



**PREDIÇÃO E ASSOCIAÇÃO GENÔMICA VIA
HAPLÓTIPOS EM RUMINANTES**

ANDRÉ CAMPÊLO ARAUJO

2022



**UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA**

**PREDIÇÃO E ASSOCIAÇÃO GENÔMICA VIA
HAPLÓTIPOS EM RUMINANTES**

Autor: André Campêlo Araujo
Orientador: Prof. Dr. Paulo Luiz Souza Carneiro

ITAPETINGA
BAHIA – BRASIL
Março de 2022

ANDRÉ CAMPÊLO ARAUJO

**PREDIÇÃO E ASSOCIAÇÃO GENÔMICA VIA HAPLÓTIPOS EM
RUMINANTES**

Tese apresentada, como parte das exigências para obtenção do título de DOUTOR EM ZOOTECNIA, no Programa de Pós-Graduação em Zootecnia da Universidade Estadual do Sudoeste da Bahia.

Orientador: Prof. Dr. Paulo Luiz Souza Carneiro
Co-orientadores: Prof. Dr. Luiz Fernando Brito
Prof. Dr. Carlos Henrique Mendes Malhado

ITAPETINGA
BAHIA – BRASIL
Março de 2022

636.0821 Araujo, André Campêlo.
A687p Predição e associação genômica via haplótipos em ruminantes. / André Campêlo Araujo. – Itapetinga-BA: UESB, 2022.

174f.

Tese apresentada, como parte das exigências para obtenção do título de DOUTOR EM ZOOTECNIA, no Programa de Pós-Graduação em Zootecnia da Universidade Estadual do Sudoeste da Bahia. Sob a orientação do Prof. D. Sc. Paulo Luiz Souza Carneiro e coorientação do Prof. D. Sc. Luiz Fernando Brito e do Prof. D. Sc. Carlos Henrique Mendes Malhado.

1. Ruminantes - Predição e associação genômica - Haplótipos. 2. Ovinos - Sistemas de produção. 3. Bovinos de corte - Sistemas de produção. I. Universidade Estadual do Sudoeste da Bahia - Programa de Pós-Graduação de Doutorado em Zootecnia, *Campus* de Itapetinga. II. Carneiro, Paulo Luiz Souza. III. Brito, Luiz Fernando. IV. Malhado, Carlos Henrique Mendes. V. Título.

CDD(21): 636.0821

Catálogo na Fonte:

Adalice Gustavo da Silva – CRB 535-5ª Região
Bibliotecária – UESB – Campus de Itapetinga-BA

Índice Sistemático para desdobramentos por Assunto:

1. Genômica – Predição e associação – Haplótipos
2. Bovinos de corte - Valor genético
3. Ovinos - Valor genético

UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA – UESB

PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA

Área de Concentração: Produção de Ruminantes

DECLARAÇÃO DE APROVAÇÃO

Título: “Predição e associação genômica via haplótipos em ruminantes”.

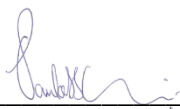
Autor: André Campêlo Araujo

Orientador: Prof. Dr. Paulo Luiz Souza Carneiro

Co-orientadores: Dr. Luiz Fernando Brito

Dr. Carlos Henrique Mendes Malhado

Aprovado como parte das exigências para obtenção do Título de Doutor EM ZOOTECNIA, ÁREA DE CONCENTRAÇÃO: PRODUÇÃO DE RUMINANTES, pela Banca Examinadora:



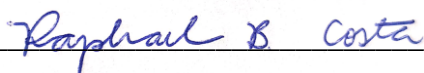
Professor Dr. Paulo Luiz Souza Carneiro



Professor Dr. Victor Breno Pedrosa



Professor Dr. Luis Fernando Batista Pinto



Professor Dr. Raphael Bermal Costa



Professor Dr. José Braccini Neto

Data de realização: 02 de março de 2022

*“Pedi, e dar-se-vos-á; buscai e achareis; batei e abrir-se-vos-á.
Pois tudo o que se pede, recebe; o que busca encontra; e, a quem bate, abrir-se-lhe-á”.*

Jesus Cristo: Mateus 7:7-8

“Insanidade é continuar fazendo a mesma coisa e esperar resultados diferentes”.

Albert Einstein

“Não é que eu seja muito inteligente, eu apenas passo mais tempo com os problemas”.

Albert Einstein

A minha família e amigos.
DEDICO.

AGRADECIMENTOS

A Deus pelo dom da vida.

À Universidade Estadual do Sudoeste da Bahia (UESB), por ter me possibilitado desenvolver este trabalho.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de estudos.

A Purdue University, pelo período de doutorado sanduiche no exterior.

A Universidade Federal de Viçosa (UFV), pelo período de doutorado sanduiche no Brasil.

Ao grupo de estudos em Genômica Aplicada a Conservação e Melhoramento (GACOM) da UESB de Jequié, por ser o primeiro grupo de estudos na área de genética e melhoramento que fiz parte e pelas amizades feitas.

Ao Brito's Lab da Purdue University, por me receber, infraestrutura, todo o suporte técnico e científico, momentos descontração e amizades feitas.

Ao Grupo de Discussão em Melhoramento Animal (GDMA) da UFV, por me receber, conhecimentos compartilhados, amizades feitas e momentos de lazer.

Aos meus orientadores, professores Paulo Carneiro e Luiz Brito, por me proporcionarem não só a possibilidade de obter o título de doutor, mas toda uma mudança de vida, serei sempre grato.

A todos que direta e indiretamente colaboraram para a realização desse trabalho e aos amigos de longa data, dizer o nome de todos necessitaria muitas páginas, mencionar o nome de alguns não seria elegante.

BIOGRAFIA

André Campêlo Araujo, filho de Antonio de Pádua Araujo Páscoa e Maria da Cruz Campêlo Araujo, nasceu em 19 de dezembro de 1991. Em 2010, ingressou na Universidade Federal do Piauí (UFPI), onde, em agosto de 2014, obteve o título de Zootecnista. Em março de 2015, iniciou o Programa de Mestrado em Zootecnia, área de concentração em Produção Animal, da UFPI, obtendo o título de mestre em fevereiro de 2017. Ingressou em março de 2018, no Programa de Doutorado em Zootecnia, área de concentração em Produção de Ruminantes, da Universidade Estadual do Sudoeste da Bahia (UESB). Em 2019, participou de um doutorado saduíche no Brasil na Universidade Federal de Viçosa. Em 2020 e 2021, participou de um doutorado saduíche nos Estados Unidos na Purdue University. Em 02 de março de 2022 submeteu-se à banca de defesa da presente Tese.

SUMÁRIO

	Página
LISTA DE FIGURAS.....	ix
LISTA DE TABELAS.....	xviii
RESUMO.....	xx
ABSTRACT.....	xxiii
I – REFERÊNCIAL TEÓRICO.....	1
1.1 Introdução.....	1
1.2 Predição genômica utilizando haplótipos.....	3
1.3 Associação genômica utilizando haplótipos.....	5
1.4 Sistemas de produção de bovinos de corte.....	7
1.5 Sistemas de produção de ovinos de corte.....	8
1.6 Referências.....	9
II – OBJETIVO GERAL.....	15
2.1 OBJETIVOS ESPECÍFICOS.....	15
III – A Comprehensive Comparison of Haplotype-Based Single-Step Genomic Predictions in Livestock Populations With Different Genetic Diversity Levels: A Simulation Study.....	16
3.1 Introduction.....	17
3.2 Material and Methods.....	19
3.3 Results.....	28
3.4 Discusssion.....	36
3.5 Conclusions.....	45
3.6 Conflict of Interest Statement.....	45
3.7 Data Availability.....	45
3.8 Author Contributions.....	45
3.9 Funding.....	46
3.10 Acknowledgements.....	46
3.11 Contribution to the Field Statement.....	46
3.12 References.....	47

IV – Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle.....	56
4.1 Introduction.....	57
4.2 Material and Methods.....	59
4.3 Results.....	66
4.4 Discussion.....	76
4.5 Conclusions.....	87
4.6 Supplementary Materials.....	87
4.7 Author Contributions.....	88
4.8 Funding.....	88
4.9 Institutional Review Board Statement.....	88
4.10 Informed Consent Statement.....	88
4.11 Data Availability Statement	89
4.12 Acknowledgments.....	89
4.13 Conflicts of Interest.....	89
4.14 References.....	89
V – Haplotype-based single-step genomic predictions for growth, wool, and reproductive traits in North American Rambouillet sheep	96
5.1 Introduction.....	98
5.2 Material and Methods.....	100
5.3 Results.....	109
5.4 Discussion.....	116
5.5 Conclusions.....	123
5.6 Conflict of Interest Statement.....	124
5.7 Data Availability.....	124
5.8 Author Contributions.....	124
5.9 Funding.....	125
5.10 Acknowledgements.....	125
5.11 References.....	125
VI – CONSIDERAÇÕES FINAIS.....	136
VII – ANEXOS.....	137
7.1 Normas da revista <i>Frontiers in Genetics</i>	137
7.2 Normas da revista <i>Genes</i>	148

7.3 Norma da revista Journal of Animal Breeding and Genetics..... 167

LISTA DE FIGURAS

	Página
<p>III – A Comprehensive Comparison of Haplotype-Based Single-Step Genomic Predictions in Livestock Populations With Different Genetic Diversity Levels: A Simulation Study</p>	
<p>Figure 1. Simulation design to obtain pure and composite sheep populations.....</p>	20
<p>Figure 2. Evaluated scenarios used in the genomic predictions with pseudo- single nucleotide polymorphisms (SNPs) from linkage disequilibrium (LD) blocks using independent and pseudo-SNPs in a single genomic relationship matrix (1H), and only pseudo-SNPs and independent and pseudo SNPs in two genomic relationship matrices (2H).....</p>	27
<p>Figure 3. Average number of blocks (Blocks) spanning two or more SNPs, markers within blocks (Blocked_SNPs), pseudo-SNPs (Pseudo_SNPs), pseudo-SNPs after quality control (PS_A_QC), non-blocked SNPs plus pseudo-SNPs after quality control (NB_PS_A_QC), and computing time to obtain the pseudo-SNPs (Duration_time) in the simulation for a trait with moderate heritability ($h^2 = 0.30$). A, B, and C show the results for haplotype blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively. Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds from two and three pure breeds, respectively. The same lower- or upper-case letters mean no statistical difference comparing populations within LD</p>	

thresholds and LD threshold across populations, respectively, at 5% significance level by the Tukey test..... 31

Figure 4. Average number of blocks (Blocks) spanning two or more SNPs, markers within blocks (Blocked_SNPs), pseudo-SNPs (Pseudo_SNPs), pseudo-SNPs after quality control (PS_A_QC), non-blocked SNPs plus pseudo-SNPs after quality control (NB_PS_A_QC), and computing time to obtain the pseudo-SNPs (Duration_time) in the simulation for a trait with low heritability ($h^2 = 0.10$). A, B, and C show the results for the haplotype blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively. Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds from two and three pure breeds, respectively. The same lower- or upper-case letters mean no statistical difference comparing populations within LD thresholds and LD threshold across populations, respectively, at 5% significance level based on the Tukey test..... 32

Figure 5. Accuracies and bias of genomic predictions based on individual SNPs and haplotypes for the simulations of traits with moderate (A) and low (B) heritability (0.30 and 0.10, respectively). Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds from two and three pure breeds, respectively. 600K: high-density panel; 50K: medium-density panel; IPS_LD01, IPS_LD03, and IPS_LD06: independent and pseudo-SNPs from blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively, in a single genomic relationship matrix; PS_LD01, PS_LD03, and PS_LD06: only pseudo-SNPs from blocks with LD threshold of 0.1, 0.3, and 0.6, respectively; and IPS_2H_LD01, IPS_2H_LD03, and IPS_2H_LD06: independent and pseudo-SNPs from blocks with LD thresholds

of 0.1, 0.3, and 0.6, respectively, in two genomic relationship matrices. Zero values for both accuracies and bias mean no results were obtained, due to poor quality of genomic information or no convergence of the genomic prediction models. The same lower-case letters mean no statistical difference comparing genomic prediction methods within population at 5% significance level based on the Tukey test..... 35

IV – Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle

Figure 1. Number of top 0.001% genomic regions for yearling temperament in American Angus cattle found by non-weighted single-step GWAS (ssGWAS) (A) and weighted ssGWAS (WssGWAS) in the second (B) and third (C) iterations. H0.15, H0.50, and H0.80: only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; NCSNP_H0.15, NCSNP_H0.50, and NCSNP_H0.80: non-clustered SNPs (NCSNP) and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively. The column colors highlight not including (blue) or including NCSNP (green). The column filling highlights different LD thresholds (0.15, 0.50, and 0.80 with a solid, square, and diamond filling, respectively)..... 68

Figure 2. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.15 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.15; A) and weighted single-step GWAS in the second (WssGWAS_2_H0.15; B) and third iterations (WssGWAS_3_H0.15; C). Green points highlighted above the red horizontal line are the top 0.001% of markers that explained

greater percentages of the total additive genetic variance for YT.
 The X-chromosome (PAR region) is represented by the
 chromosome 30..... 69

Figure 3. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.50 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.50; A) and weighted single-step GWAS in the second (WssGWAS_2_H0.50; B) and third iterations (WssGWAS_3_H0.50; C). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30..... 71

Figure 4. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.80 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.80; A) and weighted single-step GWAS in the second (WssGWAS_2_H0.80; B) and third iterations (WssGWAS_3_H0.80; C). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30..... 72

Figure 5. Manhattan plot of the variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.15 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.15; A) and weighted single-step GWAS

in the second (WssGWAS_2_NCSNP_H0.15; B) and third iterations WssGWAS_3_NCSNP_H0.15; C). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome is represented by the chromosome 30..... 74

Figure 6. Manhattan plot of the percentage of the total additive genetic variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.50 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.50; A) and weighted single-step GWAS in the second (WssGWAS_2_NCSNP_H0.50; B) and third (WssGWAS_3_NCSNP_H0.50; C) iterations. Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30..... 75

Figure 7. Manhattan plots of the total additive genetic variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.80 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.80; A) and weighted single-step GWAS in the second (WssGWAS_2_NCSNP_H0.80; B) and third (WssGWAS_3_NCSNP_H0.80; C) iterations. Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30..... 76

Figure 8. Venn diagrams showing the number of markers overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively. 77

Figure 9. Venn diagrams showing the number of genes overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD threshold of 0.15, 0.50, and 0.80, respectively..... 78

Figure 10. Venn diagrams showing the number of QTL overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively. 79

Figure 11. Absolute number of quantitative trait loci (QTL) by class overlapping with the top 0.001% markers for yearling temperament in American Angus cattle using the single-step GWAS fitting only haplotypes or non-clustered SNPs and haplotypes. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively. 80

V – Haplotype-based single-step genomic predictions for growth, wool, and reproductive traits in North American Rambouillet sheep

Figure 1. Prediction accuracies for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP fitting pseudo-haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different alpha values (0.95 or 0.50) were used to create the genomic relationship matrices..... 110

Figure 2. Prediction bias for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50).

Different alpha values (0.95 or 0.50) were used to create the genomic relationship matrices..... 112

Figure 3. Dispersion of the GEBVs for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different alpha values (0.95 or 0.50) were used to create the genomic relationship matrices..... 113

Figure 4. Mean theoretical accuracies for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different alpha values (0.95 or 0.50) were used to create the genomic relationship matrices..... 114

Figure 5. Theoretical accuracies for the genomic estimated breeding values using SNPs (TA_GEBV) and estimated breeding values (TA_EBV) per genotyped individuals for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW),

and number of lambs born (NLR). The TA_GEBV and TA_EBV were obtained using SNPs in the single-step GBLUP (H-BLUP) with alpha equal to 0.95 and pedigree-based BLUP (A-BLUP), respectively. The individuals were sorted by data birth, so that the younger individuals are in the right side of each plot..... 115

LISTA DE TABELAS

	Página
<p>III – A Comprehensive Comparison of Haplotype-Based Single-Step Genomic Predictions in Livestock Populations With Different Genetic Diversity Levels: A Simulation Study</p>	
<p>Table 1. Average (SE) effective population size based on the linkage disequilibrium (N_{eLD}) and realized inbreeding (N_{eInb}) methods, additive genetic variance (σ_a^2), residual variance (σ_a^2), and heritability (h^2) estimates of the trait in simulated sheep populations.....</p>	29
<p>IV – Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle</p>	
<p>Table 1. Scenarios used to evaluate the traditional and weighted single-step GWAS (ssGWAS and WssGWAS, respectively) using haplotypes for yearling temperament in American Angus cattle...</p>	64
<p>Table 2. Descriptive statistics of the haplotype blocks with different linkage disequilibrium (LD) thresholds used in each scenario, before and after quality control (QC), in American Angus cattle..</p>	67
<p>Table 3. Gene ontology biological terms (GO_BP) and metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) access of genes overlapped by top 0.001% markers for docility in American Angus cattle.....</p>	79
<p>V – Haplotype-based single-step genomic predictions for growth, wool, and reproductive traits in North American Rambouillet sheep</p>	

Table 1. Description of the datasets used for the genetic and genomic predictions of birth weight (BWT), post-weaning body weight (PWT), yearling body weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep.....	102
Table 2. Variance components and genetic parameters used to predict the estimated breeding values for birth weight (BWT), post-weaning body weight (PWT), yearling body weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep.....	106

RESUMO

ARAÚJO, André Campêlo. **Predição e associação genômica via haplótipos em ruminantes**. Itapetinga, BA: UESB, 2022. 174p. Tese. (Doutorado em Zootecnia, Área de Concentração em Produção de Ruminantes) *.

