040150	15	0,99	9	olfactory receptor family 1 subfamily S member 7	81,561	81,562	0,299
ENSBTAG00000040197	15	0,94	9	olfactory receptor family 10 subfamily W member 4	81,445	81,446	0,183
ENSBTAG00000047396	15	0,95	9	olfactory receptor family 5 subfamily AZ member 2	81,149	81,150	-0,113
ENSBTAG00000047968	15	6,32	9		80,953	80,959	-0,308
ENSBTAG00000049096	15	4,05	9		81,208	81,212	-0,054
ENSBTAG00000049247	15	0,94	9	olfactory receptor family 5 subfamily B member 140	81,569	81,570	0,307
ENSBTAG00000049391	15	0,99	9	olfactory receptor family 5 subfamily B member 32	81,469	81,470	0,208
ENSBTAG00000049716	15	1,24	9		81,484	81,486	0,223
ENSBTAG00000049719	15	0,94	9	olfactory receptor family 5 subfamily B member 12	81,535	81,536	0,273
ENSBTAG00000050945	15	0,94	9	olfactory receptor family 5 subfamily B member 147, pseudo	81,650	81,650	0,388
ENSBTAG00000051230	15	29,66	9	solute carrier family 43 member 1	80,744	80,773	-0,518
ENSBTAG00000051379	15	0,94	9	olfactory receptor family 5 subfamily B member 12C	81,671	81,672	0,410
ENSBTAG00000051815	15	0,94	9	olfactory receptor family 5 subfamily B member 143	81,745	81,746	0,483
ENSBTAG00000051853	15	0,94	9	olfactory receptor family 5 subfamily B member 141	81,630	81,631	0,368
ENSBTAG00000052154	15	0,95	9	olfactory receptor family 5 subfamily B member 149, pseudo	81,494	81,495	0,233
ENSBTAG00000053250	15	1,01	9	olfactory receptor family 5 subfamily B member 145	81,717	81,718	0,456
ENSBTAG00000053317	15	4,53	9	yippee like 4	80,891	80,896	-0,370
ENSBTAG00000053501	15	12,53	9	olfactory receptor family 5 subfamily B member 2	81,581	81,593	0,319
ENSBTAG00000053887	15	0,89	9	olfactory receptor family 5 subfamily B member 151, pseudo	81,601	81,602	0,339
ENSBTAG00000054501	15	1,03	9	olfactory receptor family 5 subfamily B member 139	81,697	81,699	0,436
ENSBTAG00000055170	15	0,94	9	olfactory receptor family 5 subfamily B member 146	81,514	81,515	0,253
ENSBTAG00000001572	6	45,14	10	NOP2/Sun RNA methyltransferase family member 7	59,664	59,709	-0,203
ENSBTAG00000002356	6	276,36	10	RNA binding motif protein 47	59,387	59,664	-0,480
ENSBTAG00000005078	6	12,26	10	ubiquitin C-terminal hydrolase L1	60,147	60,159	0,280
ENSBTAG00000010677	6	349,19	10	LIM and calponin homology domains 1	60,224	60,573	0,357
ENSBTAG00000027569	6	381,98	10	amyloid beta protein binding family B member 2	59,714	60,096	-0,153
ENSBTAG00000051708	6	22,87	10		59,357	59,380	42,875
ENSBTAG00000002002	14	46,05	11	ATPase family AAA domain containing 2	16,473	16,519	-0,008
ENSBTAG00000005025	14	37,96	11	family with sequence similarity 91 member A1	16,075	16,113	-0,406
ENSBTAG00000005585	14	34,26	11	annexin A13	16,166	16,200	-0,315
ENSBTAG00000005877	14	21,41	11	family with sequence similarity 83 member A	16,588	16,609	0,107
ENSBTAG00000009916	14	18,17	11	N-terminal glutamine amidase 1	16,425	16,443	-0,056
ENSBTAG00000013537	14	130,25	11	fer-1 like family member 6	15,865	15,995	-0,616
ENSBTAG00000015319	14	17,62	11	chromosome 14 C8orf76 homolog	16,566	16,584	0,085
ENSBTAG00000015334	14	22,80	11	zinc fingers and homeoboxes 1	16,534	16,557	0,053
ENSBTAG00000016194	14	32,10	11	F-box protein 32	16,329	16,361	-0,152
ENSBTAG00000016754	14	12,40	11	kelch like family member 38	16,229	16,241	-0,252
ENSBTAG00000020693	14	23,59	11	derlin 1	16,715	16,738	0,233
ENSBTAG00000021156	14	66,94	11	TBC1 domain family member 31	16,631	16,698	0,150
ENSBTAG00000026309	14	20,43	11	zinc fingers and homeoboxes 2	16,772	16,793	0,291
ENSBTAG00000044213	14	0,29	11	7SK RNA	16,288	16,288	-0,194
ENSBTAG00000048822	14	15,88	11	annexin A13	16,143	16,159	-0,338
ENSBTAG00000049059	14	84,21	11		16,869	16,953	0,387
ENSBTAG00000049478	14	8,22	11		16,327	16,335	-0,154
ENSBTAG00000051782	14	25,23	11	ATPase family AAA domain containing 2	16,455	16,480	-0,027
ENSBTAG00000026156	1	87,06	12	vestigial like family member 3	35,036	35,123	0,268
ENSBTAG00000053095	1	0,06	12	bta-mir-2285cm	34,477	34,477	-0,291
ENSBTAG00000006042	28	24,34	13	transmembrane protein 273	43,421	43,446	-0,311
ENSBTAG00000006288	28	90,91	13	V-set and transmembrane domain containing 4	43,286	43,377	-0,446
ENSBTAG00000006719	28	52,12	13	zinc finger AN1-type containing 4	44,161	44,213	0,429
ENSBTAG00000011660	28	10,51	13	microseminoprotein beta	44,061	44,072	0,329
ENSBTAG00000011694	28	21,10	13	translocase of inner mitochondrial membrane 23 homolog (yeast)	44,015	44,036	0,283
ENSBTAG00000011991	28	31,62	13	dorsal root ganglia homeobox	43,579	43,611	-0,153
ENSBTAG00000014990	28	1,57	13	solute carrier family 18 member A3	43,741	43,743	0,009