Os haplótipos são conjuntos de *loci* adjacentes que apresentam maior probabilidade de serem herdados conjuntamente, possuindo algumas vantagens quando comparados aos tradicionais polimorfismos de base única (SNPs, de *Single Nucleotide Polymorphisms*). Dentre essas vantagens podem ser citadas a possibilidade de estarem em maior desequilíbrio de ligação (LD, de *Linkage Disequilibrium*) com os *loci* de característica quantitativa (QTLs, de *Quantitative Trait Loci*) ou mutações causais e capturar efeitos epistáticos. No entanto, os haplótipos tem sido subutilizados quando comparados aos SNPs. Nesse sentido, objetivou-se avaliar o desempenho de predições genômicas de valores genéticos e associações genômicas ampla utilizando haplótipos em bovinos e ovinos em comparação ao uso de SNPs. Um primeiro estudo de simulação foi realizado para avaliar a acurácia, viés e custos computacionais para predição de valores genéticos genômicos (GEBVs, de *Genomic Estimated Breeding Values*) utilizando haplótipos em populações com diversos níveis de diversidade genética. Nesse estudo foram simulados programas de melhoramento de ovinos genotipados com um painel de SNPs de alta densidade (600K). Duas características foram simuladas, uma com moderada (0,30) e outra com baixa (0,10) herdabilidade. O melhor preditor linear não viesado genômico em passo único (ssGBLUP, de *Single-step Genomic Best Linear Unbiased Prediction*) foi utilizado para estimar os GEBVs utilizando: 1) apenas SNPs; 2) SNPs fora dos blocos e haplótipos de blocos com diferentes LD (0,1, 0,3 e 0,6) em uma única matriz de relacionamento genômico (**G**); 3) SNPs fora dos blocos e haplótipos de blocos com diferentes LD (0,1, 0,3 e 0,6) em duas matrizes **G** diferentes; e 4) apenas os haplótipos de blocos com diferente LD (0,1, 0,3 e 0,6). Todas as predições genômicas foram feitas utilizando os softwares da família BLUPf90. O aumento no tempo de análise utilizando os haplótipos foi de no máximo 7 horas quando comparado com o uso de SNPs. A

* Orientador: Dr. Paulo Luiz Souza Carneiro, UESB e Co-orientadores: Dr. Luiz Fernando Brito, Purdue University; Dr. Carlos Henrique Mendes Malhado, UESB.

acurácia e o viés de predição entre SNPs e haplótipos variou de 0,11 até 0,54 e de -0,78 até -0,08 em populações simuladas com alto e baixo tamanho efetivo populacional (N_e), respectivamente. A acurácia e o viés de predição entre os métodos baseados em SNP e haplótipos foram semelhantes nas cinco populações simuladas com N_e diferentes, indicando que ambas as abordagens podem ter desempenho semelhante dependendo da estrutura dos dados e independentemente da herdabilidade. Um segundo estudo teve o objetivo de detectar QTLs e genes candidatos para temperamento ao ano (YT, de *Yearling Temperament*) em bovinos da raça Angus dos Estados Unidos, por meio de estudos de associação genômica ampla (GWAS, de *Genome Wide Association Study*) usando-se haplótipos. Foram utilizados aproximadamente 266 K animais com fenótipos para YT, dos quais aproximadamente 70 K tinham genótipos para um SNP chip de 50 K marcadores. A GWAS em passo único tradicional (ssGWAS, de *Single-step GWAS*) e ponderada (WssGWAS, de *Weighted ssGWAS*) foram investigadas utilizando haplótipos de blocos com LD baixo, médio e alto (0,15, 0,50 e 0,80, respectivamente) incluindo ou não os SNPs fora dos blocos. A WssGWAS não apresentou vantagens comparada com a ssGWAS. A ssGWAS usando haplótipos deve incluir os SNPs fora dos blocos e usar diferentes LD para aumentar a possibilidade de encontrar regiões cromossômicas candidatas para as características de interesse zootécnico. Os genes candidatos para YT foram: *ATXN10*, *ADAM10*, *VAX2*, *ATP6V1B1*, *CRISPLD1*, *CAPRIN1*, *FA2H*, *SPEF2*, *PLXNA1* e *CACNA2D3*; e estão envolvidos em processos biológicos e vias metabólicas importantes e relacionadas a características comportamentais, interações sociais e agressividade em bovinos. No terceiro estudo, predições genômicas utilizando SNPs e haplótipos foram realizadas para características de crescimento, lã e reprodutivas em ovinos da raça Rambouillet dos Estados Unidos. O número de registros fenotípicos variou, aproximadamente, de 5 K a 28 K nas características avaliadas, que foram peso ao nascimento (BWT, de birth weight), peso pós-desmama (PWT, de post-weaning weight), peso ao ano (YWT, de yearling weight), diâmetro da fibra da lã ao ano (YFD, de yearling fiber diameter), peso da lã suja ao ano (YGFWS, de yearling greasy fleece weight) e número de cordeiros nascidos (NLB, de number of lambs born). Um total de 741 animais foram genotipados para um painel de SNPs moderado (50 K, $n = 677$) e alto (600K, $n = 64$), dos quais 32 K SNPs em comum nos dois painéis depois do controle de qualidade foram utilizados para análises posteriores. O ssGBLUP usando SNPs (H-BLUP) ou haplótipos (HAP-BLUP) de blocos com diferentes limites de LD (0,15, 0,35, 0,50, 0,65 e 0,80) foram comparados entre si e com BLUP considerando o pedigree (A-BLUP). O

peso apropriado da informação genômica (parâmetro alfa) também foi estudado. A acurácia teórica média variou de 0,499 (A-BLUP para PWT) a 0,795 (HAP-BLUP usando haplótipos de blocos com limite de LD de 0,35 e alfa igual a 0,95 para YFD). As acurácias de predição variaram de 0,143 (A-BLUP para PWT) a 0,330 (A-BLUP para YGFW), enquanto o viés de predição variou de -0,104 (H-BLUP para PWT) a 0,087 (HAP-BLUP usando haplótipos de blocos com limite de LD de 0,15 e alfa igual a 0,95 para YGFW). A dispersão dos GEBVs variou de 0,428 (A-BLUP para PWT) a 1,035 (A-BLUP para YGDW). O uso de informações genômicas de SNPs ou haplótipos proporcionou acurácias de predição e teóricas semelhantes ou superiores e reduziu a dispersão do GEBVs para características de crescimento, lã e reprodutivas em ovelhas Rambouillet, enquanto o viés de predição não mostrou melhorias claras quando comparado às predições com pedigree.

PALAVRA-CHAVE: blocos de haplótipos, desequilíbrio de ligação, genes candidatos, ovinos e bovinos, valor genético

ABSTRACT

ARAÚJO, André Campêlo. **Genomic prediction and association via haplotypes in ruminants**. Itapetinga, BA: UESB, 2022. 174p. Thesis. (Doctorate in Animal Science, Area of Concentration in Ruminant Production)*.

Haplotypes are sets of adjacent loci that are more likely to be inherited together, having some advantages when compared to traditional Single Nucleotide Polymorphisms (SNPs). Among these advantages, the possibility of being in greater linkage disequilibrium (LD) with the quantitative trait loci (QTLs) or causal mutations and capturing epistatic effects are the main ones. However, haplotypes have been underutilized compared to SNPs. In this context, the objective was to evaluate the performance of genomic predictions of breeding values and genome wide associations using haplotypes in cattle and sheep compared to the use of SNPs. A first simulation study was carried out to evaluate the accuracy, bias and computational costs in predicting the genomic estimated breeding values (GEBVs) using haplotypes in populations with different levels of genetic diversity. In this study, breeding schemes for sheep genotyped with a high-density SNP panel (600 K) were simulated. Two traits were simulated, one with moderate (0.30) and the other with low (0.10) heritability. The Single-step Genomic Best Linear Unbiased Predictor method (ssGBLUP) was used to estimate the GEBVs considering: 1) only SNPs; 2) out-of-block SNPs and haplotypes from blocks with different LD (0.1, 0.3, and 0.6) in a single genomic relationship matrix (**G**); 3) out-of-block SNPs and haplotypes from blocks with different LD (0.1, 0.3, and 0.6) in two different **G** matrices; and 4) only haplotypes from blocks with different LD (0.1, 0.3 and 0.6). All the genomic predictions were done using the BLUPf90 softwares. The analysis time increased up to 7 hours by fitting the haplotypes compared to SNPs. The prediction accuracy and bias between SNPs and haplotypes ranged from 0.11 to 0.54 and from -0.78 to -0.08 in simulated populations with high and low effective population size (N_e), respectively. The genomic prediction accuracy and bias between the SNP and haplotype-based methods were similar across the five populations simulated with different N_e ,

* Orientador: Dr. Paulo Luiz Souza Carneiro, UESB e Co-orientadores: Dr. Luiz Fernando Brito, Purdue University; Dr. Carlos Henrique Mendes Malhado, UESB.

indicating that both approaches may perform similarly depending on the data structure and regardless of heritability. A second study aimed to detect QTLs and candidate genes for yearling temperament (YT) in Angus cattle from the United States through genome-wide association studies (GWAS) using haplotypes. Approximately 266 K animals with phenotypes for YT were used, of which approximately 70 K had genotypes for a SNP chip of 50 K markers. Traditional (ssGWAS, from Single-step GWAS) and weighted (WssGWAS, from Weighted ssGWAS) single-step GWAS were investigated using block haplotypes with low, medium and high LD (0.15, 0.50 and 0.80, respectively) including or not including the SNPs outside the blocks. WssGWAS showed no advantages compared to ssGWAS. The ssGWAS using haplotypes should include the SNPs outside the blocks and use different LD to increase the possibility of finding chromosomal regions associated with YT. Candidate genes for YT are: *ATXN10*, *ADAM10*, *VAX2*, *ATP6V1B1*, *CRISPLD1*, *CAPRINI*, *FA2H*, *SPEF2*, *PLXNA1* and *CACNA2D3*; which are involved in important biological processes and metabolic pathways related to behavioral traits, social interactions and aggressiveness in cattle. In the third study, genomic predictions using SNPs and haplotypes were performed for growth, wool and reproductive traits in Rambouillet sheep from the United States. The number of phenotype records ranged from approximately 5 K to 28K in the evaluated traits, which were birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW) and number of lambs born (NLB). A total of 741 animals were genotyped for a panel of moderate (50K, n=677) and high (600K, n=64) density SNPs, of which 32K SNPs in common in both panels after quality control were used for further analyses. The ssGBLUP using SNPs (H-BLUP) or haplotypes (HAP-BLUP) from blocks with different LD thresholds (0.15, 0.35, 0.50, 0.65, and 0.80) were compared with each other and with BLUP considering the pedigree (A-BLUP). The appropriate weight of genomic information (alpha parameter) was also studied. The average theoretical accuracy ranged from 0.499 (A-BLUP for PWT) to 0.795 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.35 and alpha equal to 0.95 for YFD). The prediction accuracies ranged from 0.143 (A-BLUP to PWT) to 0.330 (A-BLUP to YGFW), while the prediction bias ranged from -0.104 (H-BLUP to PWT) to 0.087 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.15 and alpha equal to 0.95 for YGFW). The dispersion of GEBVs ranged from 0.428 (A-BLUP to PWT) to 1.035 (A-BLUP to YGDW). The use of genomic information from SNPs or haplotypes provided similar or superior prediction and theoretical accuracies and reduced the dispersion of GEBVs for

growth, wool and reproductive traits in Rambouillet sheep, while the prediction bias did not show clear improvements when compared to predictions with pedigree.

KEY-WORDS: haplotype blocks, linkage disequilibrium, candidate genes, sheep and cattle, breeding value

I – REFERENCIAL TEÓRICO

1.1 Introdução

O melhoramento genético clássico tem sido feito com base na obtenção dos valores genéticos dos indivíduos, resultante da solução das equações de modelo misto (Henderson, 1984), e seleção dos reprodutores a partir dos valores genéticos. Essa metodologia é dependente da coleta de uma grande quantidade de observações fenotípicas e informações de pedigree feitas de forma acurada em indivíduos relacionados. Ganhos genéticos relevantes tem sido obtidos com o melhoramento genético clássico, entretanto, o longo tempo para se obter os valores genéticos, as baixas acurácias de predição e os altos custos dos testes de progênie são os principais entraves ao aumento do progresso genético em ruminantes para algumas características (de difícil mensuração, que se manifestam mais tarde na vida dos animais ou com baixa herdabilidade por exemplo) utilizando meios tradicionais (Brito et al., 2020; Mcmanus et al., 2010).

A base do sucesso da produção animal em países desenvolvidos é, além de melhoras nas instalações e no manejo em geral, a presença de programas de melhoramento bem fundamentados. Além do mérito genético dos indivíduos que serão pais das próximas gerações e da herdabilidade das características, o sucesso dos programas de melhoramento depende da acurácia de predição, intensidade de seleção e da velocidade que os ganhos serão passados para progênie. Nesse sentido, a predição genômica (GP, de *Genomic Prediction*) consiste em uma ferramenta interessante para melhorar a eficiência do ganho genético, pois, as informações de densos painéis de polimorfismos de base única (SNPs, de *Single Nucleotide Polymorphisms*) são utilizadas para prever os valores genéticos genômicos dos indivíduos (Meuwissen et al., 2001), proporcionando maiores ganhos genéticos.

As informações de SNP chips com dezenas a centenas de milhares de marcadores distribuídos ao longo de todo o genoma têm se mostrado eficiente para promover ganhos tanto no melhoramento genético animal como no vegetal (Moreira et al., 2020; Lourenco et al., 2020). Dentre as contribuições da GP, destacam-se o aumento da acurácia de predição, a possibilidade de seleção mais precoce e aumento do número de indivíduos

disponíveis para a seleção, aumentando o ganho genético anual (Brito et al., 2017; Meuwissen et al., 2001; Mrode et al., 2018; Rupp et al., 2016).

A GP possibilita contabilizar o parentesco dos indivíduos em nível molecular, ou seja, em nível dos loci de características quantitativas (QTLs, de *Quantitative Trait Loci*) presentes no genoma (VanRaden, 2008). O conhecimento das regiões genômicas que afetam a variação fenotípica é imprescindível para realizar a GP, sendo o objetivo dos estudos de associação genômica ampla (GWAS, de *Genome-wide Association Studies*). Esses estudos permitem elucidar a arquitetura genética das características quantitativas, grande maioria das características de interesse zootécnico, o que pode possibilitar a inclusão dessa informação para fomentar o progresso genético (Schmid & Bennewitz, 2017).

Os SNPs se tornaram os marcadores mais utilizados tanto na GWAS quanto na GP, entretanto, haplótipos também podem ser utilizados nessas abordagens (Araujo et al. 2021, 2022), fato que precisa ser melhor avaliado. Os haplótipos são os alelos de conjuntos de *loci* adjacentes e espera-se que sejam herdados juntos, ou seja, que estão ligados em blocos (blocos de haplótipos ou haploblocos) (Gabriel et al., 2002). O uso de haplótipos na GP e GWAS proporciona algumas vantagens comparado aos SNPs, podendo estar em maior desequilíbrio de ligação (LD, de *Linkage Disequilibrium*) com os QTLs do que os SNPs e capturar efeitos epistáticos dentro dos haploblocos (Hess et al., 2017; Liang et al., 2020; Jiang et al., 2018). No entanto, a maior quantidade de etapas e recursos computacionais nas análises com haplótipos podem ser citadas como desvantagens (Araujo et al., 2021; Cuyabano et al., 2014).

A indústria da ovinocultura tem se transformado no que diz respeito a oferta de produtos, dando mais atenção à produção de carne em relação a lã. O mercado mundial da carne de ovinos em 2020 movimentou aproximadamente 7 bilhões de dólares em exportações, apresentando tendência de crescimento nos últimos 10 anos (FAOSTAT, 2021). No caso da carne bovina, as exportações mundiais movimentaram aproximadamente 39 bilhões de dólares em 2020, também, com tendência de crescimento dos últimos 10 anos (FAOSTAT, 2021). Considerando que a oferta de carne tende a crescer em 14% (5,9 e 15,7% para as carnes bovina e ovina, respectivamente) até 2030, principalmente, devido ao aumento de renda e crescimento populacional (FAO, 2021), o melhoramento genético configura-se como um componente indispensável para alcançar tais ganhos. Nesse sentido, pesquisar novas estratégias que possam otimizar os ganhos genéticos advindos das avaliações genética e seleção são importantíssimas.

1.2 Predição genômica utilizando haplótipos

A GP consiste em utilizar as informações de densos painéis de marcadores SNPs para estimar os valores genético genômicos (GEBVs, de *Genomic Estimated Breeding Values*) dos candidatos a seleção (Meuwissen et al., 2001). No entanto, para a implementação da GP, é necessário ter programas de melhoramento estabelecidos (Rupp et al., 2016) e o conhecimento do LD, que pode ser afetado pela diversidade genética e estrutura populacional dos rebanhos (Goddard, 2009).

O LD pode ser definido como a associação não aleatória entre dois loci, sendo o LD entre marcadores-QTLs a base da GS (Meuwissen et al., 2001). Maiores tamanhos efetivos populacionais (N_e) e distâncias entre os loci estão relacionados a menores LD, sendo necessário maiores quantidades de marcadores para obter melhores acurácias de predição (Brito et al., 2017). Nesse sentido, como os haplótipos podem estar em maior LD com os QTLs do que apenas os SNPs (Calus et al., 2008; Cuyabano et al. 2014, 2015; Hess, et al., 2017; Villumsen et al., 2019), vários trabalhos tem avaliado o uso haplótipos na GP.

O primeiro estudo de GP (Meuwissen et al., 2001) utilizou haplótipos em vez de SNPs como covariáveis no modelo para prever os GEBVs, mostrando o potencial dessa técnica. Subsequentemente, Calus et al. (2008) e Villumsen et al. (2009) mostraram, também, utilizando dados simulados, que os haplótipos resultaram em maiores acurácias e menores viés de predição quando comparado com SNPs. Todavia, esses mesmos autores afirmaram que a superioridade dos haplótipos sobre os SNPs é dependente de fatores relacionados com a forma de criação e o tamanho dos haploblocos.

Em relação a divisão do genoma nos haploblocos, pode-se destacar os métodos que utilizam comprimentos fixos (Villumsen et., 2009) e variáveis (Rinaldo et al., 2005). Nos métodos fixos, o genoma é dividido de acordo com o número de SNPs no haplobloco ou o comprimento em megabases, apresentando a vantagem de serem facilmente obtidos (Hess et al., 2017). No entanto, os haploblocos fixos não consideram o LD ou pontos de recombinação no seu interior (Cuyabano et al., 2014), sendo essas últimas desvantagens desse método. A construção de haploblocos com comprimentos variáveis contabiliza a co-segregação dos alelos através da identificação de *hotspots* de recombinação e considera o LD, no entanto, podem demandar maior tempo de construção e serem específicos de uma determinada população (Hayr, 2016).

Ainda dentre os fatores que podem afetar as análises considerando haplótipos, a densidade do painel de marcadores e o modelo usado também devem ser considerados. Calus et al. (2009) recomendaram o uso de painéis de marcadores mais densos para a GP, pois o uso de haplocos mais densos aumentaria a variância explicada pelos QTLs flanqueados pelos marcadores implicando em melhores resultados da predição. O uso de modelos que primeiro estimam os efeitos dos marcadores para então estimar os valores genéticos, como a maioria dos modelos Bayesianos por exemplo, tem efeito, principalmente, no tempo de análise (Araujo et al., 2021); visto que, os modelos Bayesianos apresentam desempenho semelhante ao BLUP genômico (GBLUP) para características poligênicas (Su et al., 2014), que são maioria entre as características de interesse zootécnico.

O tempo, maior quantidade de etapas e o custo computacional são desvantagens da GP usando haplótipos quando comparado aos SNPs (Cuyabano et al., 2014; Hess et al., 2017). Araujo et al. (2021) utilizando o método single-step GBLUP (ssGBLUP) observaram menores aumentos no tempo de análise para realizar a GP usando haplótipos do que os observados por Cuyabano et al. (2015) usando maiores números de SNPs principais para derivar os haplótipos com modelos BLUP Bayesiano e de mistura.

No GBLUP e ssGBLUP os valores genéticos são obtidos diretamente a partir das equações do modelo misto (MME, de *Mixed Model Equation*), no entanto, assumem o modelo infinitesimal (os marcadores explicam similar e pequena proporção da variância genética da característica) (Wang et al., 2012). A principal diferença entre GBLUP e ssGBLUP consiste em que no primeiro são analisados apenas animais genotipados ou pseudo-fenótipos de avaliações genéticas prévias (valores genéticos estimados desregredidos) devem ser utilizados para considerar informações de outros animais, enquanto no segundo, ambos animais genotipados e não genotipados são considerados diretamente, o que tem apresentado resultados semelhantes ou melhores na GP (Legarra et al., 2014). Uma vantagem dos modelos Bayesianos sobre o GBLUP e ssGBLUP seria a possibilidade de incluir informação prévia a respeito dos efeitos dos marcadores, flexibilizando a pressuposição do modelo infinitesimal (Meuwissen et al., 2001).