Chr – Cromossomo; Len – Comprimento do gene (Mb); R – Ranking do SNP; Spos – Posição inicial do gene (Gb); Epos – Posição final do gene (Gb); Dpos – Distancia do SNP até o Spos (Gb)

Tabela 12 – Descrição, cromossomo e ranking dos Genes presentes na janela de 1Mb
(Continuação)

ENSBTAG00000016814	28	44,19	13	choline O-acetyltransferase	43,760	43,805	0,028
ENSBTAG00000018915	28	44,93	13	family with sequence similarity 21, member A	44,113	44,158	0,381
ENSBTAG00000019537	28	111,11	13	membrane associated ring-CH-type finger 8	44,228	44,339	0,496
ENSBTAG00000019565	28	22,33	13	nuclear receptor coactivator 4	44,037	44,059	0,305
ENSBTAG00000021728	28	27,05	13	oxoglutarate dehydrogenase L	43,842	43,869	0,110
ENSBTAG00000023018	28	114,70	13	poly(ADP-ribose) glycohydrolase	43,900	44,015	0,168
ENSBTAG00000023028	28	248,08	13	WDFY family member 4	43,020	43,268	-0,712
ENSBTAG00000032527	28	70,72	13	ERCC excision repair 6, chromatin remodeling factor	43,633	43,704	-0,099
ENSBTAG00000043162	28	0,20	13	Small nucleolar RNA SNORA74	44,028	44,028	0,296
ENSBTAG00000045949	28	9,61	13	chromosome 28 C10orf53 homolog	43,813	43,823	0,081
ENSBTAG00000047155	28	4,30	13	chromosome 28 C10orf71 homolog	43,548	43,552	-0,185
ENSBTAG00000047402	28	4,30	13	family with sequence similarity 170 member B	43,394	43,398	-0,339
ENSBTAG00000050740	28	0,11	13	U6 spliceosomal RNA	43,719	43,719	-0,013
ENSBTAG00000054349	28	18,61	13		43,529	43,548	-0,203
ENSBTAG00000026156	1	87,06	14	vestigial like family member 3	35,036	35,123	0,265
ENSBTAG00000053095	1	0,06	14	bta-mir-2285cm	34,477	34,477	-0,295
ENSBTAG00000026156	1	87,06	15	vestigial like family member 3	35,036	35,123	0,267
ENSBTAG00000053095	1	0,06	15	bta-mir-2285cm	34,477	34,477	-0,292