A maior flexibilidade dos modelos Bayesianos quanto a arquitetura genética da característica levou à ponderação da matriz genômica de parentesco (\mathbf{G}) com o efeito dos SNPs no GBLUP (Su et al., 2014) e no ssGBLUP (Wang et al., 2012), no intuito de contornar a pressuposição do modelo infinitesimal. De forma geral, a ponderação da matriz \mathbf{G} proporciona melhores acurácias e menores viés do que a não ponderação dessa

matriz tanto no GBLUP (Su et al., 2014; Tiezzi & Maltecca, 2015) como no ssGBLUP (Lourenco et al., 2017; Zhang et al., 2016), sendo semelhante ao uso dos modelos Bayesianos em casos de arquiteturas genéticas que esses modelos seriam indicados. Poucos trabalhos avaliaram o uso haplótipos principalmente no ssGBLUP (Araujo et al., 2021; Feitosa et al., 2019; Teissier et al., 2020).

Em dados reais e utilizando o LD como método de construção dos haploblocos, Cuyabano et al. (2014) observaram que haploblocos com LD de 0,45 promoveram melhores acurácias de predição do que o ajuste de SNPs utilizando modelos de mistura Bayesianos, para proteína do leite, fertilidade e mastite em bovinos da raça Holandês. Ainda em gado Holandês, Cuyabano et al. (2015) ajustaram haplótipos construídos com base apenas em SNPs significativos em uma análise prévia, com o intuito de capturar haplótipos próximos aos QTLs, e observaram melhores acurácias de predição nessa abordagem do que com os SNP para as mesmas características.

Em uma população composta de gado de leite, Hess et al. (2017) usaram vários modelos Bayesianos (BayesA, BayesB e BayesN) para ajustar haplótipos de comprimento fixo considerando diversos limites de frequência mínima para o descarte de alelos dos haploblocos sobre características produtivas. Esses autores observaram melhores acurácias de predição quando foram ajustados haploblocos de 250 kb com descarte de haplótipos com menos de um por cento de frequência.

Deve-se notar que a grande maioria dos resultados para GP usando haplótipos, tanto em dados reais como em dados simulados, foi demonstrado em bovinos. Existem controvérsias em alguns trabalhos mostrando pouco ganho (Hess et al., 2017; Karimi et al., 2018; Mucha et al., 2019) e outros bastante promissores (Liang et al., 2020; Xu et al., 2020). Poucos trabalhos de GP ajustando haplótipos em vez de SNPs foram encontrados na literatura em pequenos ruminantes (Araujo et al., 2021; Teissier et al., 2020), demonstrando a necessidade de avaliar essas metodologias nessas espécies.

1.3 Associação genômica utilizando haplótipos

As vantagens e desvantagens, bem como, as formas de criação dos haploblocos, descritas anteriormente para a GP com haplótipos podem ser estendidas para as análises com GWAS. Entretanto, como os estudos de GWAS tem um foco diferente dos estudos de GP, as implicações do uso de haplótipos nessa abordagem, também, apresentam particularidades.

Apesar de ambas as metodologias apresentarem um objetivo comum, estimar corretamente os efeitos dos alelos, a GP tem o objetivo principal de prever os valores genéticos de forma acurada, enquanto a GWAS visa detectar as posições dos QTLs no genoma (Calus et al., 2009). Nesse sentido, Callus et al. (2009) afirmaram que os modelos de GP tentam maximizar a variância explicadas pelos QTLs presentes no genoma, enquanto que na GWAS ocorre a maximização do contraste entre uma região que tem um QTL e as outras que não tem; fazendo com que algumas vezes o modelo considerado ótimo para uma metodologia pode não ser o mesmo para a outra. Essa discussão implica no fato que, apesar de modelos usando haplótipos não apresentarem melhores acurácias de predição quando comparados aos SNPs, eles podem ser utilizados na investigação de QTLs devido o objetivo ser diferente (Araujo et al. 2021, 2022).

Usar apenas SNPs nos estudos de GWAS pode implicar em menor poder de detecção de QTLs, visto que, as regiões flanqueadas podem proporcionar informações limitadas (Guo et al., 2009; Tang et al., 2009), devido ao menor LD entre SNP-QTL comparado ao LD haplótipo-QTL (Araujo et al., 2022). No entanto, os haplótipos não demonstraram vantagens sobre os SNPs na detecção de QTLs com efeitos pequenos (Lorenz et al., 2010) e podem não capturar os QTLs em regiões genômicas com baixo nível de LD. Nesse sentido, Braz et al. (2019) encontraram dois marcadores associados com a força de cisalhamento da carne em bovinos Nelore que não foram detectados por GWAS usando haplótipos, recomendando também o uso de SNPs em estudos de associação.

Por outro lado, vários estudos tem recomendado o uso de haplótipos na GWAS pelas vantagens, mencionadas anteriormente, que essa abordagem pode oferecer. Ao usar haplótipos na GWAS, Braz et al. (2019) encontraram 4 vezes mais marcadores associados com a força de cisalhamento da carne de Nelore quando comparada ao GWAS usando os SNPs (33 haplótipos e oito SNPs, respectivamente), dos quais apenas dois não foram detectados pelos haplótipos. Aumento semelhante no número de marcadores associados quando se considera haplótipos na GWAS em vez de SNPs foi demonstrado por Bovo et al. (2021) ao analisarem o número de tetos em suínos. No entanto, esses últimos autores também recomendaram o uso de SNPs, pois, algumas regiões genômicas importantes não detectadas pelos haplótipos foram detectadas pelos SNPs.

O uso haplótipos oriundos de janelas sobrepostas com um número fixo de SNPs foi sugerido como mais poderoso na GWAS do que apenas os SNPs ou haplótipos oriundos de haploblocos com base no LD, por serem mais eficientes para regiões com

baixo LD (Guo et al., 2009). No entanto, Araujo et al. (2022) demonstraram que algumas regiões cromossômicas são exclusivas para haploblocos com diferente LD (0.15, 0.50 e 0.80 considerados baixo, médio e alto, respectivamente) e que SNPs fora dos blocos também devem ser incluídos na análise. Ainda segundo Araujo et al. (2022), essas práticas diminuiriam a perda de capacidade de dissecar a arquitetura genética usando haploblocos com base no LD devido alguns SNPs não serem alocados para nenhum bloco, descrita por Li et al. (2007). Nesse sentido, o uso de ambos SNPs e haplótipos tem sido recomendado, uma vez que, aumenta a possibilidade de encontrar QTLs associados com a características quantitativas (Araujo et al., 2022; Bovo et al., 2021; Braz et al, 2019).

1.4 Sistemas de produção de bovinos de corte

A carne bovina é uma commodity, movimentando aproximadamente 42 e 39 bilhões de dólares em importações e exportações, respectivamente, no ano de 2020 com uma tendência de crescimento nos últimos 10 anos (FAOSTAT, 2021). Assim como no caso dos sistemas de produção de ovinos, a cadeia produtiva da carne de boi apresenta diferenças de acordo sistemas de produção em cada país/ região (Greenwood, 2021).

Os maiores produtores de carne bovina são os Estados Unidos, seguido pelo Brasil, produzindo aproximadamente 12 e 10 milhões de toneladas de carne em 2020, respectivamente. Esses países figuraam também no topo do ranking de exportações (FAOSTAT, 2021). A maioria dos sistemas de produção de gado de corte nos Estados Unidos são desenvolvidos de forma extensiva e, basicamente, a pasto durante a primeira fase, com terminação em sistema intensivo utilizando dietas de alta energia (Greenwood, 2021; Vale et al., 2019). As principais raças para produção na América do Norte são o Angus, Red Angus, Hereford, Simental, Charolais, Gelbvieh, Brangus, Limousin, Beefmaster, Shorton e Brahman (Drouillard, 2018). O uso de animais cruzados para abate também é comum nos Estados Unidos, sendo esses cruzados entre animais Angus e alguma outra raça, inclusive de bovinos leiteiros (Greenwood, 2021).

O ciclo da produção de gado de corte nos Estados Unidos dura em média de oito a 12 anos e é afetado, principalmente, pelo preço do gado, período de gestação, tempo que os bezerros levam para atingir o peso de mercado e condições climáticas (USDA, 2021). A maioria dos confinamentos nos Estados Unidos estão concentrados em três estados, sendo eles Nebraska, Kansas e Texas (Drouillard, 2018), devido a maior disponibilidade de grãos com melhor qualidade nutricional (Greenwood, 2021).

O uso de tecnologias de precisão tem sido pesquisado tanto na América do Norte como em outros países no sentido de proporcionar mais ferramentas para auxiliar no sistema de produção, no intuito de aumentar a eficiência, sustentabilidade e rentabilidade (Aquilani, 2022). Dentre as aplicações das tecnologias de precisão, o monitoramento do comportamento dos animais, bem como, a possibilidade de monitorar todo o ciclo dos animais, é promissor devido o comportamento ser uma característica importante para o sistema de produção (Brito et al., 2020), com grande impacto econômico e de bem-estar (Northcut & Bowman, 2010).

Apesar de não ser o maior produtor de carne bovina do mundo, o Brasil é o maior exportador, exportando aproximadamente 1,7 milhões de toneladas de carne (*in natura* e processados), apresentando crescimento no número de exportações desde 2008 (FAOSTAT, 2021). O Brasil e os Estados Unidos compartilham o fato de a maior parte do sistema de produção ser desenvolvido a pasto (Greenwood, 2021), no entanto, a maioria do gado brasileiro é terminada a pasto embora o número de confinamentos esteja em crescimento (Vale et al., 2019). Outra diferença evidente entre o sistema de produção nesses dois países é a composição dos animais. O gado de corte brasileiro é predominantemente *Bos taurus indicus*, com a raça Nelore como o principal representante, sendo que, os animais zebuínos correspondem a aproximadamente 80% do rebanho nacional (ABCZ, 2020; Santana Júnior et al., 2016).

O maior uso de animais zebuínos nos sistemas de produção brasileiros ocorre devido a esses bovinos serem mais adaptados as condições tropicais (altas temperaturas, resistência e baixa qualidade de pasto) predominantes no Brasil (Santana Júnior et al., 2016). Não obstante, a qualidade da carne dos animais zebuínos, mais especificamente, o marmoreio e composição da gordura e maciez, é menor do que a de animais taurinos, o que tem estimulado estratégias de manejo e melhoramento genético para melhorar estas características (Braz et al., 2019; Feitosa et al., 2019), além do uso de cruzamentos com raças taurinas.

1.5 Sistemas de produção de ovinos de corte

A cadeia produtiva da ovinocultura varia de acordo com o país de origem, sendo possível caracterizar alguns tipos de sistemas de produção no mundo. Em sua grande maioria, as criações são realizadas a pasto, com baixos investimentos e rebanhos pouco numerosos. No entanto, existem países como Nova Zelândia e Austrália, que são

referências nessa atividade, tendo como base o manejo mais tecnificado e programas de melhoramento bem estabelecidos (Rupp et al., 2016; Mrode et al., 2018).

A Nova Zelândia e Austrália são responsáveis por mais de dois terços das exportações de carne de ovinos (FAOSTAT, 2021), representando maior a parte do mercado internacional, atingindo os mais diversos mercados. No caso da Nova Zelândia, foram registrados ganhos na ordem 83% em quilos de cordeiros produzidos por ovelha e mais de 28% no peso da carcaça de 1990 a 2012 (Beef and Lamb New Zealand, 2012). A Austrália é o maior produtor de lã do mundo (FAOSTAT, 2021), mas a produção da carne de cordeiro tem ganhado espaço nesse país por ser um produto que vem ganhando mercado nos últimos anos em comparação com a lã (Ferguson et al., 2014). Tanto na Nova Zelândia quanto na Austrália, o sucesso da atividade está relacionado aos investimentos em pesquisa e disseminação das tecnologias nas áreas de manejo, instalações, melhoramento genético e uso de raças especializadas para sistemas de produção específicos (Beef and Lamb New Zealand, 2012; MLA, 2016).

Apesar da Nova Zelândia e Austrália serem tradicionais na produção de ovinos, outros países estão emergindo neste setor. Nesse sentido, existem alguns países em desenvolvimento que tem implementado programas de melhoramento em ovinos, como Etiópia e Índia, tendo como base parcerias com instituições de pesquisa, investindo, principalmente, em programas de melhoramento comunitário (Haile et al., 2014). Os programas de melhoramento comunitários são alternativas interessantes para pequenos produtores, pois, terão acesso a animais melhorados que podem aumentar a produtividade (Mrode et al., 2018).

No Brasil, os programas de melhoramento de ovinos são escassos. A criação de pequenos ruminantes ocorre em sua grande maioria, em pequenas propriedades, com pouco investimento tecnológico e quase exclusivamente a campo, fazendo com que os produtores tenham pouco retorno econômico (Lôbo et al., 2010). Esse cenário proporciona vulnerabilidade pela dependência direta do clima e oferta irregular de alimento, que aliada a falta de programas de melhoramento colocam outros países na América Latina, como Uruguai e Argentina, com maior representatividade na produção de cordeiros (Morris et al., 2017).

1.6 Referências

ABCZ. **Associação Brasileira de Criadores de Zebu**. 2020. Disponível em <<https://www.abcz.org.br/a-abcz/racas-zebuinas>> Acesso em 20 de fevereiro de 2021.

AQUILANI, C.; CONFESSORE, A.; BOZZI, R.; SIRTORI, F.; PUGLIESE, C. Review: Precision Livestock Farming technologies in pasture-based livestock systems. **Animal**, v.16, p.100429, 2022.

ARAUJO, A.C.; CARNEIRO, P.L.S.; OLIVEIRA, H.R.; SCHENKEL, F.S.; VERONEZE, R.; LOURENCO, D.A.L.; BRITO, L.F. A comprehensive comparison of haplotype-based single-step genomic predictions in livestock populations with different genetic diversity levels: a simulation study. **Frontiers in Genetics**, v.12, p.729867, 2021.

ARAUJO, A.C.; CARNEIRO, P.L.S.; ALVARENGA, A.B.; OLIVEIRA, H.R.; MILLER, S.P.; RETALLICK, K.; BRITO, L.F. Haplotype-based single-step gwas for yearling temperament in american angus cattle. **Genes**, v.13, p.17, 2022.

Beef and Lamb New Zealand 2012. **Domestic trends and measuring progress against the Red Meat Sector Strategy**. Presentation to the Red Meat Sector Conference, 16 July 2012, Queestown. Disponível em <http://www.mia.co.nz/docs/mia_conference/2012/Rob%20Davidson.pdf> Acesso em 5 de agosto de 2018.

BOVO, S.; BALLAN, M.; SCHIAVO, G.; RIBANI, A.; TINARELLI, S.; UTZERI, V.J.; DALL'OLIO, S.; GALLO, M.; FONTANESI, L. Single-marker and haplotype-based genome-wide association studies for the number of teats in two heavy pig breeds. **Animal Genetics**, v.52, p.440–450, 2021.

BRAZ, C.U.; TAYLOR, J.F.; BRESOLIN, T.; ESPIGOLAN, R.; FEITOSA, F.L.B.; CARVALHEIRO, R.; BALDI, F.; ALBUQUERQUE, L.G.; OLIVEIRA, H.N. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. **BMC Genetics**, v.20, p.1–12, 2019.

BRITO, L.F.; CLARKE, S.M.; MCEWAN, J.C.; MILLER, S.P.; PICKERING, N.K.; BAIN, W.E.; DODDS, K.G.; SARGOLZAEI, M.; SCHENKEL, F.S. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. **BMC Genetics**, v.18, p.7-24, 2017.

BRITO L.F.; OLIVEIRA, H.R.; MCCONN, B.R.; SCHINCKEL, A.P.; ARRAZOLA, A.; MARCHANT-FORDE, J.N. JOHNSON J.S. Large-Scale Phenotyping of Livestock Welfare in Commercial Production Systems: A New Frontier in Animal Breeding. **Frontiers in Genetics**, v. 11, p. 793, 2020.

CALUS, M.P.L.; MEUWISSEN, T.H.E.; DE ROOS, A.P.W.; VEERKAMP, R.F. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. **Genetics**, v.178, p.553-561, 2008.

CALUS, M.P.; MEUWISSEN, T.H.; WINDIG, J.J.; KNOL, E.F.; SCHROOTEN, C.; VEREIJKEN, A.L.; VEERKAMP, R.F. Effects of the number of markers per haplotype

and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. **Genetics Selection Evolution**, v.41, p.11-21, 2009.

CUYABANO, B.C.D.; SU, G.S.; LUND, M.S. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. **BMC Genomics**, v.15, p.1171, 2014.

CUYABANO, B.C.D.; SU, G.S.; LUND, M.S. Selection of haplotype variables from a high-density marker map for genomic prediction. **Genetics Selection Evolution**, v.47, p.61, 2015.

DROUILLARD, J.S. Current situation and future trends for beef production in the United States of America. **Asian-Australasian Journal of Animal Sciences**, v.31, p.1007–1016, 2018.

FAOSTAT. **FAO (Food and Agricultural Organisation) Statistics**. 2021. Disponível em <<http://www.fao.org/3/a-I5703E.pdf>> Acesso em 15 de janeiro de 2021.

FEITOSA, F.L.B.; PEREIRA, A.S.C.; AMORIM, S.T.; PERIPOLLI, E.; SILVA, R.M.O.; BRAZ, C.U.; FERRINHO, A.M.; SCHENKEL, F.S.; BRITO, L.F.; ESPIGOLAN, R.; ALBUQUERQUE, L.G.; BALDI, F. Comparison between haplotype-based and individual SNP-based genomic predictions for beef fatty acid profile in Nellore cattle. **Journal of Animal Breeding Genetics**, v.137, p.468-476, 2019.

FERGUSON, D.M.; SCHREURS, N.M.; KENYON, P.R.; JACOB, R.H. Balancing consumer and societal requirements for sheep meat production: an Australian perspective. **Meat Science**, v.98, n.3, p.477–483, 2014.

GABRIEL, S.B.; SCHAFFNER, S.F.; NGUYEN, H.; MOORE, J.M.; ROY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU-CORDERO, S.C.; ROTIMI, C.; ADEYEMO, A.; COOPER, R.; WARD, W.; LANDER, E.S.; DALY, M.J.; ALTSHULER, D. The Structure of Haplotype Blocks in the Human Genome. **Science**, v.296, p.2225-2229, 2002.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetics**, v.136, p. 245–57, 2009

GREENWOOD, P.L. Review: An overview of beef production from pasture and feedlot globally, as demand for beef and the need for sustainable practices increase. **Animal**, v.15, p.100295, 2021.

GUO, Y.; LI, J.; BONHAM, A.J.; WANG, Y.; DENG, H. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: A comparison of association-mapping strategies. **European Journal of Human Genetics**, v.17, p.785–792, 2009.

HAILE, A.; DESSIE, T.; RISCHKOWSKY, B. **Performance of indigenous sheep breeds managed under community based breeding programs in the high lands of Ethiopia: preliminary results**. ICARDA, Addis Ababa, 2014.

- HESS, M.; DRUET, T.; HESS, A.; GARRICK, D. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. **Genetics Selection Evolution**, v.49, p. 54, 2017.
- JIANG, Y.; SCHMIDT, R.H.; REIF, J.C. Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. **Genes Genomes Genetics**, v. 8, p.1687-1699, 2018.
- KARIMI, Z.; SARGOLZAEI, M.; ROBINSON, J.A.B.; SCHENKEL, F.S. Assessing haplotype-based models for genomic evaluation in Holstein cattle. **Canadian Journal Animal Science**, v.98, p.750-760, 2018.
- LEGARRA, A.; CHRISTENSEN, O.F.; AGUILAR, I.; MISZTAL, I. Single Step, a general approach for genomic selection. **Livestock Science**, v.166, p.54-65, 2014.
- LI, Y.; SUNG, W.K.; LIU, J.J. Association mapping via regularized regression analysis of single-nucleotide polymorphism haplotypes in variable-sized sliding windows. **American Journal of Human Genetics**, v. 80, p.705–715, 2007.
- LIANG, Z.; TAN, C.; PRAKAPENKA, D.; MA, L.; DA, Y. Haplotype Analysis of Genomic Prediction Using Structural and Functional Genomic Information for Seven Human Phenotypes. **Frontiers in Genetics**, v.11, p.1, 2020.
- LÔBO, R. N. B.; FACÓ, O.; LÔBO, A. M. B. O.; VASQUES VILLELA, L. C. Brazilian goat breeding programs. **Small Ruminant Research**, v. 89, p. 149–154, 2010.
- LORENZ, A.J.; HAMBLIN, M.T.; JANNINK, J-L. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. **PLoS One**, v.5, p.e14079, 2010.
- LOURENCO, D.A.L.; FRAGOMENI, B.O.; BRADFORD, H.L.; MENEZES, I.R.; FERRAZ, J.B.S.; AGUILAR, I.; TSURUTA, S.; MISZTAL, I. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. **Journal of Animal Breeding and Genetics**, v.134, p.463–471, 2017.
- LOURENCO, D.; LEGARRA, A.; TSURUTA, S.; MASUDA, Y.; AGUILAR, I.; MISZTAL, I. Single-step genomic evaluations from theory to practice: Using SNP chips and sequence data in BLUPF90. **Genes**, v. 11, p.790, 2020.
- MCMANUS, C.; PAIVA, S. R.; ARAÚJO, R. O. Genetics and breeding of sheep in Brazil, **Revista Brasileira de Zootecnia**, v.39, p. 236-246, 2010.
- MEUWISSEN, T.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819-1829, 2001.
- MOREIRA, F.F.; OLIVEIRA, H.R.; VOLENEC, J.J.; RAINEY, K.M.; BRITO, L.F. Integrating high-throughput phenotyping and statistical genomic methods to genetically improve longitudinal traits in crops. **Frontiers in Genetics**, v.11, p.681, 2020.
- MORRIS, S. T. Overview of sheep production systems. Advances in Sheep Welfare. In: FERGUSON, D. M.; LEE, C.; FISHER, A. (Ed.), **Advances in Sheep Welfare**. Amsterdã: Elsevier, 2017. p. 19-35.

MRODE, R.; TAREKEGN, G.M.; MWACHARO, J.M.; DJIKENG, A. Invited review: Genomic selection for small ruminants in developed countries: how applicable for the rest of the world? **Animal**, v.12, p.1333-1340, 2018.

MUCHA, A.; WIERZBICKI, H.; KAMIŃSKI, S.; OLEŃSKI, K.; HERING, D. Highfrequency Marker Haplotypes in the Genomic Selection of Dairy Cattle. **Journal of Applied Genetics**, v.60, p.179–186, 2019.

NORTHCUT, S.; BOWMAN, B. By the Numbers: Docility Genetic Evaluation Research. Disponível em <<http://www.angus.org/nce/documents/bythenumbersdocility.pdf>> Acessado em 12 agosto de 2021.

RINALDO, A.; BACANU, S.A.; DEVLIN, B.; SONPAR, V.; WASSERMAN, L.; ROEDER, K. Characterization of multilocus linkage disequilibrium. **Genetic Epidemiology**, v.28, p.193-206, 2005.

RUPP, R.; MUCHA, S.; LARROQUE, H.; MCEWAN, J.; CONINGTON, J. Genomic application in sheep and goat breeding. **Animal Frontiers**, v.6, p.39–44, 2016.

SANTANA JÚNIOR, M.L.; PEREIRA, R.J.; BIGNARDI, A.B.; AYRES, D.R.; MENEZES, G.R.O.; SILVA, L.O.C.; LEROY, G.; MACHADO, C.H.C.; JOSAHKIAN, L.A.; ALBUQUERQUE, L.G. Structure and genetic diversity of Brazilian Zebu cattle breeds assessed by pedigree analysis. **Livestock Science**, v.187, p.6–15, 2016.

SCHMID, M.; BENNEWITZ, J. Invited review: Genome-wide association analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs. **Archives Animal Breeding**, v. 60, p. 335–346, 2017.

SU, G.; CHRISTENSEN, O.F.; JANSS, L.; LUND, M.S. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. **Journal of Dairy Science**, v.97, p.6547–6559, 2014.

TANG, R.; FENG, T.; SHA, Q.; ZHANG, S. A variable-sized sliding-window approach for genetic association studies via principal component analysis. **Annals of Human Genetics**, v. 73, p.631–637, 2009.

TEISSIER, M.; LARROQUE, H.; BRITO, L.F.; RUPP, R.; SCHENKEL, F.S.; ROBERT-GRANIÉ, C. Genomic predictions based on haplotypes fitted as pseudo-SNPs for milk production and udder type traits and somatic cell score in French dairy goats. **Journal Dairy Science**, v.103, p.11559-11573, 2020.

TIEZZI, F.; MALTECCA, C. Genomic prediction using a weighted relationship matrix to account for trait architecture in US Holstein cattle. **Genetics Selection Evolution**, v. 47, p. 24-37, 2015.

USDA, 2021. **Cattle & beef. Sector at a glance**. Disponível em <<https://www.ers.usda.gov/topics/animal-products/cattle-beef/sector-at-a-glance/>> Acessado em 22 de janeiro de 2022.

VALE, P.; GIBBS, H.; VALE, R.; CHRISTIE, M.; FLORENCE, E.; MUNGER, J.; SABAINI, D. The expansion of intensive beef farming to the Brazilian Amazon. **Global Environmental Change**, v.57, p.101922, 2019

VANRADEN, P. Efficient methods to compute genomic predictions. **Journal Dairy Science**, v.91, p.4414-4423, 2008.

VILLUMSEN, T.M.; JANSS, L.; LUND, M. The importance of haplotype length and heritability using genomic selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v.126, p.3-13, 2009.

WANG, H.; MISZTAL, I.; AGUILAR, I.; LEGARRA, A.; MUIR, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, v. 94, p. 73-83, 2012.

Xu, L.; Gao, N.; Wang, Z.; Xu, L.; Liu, Y.; Chen, Y.; Xu, L.; Gao, X., Zhang, L.; Gao, H.; Zhu, B.; Li, J. Incorporating Genome Annotation into Genomic Prediction for Carcass Traits in Chinese Simmental Beef Cattle. **Frontiers in Genetics**, v.11, p.481, 2020.

ZHANG, X.; LOURENCO, D.; AGUILAR, I.; LEGARRA, A.; MISZTAL, I. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. **Frontiers in Genetics**, v. 7, p. 151-165, 2016.

II – OBJETIVO GERAL

Objetivou-se avaliar o desempenho de predições e associações genômicas utilizando haplótipos em ruminantes de interesse zootécnico em comparação ao uso de SNPs.

2.1 Objetivos específicos

- ✓ Avaliar o tempo de análise com o uso haplótipos e SNPs;
- ✓ Comparar a acurácia de predição dos GEBVs calculados a partir de SNPs e haplótipos utilizando o método ssGBLUP;
- ✓ Comparar viés de predição dos GEBVs calculados a partir de SNPs e haplótipos;
- ✓ Comparar a dispersão dos GEBVs calculados a partir de SNPs e haplótipos;
- ✓ Avaliar os métodos ssGWAS e WssGWAS com o uso de haplótipos;
- ✓ Avaliar diferentes formas de construir os haplótipos e inclui-los nos modelos;
- ✓ Avaliar a influência da diversidade genética das populações nas predições genômicas.

III – CAPÍTULO I

Artigo publicado na revista *Frontiers in Genetics*

Doi: <https://doi.org/10.3389/fgene.2021.729867>

A Comprehensive Comparison of Haplotype-Based Single-Step Genomic Predictions in Livestock Populations With Different Genetic Diversity Levels: A Simulation Study

Andre C. Araujo^{1,2}, Paulo L. S. Carneiro³, Hinayah R. Oliveira^{2,4}, Flavio S. Schenkel⁴, Renata Veroneze⁵, Daniela A. L. Lourenco⁶, Luiz F. Brito^{2*}

¹Postgraduate Program in Animal Sciences, State University of Southwestern Bahia, Itapetinga, BA, Brazil

²Department of Animal Sciences, Purdue University, West Lafayette, IN, USA

³Department of Biology, State University of Southwestern Bahia, Jequié, BA, Brazil

⁴Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada

⁵Department of Animal Sciences, Federal University of Viçosa, Viçosa, MG, Brazil

⁶Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA

***Corresponding author:** Luiz F. Brito

Address: Department of Animal Sciences, Purdue University, West Lafayette, 47907, IN, USA

E-mail: britol@purdue.edu

Phone number: +1 765 586 2515

Keywords: effective population size, genomic estimated breeding value, haplotype blocks, linkage disequilibrium, pseudo-SNP

Abstract

The level of genetic diversity in a population is inversely proportional to the linkage disequilibrium (LD) between individual single nucleotide polymorphisms (SNPs) and quantitative trait loci (QTLs), leading to lower predictive ability of genomic breeding values (GEBVs) in high genetically diverse populations. Haplotype-based predictions could outperform individual SNP predictions by better capturing the LD between SNP and QTL. Therefore, we aimed to evaluate the accuracy and bias of individual-SNP- and haplotype-based genomic predictions under the single-step-genomic best linear unbiased prediction (ssGBLUP) approach in genetically diverse populations. We simulated purebred and composite sheep populations using literature parameters for moderate and low heritability traits. The haplotypes were created based on LD thresholds of 0.1, 0.3, and 0.6. Pseudo-SNPs from unique haplotype alleles were used to create the genomic relationship matrix (\mathbf{G}) in the ssGBLUP analyses. Alternative scenarios were compared in which the pseudo-SNPs were combined with non-LD clustered SNPs, only pseudo-SNPs, or haplotypes fitted in a second \mathbf{G} (two relationship matrices). The GEBV accuracies for the moderate heritability-trait scenarios fitting individual SNPs ranged from 0.41 to 0.55 and with haplotypes from 0.17 to 0.54 in the most ($N_e \cong 450$) and less ($N_e < 200$) genetically diverse populations, respectively, and the bias fitting individual SNPs or haplotypes ranged between -0.14 and -0.08 and from -0.62 to -0.08, respectively. For the low heritability-trait scenarios, the GEBV accuracies fitting individual SNPs ranged from 0.24 to 0.32, and for fitting haplotypes, it ranged from 0.11 to 0.32 in the more ($N_e \cong 250$) and less ($N_e \cong 100$) genetically diverse populations, respectively, and the bias ranged between -0.36 and -0.32 and from -0.78 to -0.33 fitting individual SNPs or haplotypes, respectively. The lowest accuracies and largest biases were observed fitting only pseudo-SNPs from blocks constructed with an LD threshold of 0.3 ($P < 0.05$), whereas the best results were obtained using only SNPs or the combination of independent SNPs and pseudo-SNPs in one or two \mathbf{G} matrices, in both heritability levels and all populations regardless of the level of genetic diversity. In summary, haplotype-based models did not improve the performance of genomic predictions in genetically diverse populations.

1 Introduction

Genomic selection (GS) (Meuwissen et al., 2001) is now routinely used worldwide in livestock and plant breeding programs (Lourenco et al., 2020; Moreira et al., 2020). GS enables the prediction of more accurate genomic estimated breeding values (GEBVs) at earlier stages compared

to the traditional pedigree-based evaluation (Guarini et al., 2018, 2019; Brito et al., 2017a). The advantages of GS compared to the pedigree-based are even greater for lowly-heritable traits, traits measured late in life, and sex-limited or expensive-to-measure traits (Daetwyler et al., 2012; Lourenco et al., 2020).

Over the past 15–20 years, several statistical methods have been proposed aiming to obtain more accurate and less biased GEBVs. Among the available methods, the single-step genomic best linear unbiased prediction (ssGBLUP; Legarra et al., 2009; Aguilar et al., 2010) is widely used to perform genomic predictions in livestock. This method enables the simultaneous evaluation of both genotyped and non-genotyped individuals and has similar or better statistical properties and predictive ability compared to other approaches such as pedigree-based BLUP and multi-step GBLUP (Legarra et al., 2014; Aguilar et al., 2010; Guarini et al., 2018; Piccoli et al., 2020).

Although the pioneer GS study (i.e., Meuwissen et al., 2001) fitted single nucleotide polymorphism (SNP) haplotypes as covariates in the models, subsequent studies were mainly performed based on individual SNPs. This is most likely due to the additional analytic steps and higher computational requirements when fitting haplotype-based models. In this sense, it is important to first define the haplotype blocks or haploblocks, which are sizable regions of the genome with little evidence of historical recombination (Gabriel et al., 2002), i.e., a genomic region between two or more marker loci. More recently, the use of haplotypes as covariates in genomic evaluations rather than single SNPs has been further investigated due to many potential advantages. Haplotypes are more polymorphic than individual SNPs because they can be multi-allelic (Meuwissen et al., 2014) and they can be in stronger linkage disequilibrium (LD) with Quantitative Trait Loci (QTLs) compared to individual SNPs with low minor allele frequency (MAF) (Hess et al., 2017). In this context, the potential stronger LD between haplotypes and QTL in comparison to individual SNPs can yield more accurate GEBVs (Calus et al., 2008; Cuyabano et al., 2014; 2015). Moreover, haplotype alleles have the potential to capture epistatic effects within blocks and the QTL can be flanked by SNPs that delimit the haploblock (Hess et al., 2017; Jiang et al., 2018; Karimi et al., 2018).

Previous studies based on simulated data have shown that fitting haplotypes can substantially improve the performance of genomic predictions compared to individual SNP-based methods (Calus et al., 2008; Villumsen et al., 2009). However, none or only small increases in the predictive ability of GEBVs have been observed in practice (e.g., Cuyabano et al., 2014; 2015; Hess et al., 2017; Karimi et al., 2018; Mucha et al., 2019; Won et al., 2020). The large majority of the studies evaluating haplotype-based models were done in dairy cattle populations (real or simulated

datasets), which usually have high LD levels between SNP markers and lower genetic diversity (N_e lower than 100; Makanjuola et al., 2020). Haplotype-based genomic predictions in populations with increased genetic diversity, on the other hand, have not been widely explored yet, and the knowledge of their possible advantages is limited (Feitosa et al., 2019; Teissier et al., 2020).

Different from intensively selected populations and pure breeds, which present low genetic diversity (e.g., Holstein dairy cattle), genetically diverse populations (e.g., relatively recent breeding programs in small ruminants and crossbred or composite populations) may have more alleles segregating in the haplotype blocks and greater complexity in the interactions among haplotype allele effects within haploblocks. Thus, we hypothesize that haplotype-based methods could result in more accurate and less biased GEBV prediction when compared to SNP-based models in populations with high genetic diversity because of their development process (e.g., relatively lower selection pressures, crossbreeding) and more complex haplotype structure than observed in populations with low genetic diversity. Simulated data is an interesting approach to investigate this hypothesis because the true breeding values (TBVs) are known (Morris et al., 2019; Oliveira et al., 2019). Therefore, we simulated sheep populations with different genetic diversity levels to test our hypothesis. Sheep is a good model due to the large genetic diversity in commercial populations, with N_e ranging from less than 50 to over 1,000 (Kijas et al., 2012; Brito et al., 2017b; Stachowicz et al., 2018). Hence, the main objective of this study was to evaluate the accuracy and bias of GEBVs in genetically diverse populations, using ssGBLUP when: 1) only individual SNPs are used to construct a single genomic relationship matrix (\mathbf{G}); 2) non-clustered (out of haploblocks) SNPs and haplotypes (fitted as pseudo-SNPs) are used to construct a single \mathbf{G} ; 3) only haplotypes are used to construct a single \mathbf{G} ; and 4) non-clustered SNPs and haplotypes are used to construct two \mathbf{G} matrices. We also compared the impact of different SNP panel densities and haploblock-building methods on the performance of genomic prediction, as these factors could impact the accuracies and bias of genomic predictions.

2 Materials and Methods

The approval of Institutional Animal Care and Use Committee was not required because this study only used computationally simulated datasets.

2.1 Data simulation

2.1.1 Population structure. The simulation was performed to mimic datasets of purebred and composite sheep populations (Kijas et al., 2012; Prieur et al., 2017; Brito et al., 2017a; Oliveira et al., 2020). The QMSim software (Sargolzaei and Schenkel, 2009) was used to simulate a

historical population initially with 80,000 individuals (40,000 males and 40,000 females). Then, a population bottleneck was simulated, reaching 50,000 individuals (25,000 males and 25,000 females) in the 1,000th generation. After that, there was an increase in the population to 60,000 individuals, with 20,000 males and 40,000 females in the 1,500th generation. There was random mating in the historical population, with gametes randomly sampled from the pool of males and females present in each generation. Mutation and genetic drift were considered in the historical population to create the initial LD. The complete simulation design is summarized in Figure 1.

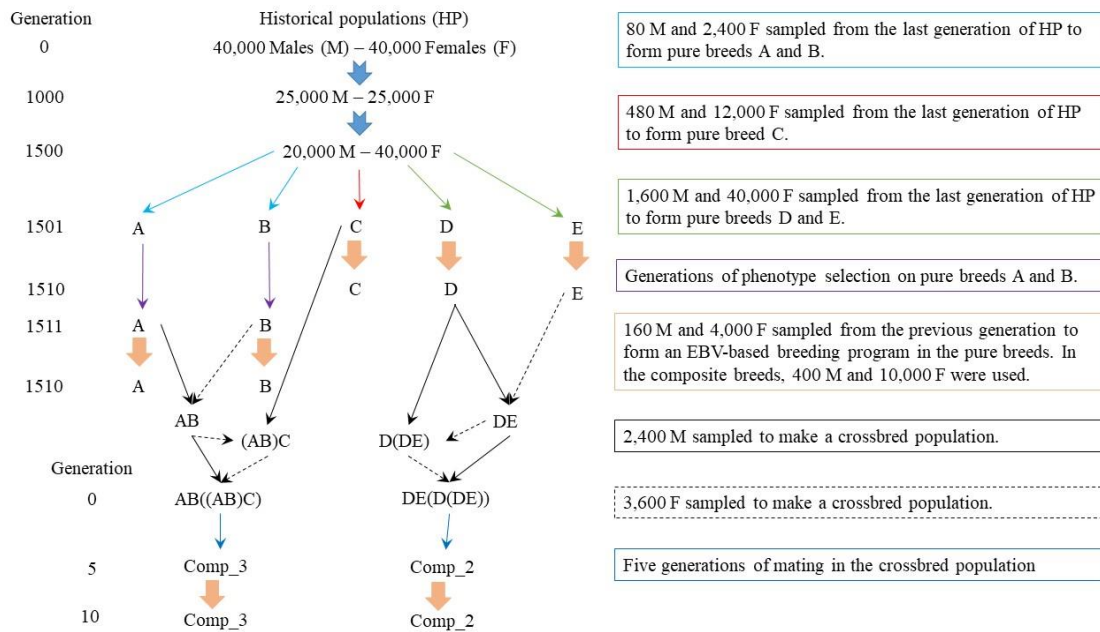


Figure 1. Simulation design to obtain pure and composite sheep populations.