Chr – Cromossomo; Len – Comprimento do gene (Mb); R – Ranking do SNP; Spos – Posição inicial do gene (Gb); Epos – Posição final do gene (Gb); Dpos – Distancia do SNP até o Spos (Gb)

Tabela 13 – Genes ortólogos

ID	Descrição	BgRatio	pvalue	p.adjust	qvalue
GO:0008033	tRNA processing	45/4490	0,01	0,27	0,25
GO:0051603	proteolysis involved in protein catabolic process	163/4490	0,01	0,27	0,25
GO:0006399	tRNA metabolic process	57/4490	0,01	0,27	0,25
GO:0030163	protein catabolic process	213/4490	0,03	0,27	0,25
GO:0007420	brain development	82/4490	0,03	0,27	0,25
GO:0060322	head development	89/4490	0,03	0,27	0,25
GO:0002792	negative regulation of peptide secretion	10/4490	0,03	0,27	0,25
GO:0002920	regulation of humoral immune response	10/4490	0,03	0,27	0,25
GO:0009953	dorsal/ventral pattern formation	10/4490	0,03	0,27	0,25
GO:0043112	receptor metabolic process	10/4490	0,03	0,27	0,25
GO:0070266	necroptotic process	10/4490	0,03	0,27	0,25
GO:0090278	negative regulation of peptide hormone secretion	10/4490	0,03	0,27	0,25
GO:1903078	positive regulation of protein localization to plasma membrane	10/4490	0,03	0,27	0,25
GO:0002253	activation of immune response	92/4490	0,03	0,27	0,25
GO:0015837	amine transport	11/4490	0,03	0,27	0,25
GO:0030195	negative regulation of blood coagulation	11/4490	0,03	0,27	0,25
GO:0051937	catecholamine transport	11/4490	0,03	0,27	0,25
GO:0051952	regulation of amine transport	11/4490	0,03	0,27	0,25
GO:0015844	monoamine transport	12/4490	0,04	0,27	0,25
GO:0031648	protein destabilization	12/4490	0,04	0,27	0,25
GO:0032527	protein exit from endoplasmic reticulum	12/4490	0,04	0,27	0,25
GO:0046888	negative regulation of hormone secretion	12/4490	0,04	0,27	0,25
GO:1900047	negative regulation of hemostasis	12/4490	0,04	0,27	0,25
GO:1904377	positive regulation of protein localization to cell periphery	12/4490	0,04	0,27	0,25
GO:0008589	regulation of smoothened signaling pathway	13/4490	0,04	0,27	0,25
GO:0030902	hindbrain development	13/4490	0,04	0,27	0,25
GO:0050819	negative regulation of coagulation	13/4490	0,04	0,27	0,25
GO:0061045	negative regulation of wound healing	13/4490	0,04	0,27	0,25
GO:0090114	COPII-coated vesicle budding	13/4490	0,04	0,27	0,25
GO:1903035	negative regulation of response to wounding	13/4490	0,04	0,27	0,25
GO:0010648	negative regulation of cell communication	252/4490	0,04	0,27	0,25
GO:0023057	negative regulation of signaling	253/4490	0,04	0,27	0,25
GO:0002448	mast cell mediated immunity	14/4490	0,04	0,27	0,25
GO:0043303	mast cell degranulation	14/4490	0,04	0,27	0,25
GO:0050829	defense response to Gram-negative bacterium	14/4490	0,04	0,27	0,25
GO:0062208	positive regulation of pattern recognition receptor signaling pathway	14/4490	0,04	0,27	0,25
GO:0097300	programmed necrotic cell death	14/4490	0,04	0,27	0,25
GO:0034470	ncRNA processing	109/4490	0,04	0,27	0,25
GO:0002279	mast cell activation involved in immune response	15/4490	0,05	0,27	0,25
GO:0045576	mast cell activation	16/4490	0,05	0,27	0,25
GO:0050832	defense response to fungus	16/4490	0,05	0,27	0,25