Five random samples from the last historical population were selected to create five pure breeds, called A, B, C, D, and E (Figure 1). The combination of different founder population sizes (2,480 animals for the breeds A and B, 12,480 for the breed C, and 41,600 for the breeds D and E) and generations of phenotypic selection (10 for the breeds A and B, and one generation for the breeds C, D, and E) were used to achieve different LD patterns and, consequently, different N_e in the most recent populations. There were random matings and exponential increase in the number of females in a rate of 0.10 for the breeds A and B and 0.15 for the breeds C, D, and E. During the generations of phenotypic selection, it can be considered that the breeds were separated geographically, restricting the mating within each population. Subsequently, the pure breeds were divergently selected based on estimated breeding values (EBVs) predicted using BLUP, with breeds A, C, and D selected for increasing and breeds B and E for decreasing the EBVs for the simulated trait. All breeds were selected based on the EBVs during 10 generations. The male/female ratio in the EBV-selected populations was 1/25, with a replacement rate of 40% for males and 20% for females. There were single, double, and triple births, with the odds of 30%, 50%, and 20%, respectively, to be similar with the ones observed in sheep flocks. The number of individuals in

each generation of EBV-based selection were tested and at the end were greater than 7,000 to allow a reasonable number of selection candidates in each generation.

Crosses were made to obtain composite breeds, which had two or three pure breeds as the starting point (Figure 1). Two composite populations were created based on either two breeds (Comp_2), which had 62.5% of breed D and 37.5% of breed E (Figure 1), or three breeds (Comp_3), which had 37.5% of breed A, 37.5% of breed B, and 25.0% of breed C (Figure 1). Random mating was restricted within each crossbreed population for five generations. According to Rasali et al. (2006), five-to-six generations are sufficient to stabilize the frequencies of linked genes in new populations. Thereafter, the composite breeds were divergently selected using EBVs for the next 10 generations, with Comp_2 and Comp_3 divergently selected for decreasing and increasing performance, respectively. Mating type, sire and dam replacements, and the number of births per dam in the composite breeds were the same as those previously described for the pure breeds. The number of individuals per generation in the composite breeds (during the selection based on EBVs) was more than 18,000, to keep a higher N_e on those populations compared to the pure breeds.

2.1.2 Effective population size in the recent populations. The number of generations in the pure breeds during the expansion of the recent populations were modified accordingly to achieve the LD patterns corresponding to N_e of ~100, ~250, and ~500. The N_e was calculated using the LD and the realized inbreeding in the recent populations for pure and composite breeds under EBV-based selection. With the LD approach, N_e was estimated using the formula: $N_{eLD} = (4c)^{-1}\{[E(r^2)]^{-1} - 2\}$, which is a re-arrangement of the estimator $E(r^2) = (4N_{ec} + 2)^{-1}$ proposed by Sved (1971), where $E(r^2)$ is the expected LD for a population with effective size N_e , c is the genetic distance (chromosome segment size in Morgans—M) within autosomal chromosomes. It was considered that one Mb corresponds to a centimorgan (cM) when calculating the c value, as this is an acceptable approximation in sheep (Prieur et al., 2017). Lastly, populations were simulated to have an LD of approximately 0.024, 0.010, and 0.005 for SNPs spaced apart by 10 Mb, which correspond to the values of $E(r^2)$ for $N_e = 100, 250, \text{ and } 500$, respectively. A 10 Mb distance corresponds to an N_e that existed five generations ago (considered as current N_e), based on the relationship $t = 1/2c$ proposed by Hayes et al. (2003), where t is the number of generations ago and c is as previously defined. Estimation of LD was performed considering only SNPs with MAF higher than 0.05 using the r^2 metric (Hill and Robertson, 1968). We also estimated the N_e based on the realized inbreeding five generations ago using the formula (Falconer and Mackay, 1996): $N_{eInb} = 1/2\Delta F$, where $\Delta F = (F_n - F_{n-1})/(1 - F_{n-1})$ and F_n is the average inbreeding in the n th generation. The average inbreeding per generation was obtained from the QMSim software outputs (Sargolzaei and Schenkel, 2009).

2.1.3 Simulated traits. We simulated two traits with initial heritability levels of 0.30 and 0.10 (global parameters for the QMSim software; Sargolzaei and Schenkel, 2009), to represent moderate (MH2) and low (LH2) additive genetic effects, respectively, affecting the total phenotypic variability of the trait. The phenotypic variance was set to 100 in both simulations. The heritability was estimated in the recent populations based on pedigree and phenotype information using the AIREMLf90 software (Misztal et al., 2018) to verify if the desired values were achieved. All simulations were replicated five times using different seed values in order to simulate different populations. Only additive genetic effects were simulated due to the QMSim software (Sargolzaei and Schenkel, 2009) capabilities.

2.1.4 Genome and data editing. The genome was simulated with 26 autosomal chromosomes with size varying between 43 and 301 cM (a total of 2,656 cM), mimicking the sheep genome (Supplementary File 1). The number and size of chromosomes were defined based on information obtained from the most recent sheep reference genome (assembly OAR_v4.0) available in the NCBI platform (www.ncbi.nlm.nih.gov/genome?term=ovis%20aries). The genome simulation was also performed using the QMSim software (Sargolzaei and Schenkel, 2009).

A total of 3,057 QTLs were simulated, spanning the whole autosomal genome. The number of QTLs per chromosome varied between 51 and 391 (Supplementary File 1), which was chosen based on the information published in the AnimalQTLdb (AnimalQTLdb, 2019). QTLs with the number of alleles varying from two to six were simulated to evaluate the advantages of using haplotype-based approaches. All simulated markers were bi-allelic to mimic SNP markers, and the total number of SNPs was set to 576,595 (Supplementary File 1; similar number of autosomal SNPs included in the Ovine Infinium® HD SNP Beadchip 600K; FarmIQ, 2013; Kijas et al., 2014) sampled from the segregating loci ($MAF \geq 0.05$) in the last historical generation. The information on the number of markers in each chromosome was obtained from the SNPchiMp v.3 platform (Nicolazzi et al., 2015). Both QTL and markers were randomly distributed within chromosome and placed in different chromosomal positions, i.e., simulated QTLs were not among the SNPs, so that the genomic predictions rely only on the LD between them.

The additive genetic effects of the QTL were sampled from a gamma distribution with the shape parameter equal to 0.4, whereas no effects were simulated for the SNP markers. The initial allele frequencies assumed for QTL and markers (generation 0 of the historical population) were 0.5. The QTL heritability on the MH2 and LH2 traits was equal to 50% and 10% of the trait heritability, i.e., 0.15 and 0.01, respectively. The remaining genetic variance not explained by the QTLs was attributed to the polygenic effect. Recurrent mutation rates on the order of 1×10^{-4} were

simulated for the QTL and markers. Rates of 0.05 and 0.01 were used for the occurrence of missing genotypes and genotyping errors, respectively.

Quality control (QC) was performed in the genotype file of each simulated recent population for each replicate, using the PREGSf90 software from the BLUPf90 family programs (Misztal et al., 2018). In this step, SNPs with no extreme departure from Hardy–Weinberg equilibrium (difference between observed and expected frequency of heterozygous less than 0.15) and $MAF \geq 0.01$ were maintained. All SNPs passed this QC for all populations, indicating that there was enough variability on the simulated SNP chip panel.

2.2 Haplotype blocks construction

The FImpute v.3.0 software (Sargolzaei et al., 2014) was used to phase the genotypes (i.e., to infer SNP allele inheritance). Subsequently, the haploblocks were constructed using different LD thresholds (variable haploblock sizes), as described below. The r^2 metric (Hill and Robertson, 1968) was used to calculate the LD between markers to construct the haploblocks, as this measure is less sensitive to allele frequency (Bohmanova et al., 2010). The “gpart” package (Kim et al., 2019) implemented in the R software (R Core Team, 2020) was used to build the haploblocks considering r^2 levels of 0.1 (low), 0.3 (moderate), and 0.6 (high) based on the Big-LD approach (Kim et al., 2018). Following the previous definition of haploblocks (Gabriel et al., 2002), a haploblock in this study was considered as a genomic region spanning at least two SNPs.

2.3 Prediction of GEBV

All genomic predictions were performed using the ssGBLUP method implemented in the BLUPf90 family programs (Misztal et al., 2018). Before using the BLUPf90 software, the AIREMLf90 software (Misztal et al., 2018) was used to estimate the variance components for each simulation replicate for the models described in the next sections.

2.3.1 ssGBLUP using SNPs. The model used to predict the GEBVs under this approach was:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where \mathbf{y} is an $N \times 1$ vector of phenotypes for genotyped and non-genotyped animals, \mathbf{b} is the vector of fixed effects (i.e., generation), \mathbf{u} is a random vector of GEBVs for genotyped and non-genotyped animals with $\mathbf{u} \sim N(0, \mathbf{H}\sigma_g^2)$, \mathbf{e} is the vector of random errors with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, \mathbf{X} is the incidence matrix of fixed effects, and \mathbf{Z} is the incidence matrix that relates the records to GEBVs. In the case

of ssGBLUP fitting individual SNPs, the \mathbf{H} matrix is a hybrid relationship matrix that combines the genomic and pedigree relationships (Legarra et al., 2009), and its inverse can be computed directly in the mixed model equations as follows (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix},$$

where \mathbf{A}^{-1} is the inverse of pedigree relationship matrix, \mathbf{A}_{22}^{-1} is the inverse of pedigree relationship matrix for the genotyped animals, and \mathbf{G} is the genomic relationship matrix. The \mathbf{G} matrix was constructed as in the first method proposed by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{2\sum p_i(1-p_i)},$$

where \mathbf{M} is the matrix of centered genotypes, with a dimension equal to the number of animals by the number of markers. The blending and weighting parameters for the genomic information were the default values in the PREGSf90 software (α and β equal to 0.95 and 0.05, respectively, and τ and ω equal to 1.0; Misztal et al., 2018).

2.3.2 ssGBLUP using SNPs and haplotypes combined in a single genomic relationship matrix. The model and assumptions in this approach are the same as described in section 2.3.1. However, the \mathbf{G} matrix used to construct the combined relationship in this model had both independent markers (i.e., non-blocked markers, which are SNPs out of the LD blocks) and haplotypes as pseudo-SNPs. To build the \mathbf{G} matrix using haplotype information, the haplotype alleles were first converted to pseudo-SNPs, as in Teissier et al. (2020). Using this approach, if there were five unique haplotype alleles in a haploblock, five pseudo-SNPs were created for this haploblock. At the end, the number of copies of a specific pseudo-SNP allele were counted and coded as 0, 1, or 2 for each individual, similar to the codes used in \mathbf{M} (when creating the \mathbf{G}) as previously described based on individual SNPs. The pseudo-SNPs were subjected to the same QC steps as described above for individual SNPs.

2.3.3 ssGBLUP using haplotypes. The model and assumptions in this approach were the same as described in section 2.3.1. However, only haplotypes converted to pseudo-SNPs were used to create the \mathbf{G} matrix used in the predictions, therefore, excluding non-blocked individual SNPs.

2.3.4 ssGBLUP using SNPs and haplotypes assigned to two different genomic relationship matrices. The model used for these analyses was:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u}_1 + \mathbf{Z}\mathbf{u}_2 + \mathbf{e},$$

where \mathbf{u}_1 and \mathbf{u}_2 are the random additive genetic effects of the first and second component of the overall GEBV, respectively, which, under this modeling, is equal to $\mathbf{u}_1 + \mathbf{u}_2$. All other vectors and matrices on this model are the same as described on the previous sections. The main assumption on this model is that the breeding value is divided into two uncorrelated components with their own covariance structure, being $\mathbf{u}_1 \sim N(0, \mathbf{H}_1 \sigma_{g1}^2)$ and $\mathbf{u}_2 \sim N(0, \mathbf{H}_2 \sigma_{g2}^2)$, in which \mathbf{H}_1 and \mathbf{H}_2 are the hybrid relationship matrices with the same structure of the \mathbf{H} matrix described before. The only difference between \mathbf{H}_1 and \mathbf{H}_2 is the \mathbf{G} matrix that is combined with the pedigree relationship in each one of them, named as \mathbf{G}_1 and \mathbf{G}_2 , respectively, containing the genomic relationships between the individuals based on single non-blocked SNPs and haplotypes, respectively. This parametrization was used to account for the fact that haplotypes and, therefore, the corresponding pseudo-SNPs, are more polymorphic than individual SNPs. Consequently, pseudo-SNPs could better capture the effect of large-sized QTL with lower allele frequency than individual SNPs and could have different distribution of their allele effects compared to individual SNPs.

2.4 Training and validation population sets

The populations used in the genomic predictions were the pure breeds B, C, and E, defined as Breed_B, Breed_C, and Breed_E, respectively, and composite breeds Comp_2 and Comp_3. Only breeds Breed_B, Breed_C, and Breed_E were presented here because the genetic background simulated, i.e., the size of the founder population and generations of selection, was more divergent for these populations (Figure 1). As breeds A and D had similar sizes of the founder populations and generations of selection when compared to breeds B and E, respectively, we observed similar results between breeds A and B and also D and E (data not shown).

The datasets (populations from the simulated EBV-based selection programs) were divided into training and validation sets to test the accuracy and bias of GEBVs. The training sets within each population were composed of 60,000 individuals with phenotypes randomly sampled from generations one to eight, and 8,000 of them also had genotypes for the simulated HD panel. The genotyped individuals in the training set were randomly sampled from generations four to seven. The validation populations were composed of 2,000 individuals randomly sampled from generations nine and ten and were also genotyped for the same panel. Generation eight was considered as a gap between training and validation populations in terms of genotypes. The whole pedigree (generations 1 to 10) was used in all analyses. As we assume that validation individuals would not have phenotypes, their GEBVs were estimated based on the relationships of the validation cohort with the training set (with phenotypes and genotypes included in the analyses).

2.5 Evaluated scenarios

Although the HD SNP panel datasets were first simulated, the main genomic predictions were performed using a medium density 50K SNP panel, which was designed based on randomly selected SNPs from the original HD panel. This step was performed because similar accuracies tend to be achieved when using a medium density SNP panel in sheep (Moghaddar et al., 2017), as well as in other species (Binsbergen et al., 2015; Ni et al., 2017; Frischknecht et al., 2018). The total number of SNPs selected for the 50K panel was 46,827, as currently available in the 50K SNP panel (for autosomal chromosomes) reported in the SNPchiMp v.3 platform (Nicolazzi et al., 2015). The markers in the 50K SNP panel were randomly sampled within each autosome, and the number of SNPs per chromosome is reported in Supplementary File 1. In addition, previous analyses showed that both SNP and haplotype-based predictions based on the HD and 50K SNP panels were not statistically different (data not shown). Therefore, the haplotype blocks for all the prediction scenarios were created based on the 50K panel and the results for the HD SNP panel were presented as an additional scenario.

At the end, 11 scenarios were evaluated, which consisted of genomic predictions using: 1) SNPs from the 600K; 2) SNPs from the 50K; 3 to 5) independent SNPs and pseudo-SNPs from haplotype blocks with LD equal to 0.1, 0.3, and 0.6 in a single relationship matrix (IPS_LD01, IPS_LD03, and IPS_LD06, respectively); 6 to 8) only pseudo-SNPs from haplotype blocks with LD equal to 0.1, 0.3, and 0.6 (PS_LD01, PS_LD03, and PS_LD06, respectively); and 9 to 11) independent SNPs and pseudo-SNPs from haplotype blocks with LD equal to 0.1, 0.3, and 0.6 in two different relationship matrices (IPS_2H_LD01, IPS_2H_LD03, and IPS_2H_LD06, respectively). All these scenarios were evaluated for two different heritability levels (moderate and low) and in each one of the five populations previously described (purebred and composite breeds with distinct N_e). Therefore, 110 different scenarios were evaluated in each one of the five replicates. A summary of the evaluated scenarios is shown in Figure 2.

2.6 Scenario comparisons

The statistics related to haplotype blocking strategies were compared between populations (pure and composite breeds) within each LD threshold to create the blocks (0.1, 0.3, and 0.6), and also, the LD thresholds were compared within each population to differentiate the haplotype block structures. These statistics are: average number of haploblocks, blocked SNPs, pseudo-SNPs before and after QC, non-blocked plus pseudo-SNPs after QC, and the additional computer time required by using pseudo-SNPs (e.g., SNPs phasing, haplotype blocking, pseudo-SNP derivation). The GEBV accuracies and bias in each prediction scenario were compared within each population, to

mimic population-specific (breed) genetic evaluation. Prediction accuracy was estimated as the Pearson correlation coefficient between the GEBVs and TBVs for the validation animals, for each replicate and scenario. Prediction bias was assessed as the deviation from one of the linear regression coefficients (β_1) of the TBVs on the GEBVs (i. e., $bias = \beta_1 - 1$; where $TBV = \beta_0 + \beta_1 \times GEBV$) in the validation population in each replicate and scenario.

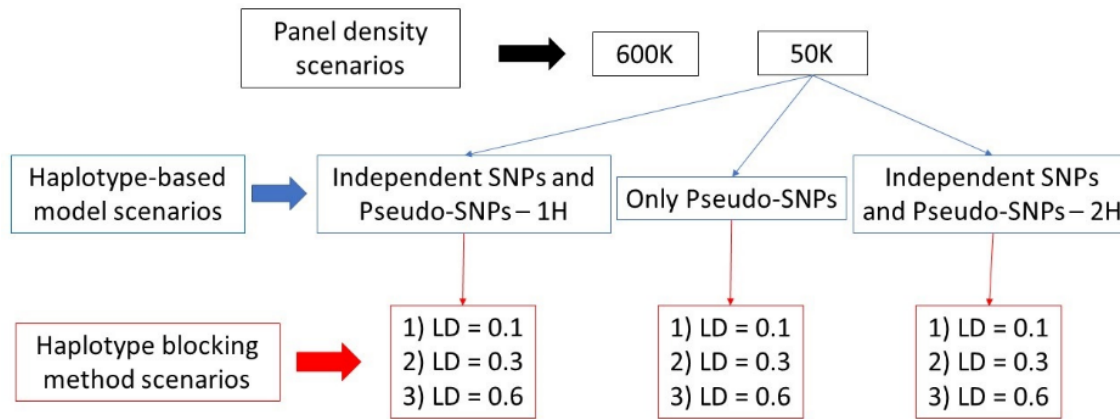


Figure 2. Evaluated scenarios used in the genomic predictions with pseudo- single nucleotide polymorphisms (SNPs) from linkage disequilibrium (LD) blocks using independent and pseudo-SNPs in a single genomic relationship matrix (1H), and only pseudo-SNPs and independent and pseudo SNPs in two genomic relationship matrices (2H).

A linear mixed model was used to test the effect of the population and LD level on the statistics from haplotype block strategies and the effect of marker information (SNP and haplotype prediction scenarios) on the accuracy and bias of GEBV prediction. The statistical model used was:

$$y_{ij} = \mu + T_i + R_j + \varepsilon_{ij}$$

where y_{ij} is the observation of the i^{th} treatment on the j^{th} repetition; T_i is the treatment effect, in which i is equal to Breed_B, Breed_C, Breed_E, Comp_2, and Comp_3 to compare the population effect over the statistics from haplotype block strategies within each LD threshold; equal to LD01, LD03, and LD06 to compare the effects of LD level over the statistics from haplotype block strategies within population; and equal to 600K, 50K, IPS_LD01, IPS_LD03, IPS_LD06, PS_LD01, PS_LD03, PS_LD06, IPS_2H_LD01, IPS_2H_LD03, and IPS_2H_LD06 to test the effect of marker information over the accuracy and bias of GEBV prediction within each population; R_j is the random effect of replicates which was assumed to follow $\sim N(0, \mathbf{B}\sigma_b^2)$; and ε_{ij} is the residual effect of the model.

Replicate was used as a random effect in the model to account for the covariance between the scenarios, as the compared averages were obtained within the simulated populations in each replicate. This was done to reduce the occurrence of false negatives (Type-II error). Different

covariance structures (**B**) were evaluated (spherical, compound symmetry, simple autoregressive process, and unstructured covariance) to explain the covariances between replicates, and the structure that presented the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values was used in the final models for comparison purposes. After defining the appropriate covariance structure (which was not the same for all scenarios, with unstructured covariance being the best in the major part of the scenarios), the means of the T_i levels were compared using the Tukey test at 5% of significance level. The “nlme” (Pinheiro et al., 2021) and “emmeans” (Lenth, 2021) R packages were used to fit the models and compare the means, respectively, in the R environment (R Core Team, 2020).

3 Results

3.1 Genetic diversity and genetic parameters in the simulated populations.

After the simulation process, several different N_e levels were observed in the recent populations studied (generations 1 to 10 of pure and composite breeds under EBV-based selection). The total additive genetic effect variances estimated with the models that used two **H** matrices (section 2.3.4), taken as $\sigma_{g1}^2 + \sigma_{g2}^2$, and the residual variances were similar to the variances estimated with the models that fitted a single **H** matrix (sections 2.3.1, 2.3.2, and 2.3.3) and similar to the variances estimated with the model that used only the pedigree relationship matrix (section 2.1.3; Supplementary Files 3 and 4). Therefore, for simplicity, only the genetic parameters estimated based on the pedigree relationship matrix are presented in Table 1. A population structure analysis based on principal components (PCs) of the **G** matrix using the SNPs from the 50K panel was also performed (Supplementary File 2). Individuals within the population were close to each other, and no clear clusters between populations existed at 95% confidence level based in the approximated unbiased test from a hierarchical clustering method using 10,000 bootstrap samples (Shimodaira, 2002; Supplementary File 2).

3.1.1 N_e and genetic parameters for the simulation of a trait with moderate heritability. The average N_{eLD} ranged between 110 and 644 (Breed_B and Comp_2, respectively), while the N_{eInb} varied from 159 to 373 (Breed_B and composite breeds, respectively), being lower in pure breeds independently of the N_e measure (Table 1 and Supplementary File 3). The average additive genetic variance in the MH2 scenarios ranged from 25.82 (Comp_2) to 28.09 (Breed_C), while the residual variances ranged from 70.85 (Breed_C) to 73.07 (Comp_2). Average heritability estimates ranging from 0.26 (Comp_2) to 0.29 (Breed_C) were observed across populations, which

are close to the global simulation parameters (heritability and phenotypic variance equal to 0.30 and 100, respectively).

Table 1. Average (SE) effective population size based on the linkage disequilibrium (N_{eLD}) and realized inbreeding (N_{eInb}) methods, additive genetic variance (σ_a^2), residual variance (σ_e^2), and heritability (h^2) estimates of the trait in simulated sheep populations.

Simulation	Population ¹	N_{eLD} ²	N_{eInb} ³	σ_a^2	σ_e^2	h^2
Moderate h^2 (0.30)	Breed_B	110 (6)	190 (17)	27.12 (0.27)	71.54 (0.10)	0.27 (0.00)
	Breed_C	379 (8)	260 (15)	28.09 (0.25)	70.85 (0.26)	0.29 (0.00)
	Breed_E	359 (5)	192 (6)	27.45 (0.35)	72.42 (0.34)	0.28 (0.00)
	Comp_2	644 (15)	446 (7)	25.82 (0.37)	73.07 (0.25)	0.26 (0.00)
	Comp_3	466 (40)	447 (53)	26.80 (0.62)	72.88 (0.50)	0.27 (0.00)
Moderate h^2 (0.10)	Breed_B	125 (8)	94 (11)	9.17 (0.26)	90.30 (0.38)	0.09 (0.00)
	Breed_C	272 (11)	120 (11)	9.31(0.28)	89.91(0.23)	0.09 (0.00)
	Breed_E	251 (22)	119 (19)	9.31 (0.23)	90.38 (0.26)	0.09 (0.00)
	Comp_2	522 (32)	259 (40)	8.42 (0.27)	91.13 (0.27)	0.08 (0.00)
	Comp_3	407 (32)	235 (38)	8.00 (0.29)	91.90 (0.23)	0.08 (0.00)

¹Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds based on two and three pure breeds, respectively.

²Estimated based on the re-arranged estimator present in Sved (1971).

³Estimated based on the formula presented by Falconer and Mackay (1996).

3.1.2 Ne and genetic parameters for the simulation of a low heritability trait. The average N_{eLD} ranged from 125 (Breed_B) to 522 (Comp_2), while N_{eInb} ranged between 94 and 259 for these same populations (Table 1 and Supplementary File 4). Average additive genetic variances ranging from 8.00 (Comp_3) to 9.31 (Breed_C and Breed_E) were observed. The average residual variances ranged from 90.30 (Breed_B) to 91.90 (Comp_3). In the LH2 scenarios, the average heritabilities were equal to 0.09 in the pure breeds and 0.08 in the composite breeds, which are close to the global simulation parameters (heritability and phenotypic variance equal to 0.10 and 100, respectively).

3.2 Statistics from haplotype blocks and pseudo-SNPs: moderate heritability trait

3.2.1 Number of blocks. The average number of blocks with two or more SNPs and the LD threshold equal to 0.1 ranged from 7,709.6 (Comp_2) to 8,607.6 (Comp_3), with Comp_2 and Breed_B showing similar and significantly lower number of blocks with this LD threshold level than the other populations (Figure 3A and Supplementary File 5). With the LD threshold equal to 0.3, the average number of blocks ranged from 145.0 (Comp_2) to 3,574.6 (Breed_B), and Breed_B showed significantly larger mean compared to the other populations (Figure 3B and Supplementary File 5). Only Breed_B had blocks with an LD threshold equal to 0.6, with an average equal to 23.8, which was statistically different from all the other populations (Figure 3C and Supplementary File 5). Within each population, the mean number of blocks from LD threshold levels of 0.1, 0.3, and

0.6 were statistically different for all populations, with the LD threshold equal to 0.1 being the largest, followed by the LD threshold equal to 0.3, and the 0.6 level yielding the lowest number of blocks.

3.2.2 Number of blocked SNPs. The average number of blocked SNPs for the LD threshold equal to 0.1 varied between 17,122.2 (Comp_2) and 19,199.8 (Comp_3) (Figure 3A and Supplementary File 5), and for Comp_2, it was significantly lower than all the other populations. The average number of SNPs within blocks with an LD threshold equal to 0.3 ranged from 340.4 (Comp_2) to 8,195.4 (Breed_B) (Figure 3B and Supplementary File 5). The number of blocked SNPs for Breed_B was significantly higher than for the other populations (which did not differ among them). The average number of blocked SNPs with LD threshold equal to 0.6 in Breed_B was 56.8 (Figure 3C and Supplementary File 5) and was significantly greater, as no blocks were created for all the other populations.

3.2.3 Number of pseudo-SNPs after quality control. After QC, the average number of pseudo-SNPs from blocks with an LD threshold equal to 0.1 was reduced, ranging from 35,524.6 (Comp_2) to 39,713 (Breed_E) (Figure 3A and Supplementary File 5). In general, Breed_B and Comp_2 were statistically similar and had lower averages compared to all other populations. The average number of pseudo-SNPs after QC with haploblocks constructed with the LD threshold of 0.3 was between 718.6 (Comp_2) and 16,259.4 (Breed_B), in which only Breed_B was statistically different from all other populations (Figure 3B and Supplementary File 5). With an LD threshold equal to 0.6, the average number of pseudo-SNPs for Breed_B was 91 and no pseudo-SNPs were generated with this LD threshold for all the other populations (Figure 3C and Supplementary File 5). The average number of pseudo-SNPs before QC is also shown in Figure 3A and Supplementary File 5.

3.2.4 Number of non-blocked SNPs plus pseudo-SNPs after quality control. The average number of non-blocked plus pseudo-SNPs after QC varied from 64,987.0 (Breed_B) to 67,367.2 (Breed_E) when using blocks with an LD threshold of 0.1 (Figure 3A and Supplementary File 5). Breed_B and Comp_2 showed lower averages compared to all the other populations. Regarding the LD threshold of 0.3, the number of non-blocked plus pseudo-SNPs after QC ranged from 47,205.2 (Comp_2) to 54,891.0 (Breed_B) (Figure 3B and Supplementary File 5). For this LD threshold, the Breed_B average was statistically greater than all the other populations. The average number of non-blocked plus pseudo-SNPs after QC was equal to 46,867.8 for Breed_B and 46,827 for all the other populations when using an LD threshold of 0.6 to create the haploblocks (Figure 3C and Supplementary File 5).

3.2.5 Additional time to create pseudo-SNPs. The average computing time to create the pseudo-SNPs (also considering the haplotype phasing and blocking) was between 8,800.6 s (2 h and 26 min; Comp_2) and 22,650.0 s (6 h and 18 min; Breed_B) with the LD threshold of 0.1 (Figure 3A and Supplementary File 5). For this LD threshold, the computing time for Breed_B was statistically similar to that in Breed_C, but significantly different from all the other populations. When using an LD threshold of 0.3 to create the blocks, the average computing time ranged from 675.4 s (11 min; Comp_2) to 2,935.0 s (49 min; Breed_B) (Figure 3B and Supplementary File 5). The computing time for Breed_B was statistically higher than all the other populations, which were not statistically different among them. The average computing time for pseudo-SNPs from blocks with an LD threshold equal to 0.6 ranged from 591.4 (10 min) to 666.8 s (11 min) (Breed_C and Breed_B, respectively; Figure 3C and Supplementary File 5), and no statistical differences were observed across populations. The computing time compared across LD thresholds within the population showed that LD thresholds of 0.3 and 0.6 were statistically similar and lower than with the LD threshold of 0.1.

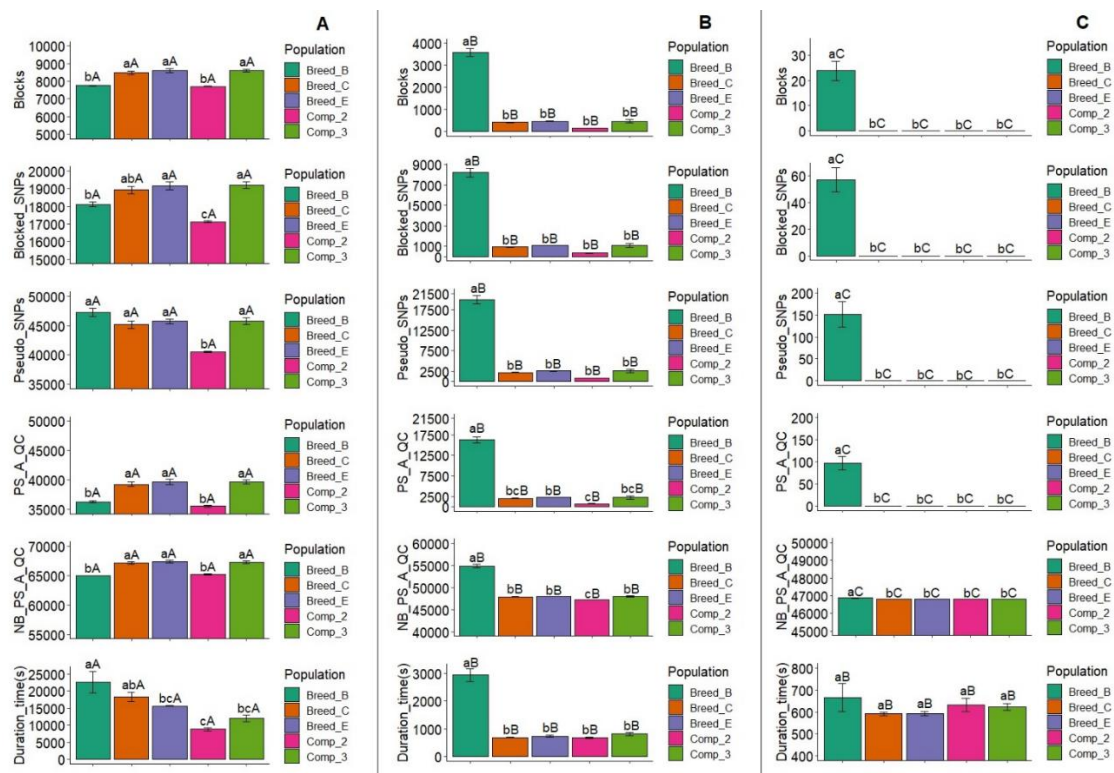


Figure 3. Average number of blocks (Blocks) spanning two or more SNPs, markers within blocks (Blocked_SNPs), pseudo-SNPs (Pseudo_SNPs), pseudo-SNPs after quality control (PS_A_QC), non-blocked SNPs plus pseudo-SNPs after quality control (NB_PS_A_QC), and computing time to obtain the pseudo-SNPs (Duration_time) in the simulation for a trait with moderate heritability ($h^2 = 0.30$). A, B, and C show the results for haplotype blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively. Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds from two and three pure breeds, respectively. The same lower- or upper-case letters mean no

statistical difference comparing populations within LD thresholds and LD threshold across populations, respectively, at 5% significance level by the Tukey test.

3.3 Statistics from haplotype blocks and pseudo SNPs: low heritability trait

We have also checked the statistics from haplotype blocks and pseudo-SNPs in the low heritability trait scenarios because the simulation was done for each heritability level at a time. In general, the number of blocks, blocked SNPs, pseudo-SNPs before and after the QC, the number of non-blocked plus pseudo-SNPs after QC, and computing time to generate the pseudo-SNPs for a trait with a low heritability were similar to those for a trait with moderate heritability and are shown in Figure 4 and Supplementary File 6. The results for the statistical comparisons in each one of these metrics for both populations, within each LD threshold, and for LD thresholds across populations were also similar between the LH2 and MH2 scenarios. The exceptions for the statistical comparisons under LH2 scenario was that the number of blocks in Breed_C and Breed_E would show a similar or lower average number of blocks, blocked SNPs, pseudo-SNPs after QC, and number of non-blocked plus pseudo-SNPs after QC than Breed_B, whereas the opposite would occur under the MH2 scenario. However, as pointed out before, the values were similar across the LH2 and MH2 scenarios. Therefore, the interpretation of the statistical comparisons for haplotype blocks in the MH2 scenario are also extended to LH2.

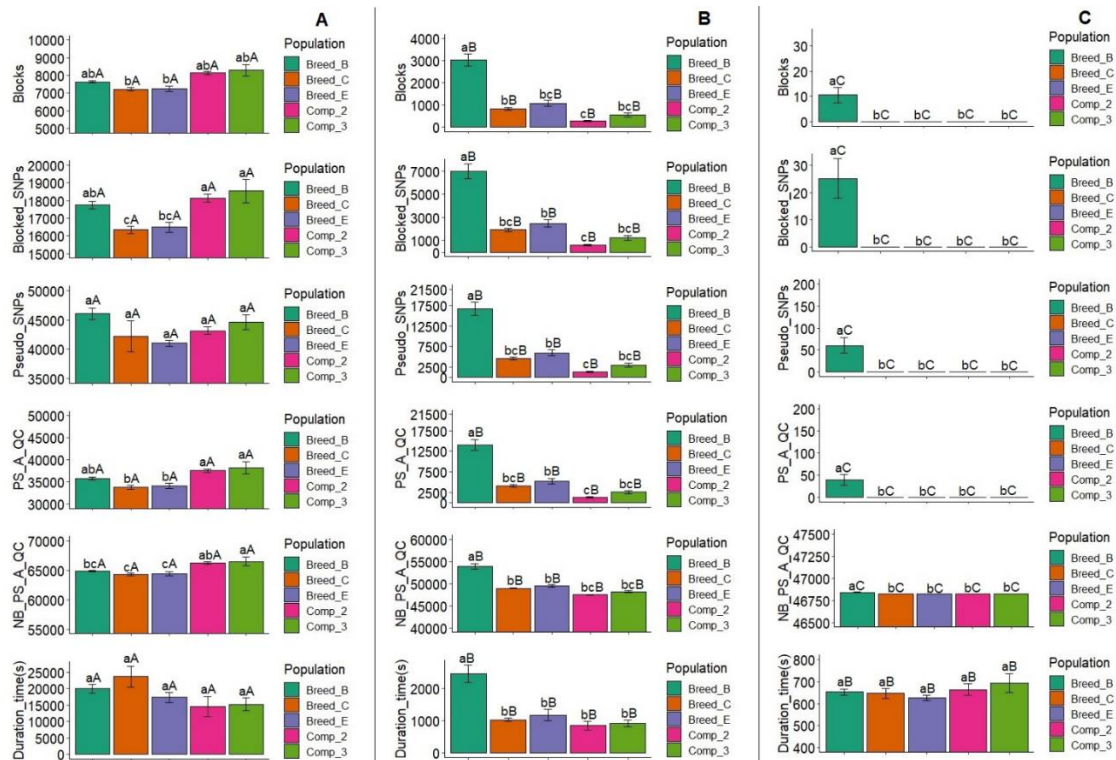


Figure 4. Average number of blocks (Blocks) spanning two or more SNPs, markers within blocks (Blocked_SNPs), pseudo-SNPs (Pseudo_SNPs), pseudo-SNPs after quality control (PS_A_QC), non-blocked SNPs plus pseudo-SNPs after quality control (NB_PS_A_QC), and computing time to obtain the pseudo-SNPs (Duration_time) in the simulation for a trait with

low heritability ($h^2 = 0.10$). **A, B, and C** show the results for the haplotype blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively. **Breed_B, Breed_C, and Breed_E**: simulated pure breeds with different genetic backgrounds; **Comp_2 and Comp_3**: composite breeds from two and three pure breeds, respectively. The same lower- or upper-case letters mean no statistical difference comparing populations within LD thresholds and LD threshold across populations, respectively, at 5% significance level based on the Tukey test

3.4 Accuracy and bias of genomic predictions: moderate heritability trait

3.4.1 Pure breed with lower genetic diversity (Breed_B). The average accuracy for GEBVs based on individual SNPs in the Breed_B was 0.54 and 0.55 for the 50K and 600K panels, respectively, whereas it varied from 0.48 (pseudo-SNPs from blocks with an LD threshold of 0.3, PS_LD03) to 0.54 (independent SNPs and pseudo-SNPs from blocks with an LD threshold of 0.6, IPS_LD06) using haplotypes (Figure 5A, Supplementary File 7). In general, genomic predictions that used pseudo-SNPs and independent SNPs in one or two relationship matrices did not statistically differ from those with SNPs in the 50K and 600K panels. Using only pseudo-SNPs in the genomic predictions showed significantly lower accuracy than all other methods, when considering an LD threshold equal to 0.1 and 0.3 to create the blocks (PS_LD01 and PS_LD03, respectively). No predictions with PS_LD06 and IPS_2H_LD06 (independent SNPs and pseudo-SNPs from blocks with an LD threshold of 0.6 in two relationship matrices) were performed due to the low correlations observed between off-diagonal elements in \mathbf{A}_{22} and \mathbf{G} constructed with only pseudo-SNPs from haploblocks with an LD threshold of 0.6 (Supplementary File 8). The average GEBV bias was equal to -0.09 and -0.08 for the 50K and 600K SNP panels, respectively, whereas it ranged between -0.20 (PS_LD03) and -0.08 (IPS_2H_LD01) with haplotypes. No statistical differences were observed in the average bias when the two SNP panel densities or the independent and pseudo-SNP in one or two relationship matrices were used. PS_LD01 and PS_LD03 generated statistically more biased GEBVs than all the other scenarios.

3.4.2 Pure breed with medium-size founder population and moderate genetic diversity (Breed_C). The average accuracy observed in the Breed_C was equal to 0.53 and 0.54 with the 50K and 600K, respectively, while with haplotypes, it ranged from 0.25 (PS_LD03) to 0.52 (IPS_LD03) (Figure 5A, Supplementary File 7). Similar to Breed_B, the PS_LD01 and PS_LD03 models yielded statistically less accurate GEBVs than all the other models, with PS_LD03 being the worst one. Fitting pseudo-SNPs and independent SNPs in one or two relationship matrices did not have statistical differences when compared with individual-SNP predictions. The IPS_2H_LD03 scenario did not converge during the genetic parameter estimation, and no pseudo-SNPs were generated for any haplotype method that used an LD threshold of 0.6 (IPS_LD06, PS_LD06, and IPS_2H_LD06). Consequently, no results were obtained for these scenarios.

Average GEBV bias equal to -0.05 and -0.02 were observed for the 50K and 600K SNP panels, whereas in the haplotype-based predictions, it ranged from -0.49 (PS_LD03) to -0.03 (IPS_2H_LD01). PS_LD01 and PS_LD03 were statistically more biased than all the other scenarios (statistically similar among them).

3.4.3 Pure breed with larger founder population and moderate genetic diversity (Breed_E). The average accuracy was equal to 0.52 and 0.53 for the 50K and 600K SNP panel, respectively, while the haplotype-based approach yielded accuracy varying between 0.28 (PS_LD03) and 0.51 (IPS_LD03) in Breed_E (Figure 5A, Supplementary File 7). Using only pseudo-SNPs from haplotype blocks with an LD threshold of 0.3 (PSLD03) yielded the less accurate genomic predictions, being statistically lower than all the other models (with similar accuracy among them). No blocks with an LD threshold equal to 0.6 were created in this population, and therefore, no predictions were obtained with the models that would use pseudo-SNPs from these blocks. For the GEBV bias, averages of -0.09 and -0.06 were observed for the 50K and 600K panels, respectively, ranging from -0.53 (PS_LD03) to -0.09 (IPS_2H_LD01) when haplotypes were fitted. Similar to the accuracy findings, the PSLD03 showed statistically lower average GEBV bias of prediction compared to all other models, showing the more biased predictions.

3.4.4 Composite breed from two populations with high genetic diversity (Comp_2). The average accuracy for the 50K and 600K SNP panels in Comp_2 were 0.41 and 0.42, respectively, with haplotype-based predictions ranging from 0.17 (PSLD03) to 0.41 (IPS_LD03) (Figure 5A, Supplementary File 7). As observed in the pure breeds, there were no statistical differences between the predictions with SNPs based on both SNP density panels and the scenarios that fitted pseudo-SNPs and independent SNPs in one or two relationship matrices. Using only pseudo-SNPs to create the **G** matrix also provided statistically lower accuracy, with PS_LD03 yielding the worst results. No predictions were made with IPS_2H_LD03 in this population because of convergence problems during the genetic parameter estimation process. No pseudo-SNPs were obtained with the LD threshold of 0.6 and, consequently, no subsequent genomic prediction results. Average GEBV bias of -0.14 and -0.10 was observed for the 50K and 600K SNP panels, respectively, while the average GEBV bias ranged from -0.62 (PS_LD03) to -0.15 (IPS_2H_LD01) when fitting haplotypes. Statistically, more biased predictions were obtained only when pseudo-SNPs from haplotype blocks with an LD threshold of 0.3 were used (PS_LD03).

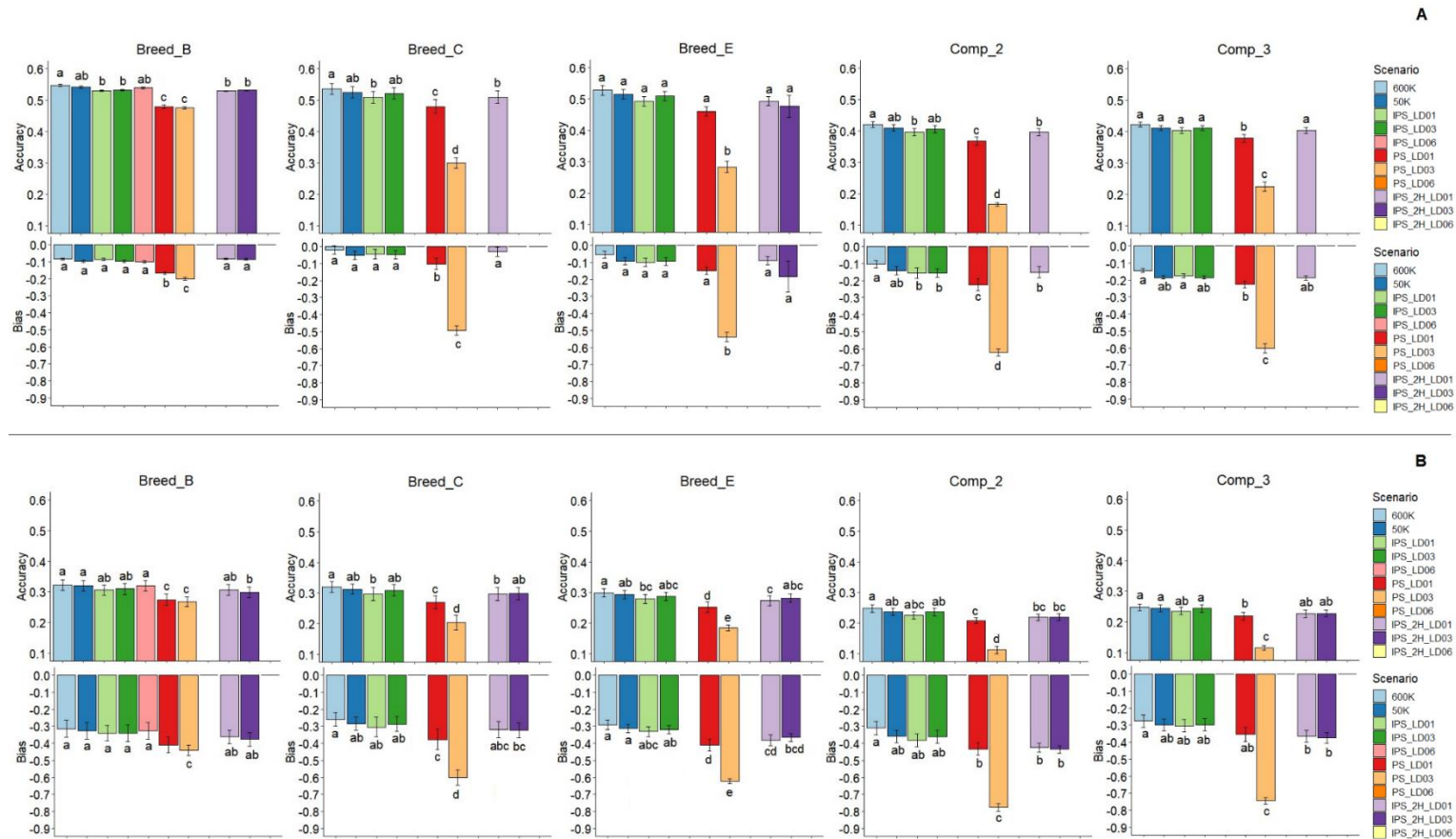


Figure 5. Accuracies and bias of genomic predictions based on individual SNPs and haplotypes for the simulations of traits with moderate (A) and low (B) heritability (0.30 and 0.10, respectively). Breed_B, Breed_C, and Breed_E: simulated pure breeds with different genetic backgrounds; Comp_2 and Comp_3: composite breeds from two and three pure breeds, respectively. 600K: high-density panel; 50K: medium-density panel; IPS_LD01, IPS_LD03, and IPS_LD06: independent and pseudo-SNPs from blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively, in a single genomic relationship matrix; PS_LD01, PS_LD03, and PS_LD06: only pseudo-SNPs from blocks with LD threshold of 0.1, 0.3, and 0.6, respectively; and IPS_2H_LD01, IPS_2H_LD03, and IPS_2H_LD06: independent and pseudo-SNPs from blocks with LD thresholds of 0.1, 0.3, and 0.6, respectively, in two genomic relationship matrices. Zero values for both accuracies and bias mean no results were obtained, due to poor quality of genomic information or no convergence of the genomic prediction models. The same lower-case letters mean no statistical difference comparing genomic prediction methods within population at 5% significance level based on the Tukey test

3.4.5 Composite breed from three populations with high genetic diversity (Comp_3). The average accuracy for the 50K and 600K SNP panels were 0.41 and 0.42, respectively, and with haplotype-based predictions, they ranged from 0.22 (PS_LD03) to 0.41 (IPS_LD03) (Figure 5A, Supplementary File 7). The PS_LD01 and PS_LD03 scenarios yielded statistically lower accuracy than all the other methods (statistically similar among them). Similarly to Comp_2, no genomic predictions were performed for the IPS_2H_LD03 and models fitting pseudo-SNPs from blocks with an LD threshold of 0.6. The average GEBV bias was -0.19 and -0.14 for the 50K and 600K SNP panels, respectively, and ranged from -0.60 (PS_LD03) to -0.18 (IPS_LD01) for the haplotype-based predictions. Using only pseudo-SNPs from LD blocks constructed based on an LD threshold of 0.3 resulted in more biased GEBV predictions for the Comp_3 population.

3.5 Accuracy and bias of genomic predictions: low heritability trait

The effects of fitting haplotypes in the genomic predictions under the LH2 scenarios were similar to those observed in the MH2 scenarios for all populations, with also similar average results (Figure 5B and Supplementary File 9). Therefore, the interpretations of the results for MH2 can be extended to the LH2 scenario, in which the worst results were observed for the PS_LD03 and similar accuracy and bias using SNPs or haplotypes (with independent SNPs) were observed. The GEBVs from the LH2 scenarios were less accurate and more biased than those from the MH2 scenarios within populations (e.g., lower accuracy and greater bias in LH2 within Breed_B), as would be expected due to the lower heritability of the trait. No GEBV predictions were made for the PS_LD06 and IPS_2H_LD06 for Breed_B due to the low correlation between the off-diagonal elements of the \mathbf{A}_{22} and \mathbf{G} created with pseudo-SNPs from blocks with an LD threshold of 0.6 (Supplementary File 10). No results for all scenarios fitting pseudo-SNPs from blocks with an LD threshold of 0.6 were obtained for Breed_C, Breed_E, Comp_2, and Comp_3 because no blocks were created based on this threshold.

4. Discussion

We hypothesized that the predicted GEBV in populations with higher genetic diversity, such as composite sheep breeds (e.g., Kijas et al., 2012; Brito et al., 2017b; Oliveira et al., 2020), could benefit from the use of haplotype-based rather than SNP-based genomic predictions, by obtaining GEBVs with higher accuracy and lower bias of

prediction. Therefore, we investigated the impact of including haplotype information in ssGBLUP for populations with high genetic diversity, assessed based on the N_e metric, and different genetic background. Furthermore, we evaluated the performance of haplotype-based models by fitting the haplotypes as pseudo-SNPs in different ways under the ssGBLUP framework. For that, we considered only pseudo-SNPs to construct the genomic relationships and also two different relationship matrices (i.e., derived from individual SNPs and pseudo-SNPs from haplotype blocks), assuming no correlation between them. To evaluate our hypothesis, simulated data was used to calculate the true accuracy and bias of genomic predictions for simulated traits with moderate and low heritability level. These two sets of heritability levels comprise the major part of traits of interest in livestock breeding programs (e.g., growth, carcass, feed efficiency, reproductive performance, disease resistance, overall resilience).

4.1 Genetic diversity and genetic parameters

The genetic diversity and variance components were assessed in the subsets of the data used for the predictions to verify the consistency of the initial simulation parameters. In addition to the first three recent N_e idealized at the beginning of this study (100, 250, and 500), several other genetic diversity measures were obtained after the simulation process was finalized, which are measures of recent N_e (until five generations ago) based on LD (N_{eLD}) and on realized inbreeding (N_{eInb}) (Table 1 and Supplementary Files 3 and 4). N_{eLD} would be more useful in the absence of accurate pedigree information, as it relies on the $E(r^2)$ estimation in a pre-defined chromosomal segment size and was proposed for simpler population structures (e.g., random mating and no selection; Sved, 1971). However, we also calculated N_{eInb} as an alternative indicator of N_e , because this estimate is based on the realized inbreeding and relies on the actual increase in population autozygosity (Falconer and Mackay, 1996).

One thousand and six hundred individuals from each one of the five populations (8,000 in total) were used to obtain the principal components (PCs) shown in Supplementary File 2, which actually explained a small proportion of the overall variance (1.71% and 2.13% for the first two and first three PCs, respectively). MacVean (2009) highlighted several situations that can affect the structure and spatial distribution of the PCA using SNPs (e.g., current and recurrent bottlenecks, admixture, waves of expansion, sample size) and potentially cause bias in the scatter with the first PCs, especially if they

explain a little proportion of the overall variance. Rao (1964) also indicated that inferences about structural relationships using the first PCs are only recommended when they explain a substantial amount of variation, which was not our case. Also, Deniskova et al. (2019) found a sheep population with a lower N_e (176) more scattered in the first two PCs than populations with higher N_e (>500), indicating the need for a third PC to observe differences within the high genetically diverse, similar to what we observed in this current study. The authors mentioned that a small founder population could be the reason for the lower N_e in the more scattered population along the first two PCs, and the Breed_B in our study (lower N_e) also had the smallest founder population. Another important point to highlight is that when using commercially available SNP chips, there tends to be ascertainment bias in the design of the SNP panels, which then contributes to a greater differentiation among populations (depending if they contributed or not to the SNP panel design) and crossbred/composite animals tend to have greater SNP diversity and be more scattered in the plots. This does not tend to happen when using simulated datasets. In summary, as it is not recommended to make inferences with PCs that are not significant (Rao, 1964; MacVean, 2009), the N_e should be used to make conclusions about the genetic diversity of the simulated populations, with the PCs used only for the illustration of the population structure.

Both N_e measures showed values close to those observed for some terminal and composite sheep breeds (125 to 974) as reported by Brito et al. (2017b), indicating that the simulation analyses resulted in datasets mimicking the genetic structure of commercial sheep populations. In addition to sheep, other species also present similar genetic diversity levels to some of the simulated populations used in this research, such as goats (N_e from 38 to 149; Brito et al., 2015), beef cattle (N_e from 153 to 220; Biegelmeyer et al., 2016), and dairy cattle (N_e from 58 to 120; Mekanjuola et al., 2020). The genetic parameters estimated after the simulation process were similar and consistent among replicates across all recent populations used for the subsequent analyses in both scenarios (MH2 and LH2; Table 1 and Supplementary Files 3 and 4).

4.2 Statistics from haplotype blocks and pseudo-SNPs

The differences observed on the haplotype block statistics across the simulated populations within LD thresholds and also across LD thresholds within populations are a consequence of the genetic events experienced by them. The number and size of the LD blocks can vary according to recombination hotspots and evolutionary events such as mutation, selection, migration, and random drift (McVean et al., 2004). In this context, a lower number of blocks with high LD thresholds would be expected in more genetically diverse populations, simply because in these populations, a large number of SNPs are expected to be excluded from all haploblocks, left to be considered as individual SNP effects. This was observed in Breed_B (less diverse, N_e ranging from 94 to 159) having a larger number of blocks not only when 0.6 was used as the LD threshold but also when the LD threshold was set to 0.3 in both MH2 and LH2 scenarios (Figures 3 and 4 and Supplementary Files 5 and 6).

The average number of blocks was similar (LH2, Figure 4 and Supplementary File 7) or even lower (MH2, Figure 3 and Supplementary File 6) in Breed_B compared to the other populations when the LD threshold was set to 0.1. The Big-LD method used in this study defines the LD blocks by using weights estimated based on the number of SNPs from all possible overlapping intervals (Kim et al., 2018). Therefore, low LD thresholds could imply in similar intervals to derive the independent blocks regardless of the level of genetic diversity in populations derived from the same historical population (i.e., same species). When setting low LD thresholds to construct the LD-blocks, more intervals of linked SNPs are obtained as the number of blocks increase with less SNPs excluded (and vice versa). Therefore, this might explain the distribution of the number of blocks across populations with an LD threshold of 0.1. Consequently, a greater number of blocks are expected, as observed when comparing the number of blocks across LD thresholds (the number of blocks with an LD threshold of $0.1 > 0.3 > 0.6$, Figures 3 and 4 and Supplementary Files 5 and 6).

The number of blocked SNPs and pseudo-SNPs before and after QC in both MH2 and LH2 (Figures 3 and 4 and Supplementary Files 5 and 6) is a function of the genetic diversity level of the populations. Longer blocks with many SNPs are expected in less genetically diverse populations (Hayes et al., 2003; Villumsen et al., 2009; Hess et al., 2017) likely due to selection and inbreeding, whereas more pseudo-SNPs (unique haplotypes) are expected in more genetically diverse populations (Teissier et al., 2020), when the single SNPs out of the LD-clusters are not considered as a block, following the

standard definition of haplotype block (Gabriel et al., 2002). However, this also depends on the LD threshold used to create the haplotype blocks, as this pattern was clear only when LD was greater than 0.1.

Independently of the LD level used to create the blocks, the relative reduction in the number of pseudo-SNPs after QC was greater on the less genetically diverse population, with approximately 40% in Breed_B when the LD threshold was set to 0.6. The greatest reduction of pseudo-SNPs in populations with less genetic diversity was due to the low frequency of the haplotypes in this research, which agrees with the literature [e.g., based on simulated data (Villumsen et al., 2009); in dairy cattle populations (Hess et al., 2017; Karimi et al., 2018); and in dairy goats (Teissier et al., 2020)].

The additional computing time needed for genotype phasing, creating the haplotype blocks and the covariates for the models (Feitosa et al., 2019; Teissier et al., 2020), and running the genomic predictions (Cuyabano et al., 2015; Hess et al., 2017) have been indicated as the main drawbacks for the use of haplotypes in routine genomic predictions. In this study, the maximum additional computing time observed was approximately 7 h (23,663.6 s, Breed_B with LD equal to 0.1 under the LH2 scenario—Figure 4A and Supplementary File 6). Hess et al. (2017) used marker effect models under Bayesian approaches and observed additional time of up to 27.2 h for predictions with haplotypes derived from 37K SNPs with training and validation populations of about 30,000 dairy cattle individuals. Cuyabano et al. (2015) reported that genomic predictions using Bayesian approaches and haplotypes took approximately from 1 h to 46 h, depending on the number of previously associated SNPs included in the GEBV predictions (1K to 50K, respectively), with approximately 4,000 individuals in the training and validation populations. Differently from these studies, we used the ssGBLUP method, which showed consistent time for the predictions in the 50K SNP panel or when fitting haplotypes (as pseudo-SNPs) in the same \mathbf{G} matrix. This was likely observed because the GEBVs are estimated directly based on the \mathbf{G} matrix and the number of pseudo-SNPs added to the non-blocked SNPs (Figures 3 and 4 and Supplementary Files 5 and 6) was not large enough to require longer time to create the genomic relationship matrices. As we calculated GEBVs for more than 62,000 individuals (genotyped and non-genotyped) using haplotype information with a relatively low increase of time, ssGBLUP is a feasible alternative for that purpose.

Interestingly, our results suggest that the computing time to obtain pseudo-SNPs in less genetically diverse populations is higher than in more diverse populations. This could be because more diverse populations have a smaller number of intervals with a determined LD level than populations with low genetic diversity, implying in less iterations for the algorithm to create the haplotype blocks. The smaller number of candidate intervals to create the blocks, leading to a lower computing time, might also explain the differences observed when comparing the LD levels within populations, with the computing time being significantly greater with an LD threshold of 0.1, followed by 0.3 and 0.6 LD thresholds.

4.3 Accuracy and bias of genomic predictions

Genomic predictions based on whole genome sequence (WGS) data could be more advantageous because all the causal mutations are expected to be included in the data. However, practical results have shown no increase in GEBV accuracy when using WGS over HD (Binsbergen et al., 2015; Ni et al, 2017) or even medium density (~50K) SNP panels (Frischknecht et al., 2018). HD SNP panels were developed to better capture the LD between SNPs and QTLs and thus improve the ability to detect QTLs and obtain more accurate GEBVs (Kijas et al., 2014), especially in more genetically diverse populations or even across-breed genomic predictions. However, the 50K SNP panel has shown a similar predictive ability to the HD even in highly diverse populations as in sheep (Moghaddar et al., 2017). These findings corroborate with our results using the 50K SNP panel, regardless of the trait heritability. This suggests that both SNP panels (i.e., 50K and 600K) are sufficient to capture the genetic relationships of the individuals, which is the base of the genomic predictions based on the ssGBLUP method (Legarra et al., 2009; Aguilar et al., 2010; Lourenco et al., 2020). Therefore, we used the 50K SNP panel for haplotype-based genomic predictions.

Genomic predictions are expected to be more accurate with haplotypes instead of individual SNPs mainly because they are expected to be in greater LD with the QTL than are individual markers (Calus et al., 2008; Villumsen et al., 2009; Cuyabano et al., 2014; 2015; Hess et al., 2017). In this context, Calus et al. (2008) and Villumsen et al. (2009) reported better results for the haplotype-based predictions of GEBVs than individual SNPs in simulated data, highlighting the possibility of improving both the accuracy and bias of genomic predictions. The N_e of the populations used by Calus et al. (2008) and

Villumsen et al. (2009) is similar to the one in Breed_B (~100). However, in this current study, haplotype-based models provided similar or lower accuracy and they were also similar or more biased than individual SNP-based models under both MH2 or LH2 scenarios (Figure 5 and Supplementary Files 7 and 9). This might be related to the LD level between SNP-QTL and haplotype-QTL and also the amount of information used to estimate the SNP and haplotype effects. Calus et al. (2008) and Villumsen et al. (2009) had fewer individuals (~1,000), and their simulations were done with more general parameters compared to our study. The training set in this research for all populations was composed by 60,000 individuals with phenotypes, in which 8,000 of them were also genotyped. This amount of data is likely enough to estimate SNP effects and also the SNP-QTL LD properly. Thus, predictions with SNPs and haplotypes did not differ in some cases due to both of them capturing well the genetic relationships to achieve similar prediction results.

The correlations between off-diagonal, diagonal, and all elements in \mathbf{A}_{22} and \mathbf{G} created with pseudo-SNPs and independent SNPs together were similar to fit only individual SNPs in both SNP panel densities for all LD thresholds and in all populations, regardless of the heritability (Supplementary Files 8 and 10). Furthermore, the average, maximum, and minimum values of the diagonal elements in \mathbf{G} created when combining pseudo-SNPs and independent SNPs were also similar to using only individual SNPs for both SNP panel densities in all scenarios investigated. Therefore, combining haplotypes and SNPs in a single \mathbf{G} matrix captured the same information as fitting only individual SNPs, and, consequently, resulting in similar GEBV predictions.

Another reason for the similar genomic predictions when fitting individual SNPs and haplotypes might be the absence of or negligible epistatic interaction effects between SNP loci within haplotype blocks. In humans, a species with high N_e (Park et al., 2011), Liang et al. (2020) showed that epistasis was the reason for increased accuracy with haplotypes over individual SNPs for health traits. In other words, a similar accuracy between SNPs and haplotypes was observed when there was negligible epistasis effect. The same authors also pointed out that predictions using haplotypes might only be worse than fitting individual SNPs because of a possible “haplotype loss,” which can happen when SNP effects are not accurately estimated by the haplotypes. As no epistatic effects are currently simulated by QMSim (Sargolzaei and Schenkel, 2009) and, therefore, were

not simulated in the current study, different from our assumption that haplotypes could improve the predictions in more genetically diverse populations (Breed_C, Breed_E, Comp_2 and Comp_3), the accuracy and bias estimated based on haplotypes were similar or worse compared to fitting individual SNPs.

Many studies based on real datasets have shown small improvements in the performance of haplotype-based genomic predictions. For instance, Cuyabano et al. (2014) showed up to a 3.1% increase in the accuracy for milk protein when using LD-based haplotypes. Cuyabano et al. (2015) also obtained gains in accuracy of up to 1.3% using pre-selected SNPs associated with the trait combined with the haplotypes as covariates in the models for production, fertility, and health traits. Mucha et al. (2019) showed no differences in predictions with high-frequency haplotypes compared to SNPs when evaluating reproductive performance traits and somatic cell score in Polish dairy cattle. Additionally, Feitosa et al. (2019) obtained nearly the same accuracy and bias for meat fatty acid (MFA) traits in Nelore cattle when fitting individual SNPs or haplotypes. These findings indicate that, even in instances where haplotypes are better than SNPs, the improvements are negligible or small. However, considerable improvements in haplotype-based predictions have also been reported in the literature for relatively less polygenic traits with known major genes or when using biological information to construct the haplotype blocks. Won et al. (2020) reported a significant increase of 4.6% in GEBV accuracy with LD-clustering-based haplotypes for eye muscle area in Korean cattle. In Simmental cattle, Xu et al. (2020) reported increases of 9.8% in carcass weight when incorporating haplotype information based on SNPs from functionally related genomic regions. Teissier et al. (2020) reported an increase in accuracy of up to 22% when using haplotypes from fixed length or LD blocking strategies under an ssGBLUP setting. Based on these literature reports in livestock, it seems that haplotype predictions could provide better results when traits are oligogenic or affected by major genes, which are less common in livestock breeding goals. In addition, the presence of epistatic interactions in a real situation can also provide better results (Liang et al., 2020). In this sense, using biological information to create the blocks of linked markers to make haplotype predictions can be an alternative to improve the genomic predictions in genetically diverse livestock populations. Unfortunately, there are limited real datasets of enough size with both phenotypes and genotypes for populations with large N_e that could be used for validating our findings.

It is worth mentioning that haplotype-based models without including the independent SNPs (markers not assigned to any block) to create the genomic relationships always provided the worst results, regardless of the LD threshold to create the haploblocks (0.1, 0.3, and 0.6). These models were also less accurate and more biased in all the populations, regardless of the genetic diversity level and heritability (Figure 5 and Supplementary Files 7 and 9). The worst results were obtained when fitting only pseudo-SNPs from blocks with an LD threshold of 0.3 (PSLD03) and in more genetically diverse populations (Breed_C, Breed_E, Comp_2, and Comp_3). This might have occurred because fitting only pseudo-SNPs from the haploblocks with two or more SNPs is not enough to consider all the important chromosomal regions influencing the trait of interest. The number of blocks, blocked SNPs, and pseudo-SNPs that were used to make the predictions were significantly lower with the LD level of 0.3 compared to 0.1 in both simulations (Figure 3 and 4 and Supplementary Files 5 and 6), with this being likely the reason for the lowest accuracy and largest bias observed for PS_LD03. In this context, increasing the LD threshold to create the haploblocks have hampered the prediction with only haplotypes because a larger number of genomic markers were not considered to make the predictions. However, increasing the LD threshold to create the blocks and using the non-clustered SNPs together with the pseudo-SNPs did not affect the prediction results, presenting similar GEBV accuracies and bias compared to SNP-based predictions. In addition, the main differences in the properties of the \mathbf{G} matrix were observed when only pseudo-SNPs from haploblocks with bigger LD thresholds were used, with lower correlations between off-diagonal and all elements in the \mathbf{A}_{22} and \mathbf{G} matrices and differences in the maximum and minimum values of the diagonal elements of the \mathbf{G} (Supplementary Files 8 and 10). Therefore, independently of the LD threshold used to create the haploblocks, we recommend using the non-clustered SNPs with pseudo-SNPs from multi-marker haploblocks to make haplotype-based predictions, as well as in genome-wide association studies (GWAS) using haplotypes, because these variants may play an important role.

Separating the independent and pseudo-SNPs in two different random effects, with no shared covariances structures, did not significantly impact the genomic predictions, but had a computational cost. The genetic parameter estimation and GEBV prediction required more computing time using these two genetic components in the model, with more iterations and greater time in each iteration than the other models (data

not shown), sometimes leading to no convergence of the solutions (IPS_2H_LD03 in the Breed_C, Comp_2, and Comp_3 under MH2). The model with pseudo-SNPs and independent SNPs in two genetic components is more complex, and the convergence difficulty might suggest poor model parametrization, potentially because the random effects were assumed to be uncorrelated. This fact can be confirmed by high correlations (above than 0.90) between the inverted **H** matrices with non-clustered SNPs and pseudo-SNPs (data not shown). Although increased computational time was a common problem in both heritability levels, convergence was achieved in all analyses with low heritability. Our findings suggest that a single **G** matrix with individual SNPs is enough to capture the QTL variation, regardless of the genetic diversity and heritability. Nonetheless, using two uncorrelated genetic components can be useful in other situations such as fitting SNPs and structural variants (e.g., copy number variation—CNVs) in the same model.

5 Conclusions

Haplotype-based models did not improve the performance of genomic prediction of breeding values in genetically diverse populations (assumed as $N_e > 150$) under ssGBLUP settings. A medium-density 50K SNP panel provided similar results to the high-density panel for the genomic predictions using individual SNPs or haplotypes, regardless of the heritability and genetic diversity levels. ssGBLUP can be used to predict breeding values for both genotyped and non-genotyped individuals using haplotype information in large datasets with no increase in computing time when fitting a single genomic relationship matrix.

6 Conflict of Interest Statement

The authors declare no conflict of interest.

7 Data availability

The simulated datasets used and the pipelines developed to carry out this research are available upon request.

8 Author Contributions Statement

AA, PC, HO, and LB: conception of the work. AA: data simulation and data analyses. AA, PC, HO, and LB: interpretation of the results. AA, HO, and LB: drafted

the manuscript. AA, PC, HO, RV, FS, DALL, and LB: critical revision of the manuscript. AA, PC, HO, RV, FS, DALL, and LB: final approval of the version to be published.

9 Funding

This study was funded by Purdue University (West Lafayette, IN, USA), State University of Southwestern Bahia (Itapetinga, BA, Brazil), and the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil* (CAPES) Award Number 001.

10 Acknowledgements

We acknowledge the Dr. Brito's Lab at Purdue University for providing the scientific support to develop this research and researchers from Purdue University and State University of Southwestern Bahia for providing training to the first author and the infrastructure and resources needed for the research. We also acknowledge the National Development Council Scientific Technological (*Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq*) for the fellowship.

11 Contribution to the Field Statement

Genetic diversity is mainly measured by the effective population size (N_e) and is inversely proportional to the accuracies and bias of breeding values in genomic evaluations. Therefore, less accurate and more biased predictions are expected in genetically diverse populations ($N_e > 150$, such as sheep, goats, and some beef cattle populations), requiring larger training sets to obtain accurate estimates. Genomic selection has been also implemented in genetically diverse populations following the increase in the use of genomic information in livestock, but there is still a need for better strategies to improve the breeding value predictions in these populations. Improvements in the accuracies of genomic predictions have been reported when using haplotype-based models over individual SNPs, mainly because they better account for the linkage disequilibrium between QTLs and haplotypes. However, these results were obtained predominantly in less diverse populations (e.g., dairy cattle, $N_e < 150$). In this research we presented a comprehensive investigation regarding the use of SNPs and/or haplotypes for genomic prediction under the single-step genomic BLUP approach. For that, we simulated pure and composite populations with several levels of genetic diversity. An

extended literature review and recommendations about further steps in haplotype predictions in real populations are also presented.

References

- Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743. doi: <https://doi.org/10.3168/jds.2009-2730>
- AnimalQTLdb. (2019). QTL data base for sheep by number of chromosome. <https://www.animalgenome.org/cgi-bin/QTLdb/OA/summary?summ=chro&qtl=2,325&pub=158&trait=251> [Accessed April 15, 2020].
- Biegelmeyer, P., Gulias-Gomes, C. C., Caetano, A. R., Steibel, J. P., and Cardoso, F. F. (2016). Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. *BMC Genet.* 17:32. doi: 10.1186/s12863-016-0339-8
- Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., Eeuwijk, F. A., Schrooten, C., and Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 47:71. doi: <https://doi.org/10.1186/s12711-015-0149-x>
- Bohmanova, J., Sargolzaei, M., and Schenkel, F. S. (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics.* 11:421. doi: <https://doi.org/10.1186/1471-2164-11-421>
- Brito, L. F., Jafarikia, M., Grossi, D. A., Kijas, J. W., Porto-Neto, L. R., Ventura, R. V., et al. (2015). Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. *BMC Genet.* 16:67. doi:10.1186/s12863-015-0220-1.
- Brito, L. F., Clarke, S. M., Mcewan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., et al. (2017a). Prediction of genomic breeding values for growth, carcass and meat

- quality traits in a multi-breed sheep population using a HD SNP chip. *BMC Genet.* 18: 7-24. doi: <https://doi.org/10.1186/s12863-017-0476-8>
- Brito, L. F., Mcewan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., et al. (2017b). Genetic diversity of a New Zealand multi-breed sheep population and composite breeds' history revealed by a high-density SNP chip. *BMC Genet.* 18:25. doi: [10.1186/s12863-017-0492-8](https://doi.org/10.1186/s12863-017-0492-8)
- Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., and Veerkamp, R. F. (2008). Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics.* 178:553. Doi: [10.1534/genetics.107.080838](https://doi.org/10.1534/genetics.107.080838)
- Cuyabano, B. C. D., Su, G. S., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics.* 15:1171. doi: <https://doi.org/10.1186/1471-2164-15-1171>
- Cuyabano, B. C. D., Su, G. S., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47:61. doi: <https://doi.org/10.1186/s12711-015-0143-3>
- Daetwyler, H. D., Kemper, K. E., Van Der Werf, J. H. J., and Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *J. Anim. Sci.* 90:3375. doi: <https://doi.org/10.2527/jas.2011-4557>
- Deniskova, T., Dotsev, A., Lushihina, E., Shakhin, A., Kunz, E., Medugorac, I., et al. (2016). Population structure and genetic diversity of sheep breeds in the Kyrgyzstan. *Front. Genet.* 10:1. doi: <https://doi.org/10.3389/fgene.2019.01311>
- Falconer, D. S., and Mackay T. F. C. (1996). *Introduction to Quantitative Genetics*, Ed. 4. Longman, Essex, UK.
- FarmIQ. Release of a high-density SNP genotyping chip for the sheep genome. (2013). <http://www.farmiq.co.nz/whatsnew/news/release-high-densitysnp-genotyping-chip-sheep-genome> [Access June 6, 2020].
- Feitosa, F. L. B., Pereira, A. S. C., Amorim, S. T., Peripolli, E., Silva, R. M. O., Braz, C. U., et al. (2019). Comparison between haplotype-based and individual SNP-based

- genomic predictions for beef fatty acid profile in Nellore cattle. *J. Anim. Breed. Genet.* 00:1. doi: 10.1111/jbg.12463
- Frischknecht, M., Meuwissen, T. H. E., Bapst, B., Seefried, F. R., Flury, C., Garrick, D., et al. (2018). Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle. *J. Dairy. Sci.* 101:1292. doi: <https://doi.org/10.3168/jds.2017-12890>
- Gabriel, S. B, Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. *Sci.* 296:2225. doi: 10.1126/science.1069424
- Guarini, A. R., Lourenco, D. A. L., Brito, L. F., Sargolzaei, M., Baes, C. F., Miglior, F., et al. (2018). Comparison of genomic predictions for lowly heritable traits using multi-step and single-step genomic best linear unbiased predictor in Holstein cattle. *J. Dairy Sci.* 101:8076. doi: <https://doi.org/10.3168/jds.2017-14193>
- Guarini, A. R., Lourenco, D. A. L., Brito, L. F., Sargolzaei, M., Baes, C. F., Miglior, F., et al. (2019). Genetics and genomics of reproductive disorders in Canadian Holstein cattle. *J. Dairy Sci.* 102:1341. doi: <https://doi.org/10.3168/jds.2018-15038>
- Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13:635. doi: [10.1101/gr.387103](https://doi.org/10.1101/gr.387103)
- Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* 49:54. doi: [10.1186/s12711-017-0329-y](https://doi.org/10.1186/s12711-017-0329-y)
- Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226. doi: <https://doi.org/10.1007/BF01245622>
- Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3.* 8:1687. doi: <https://doi.org/10.1534/g3.117.300548>

- Karimi, Z., Sargolzaei, M., Robinson, J. A. B., and Schenkel, F. S. (2018). Assessing haplotype-based models for genomic evaluation in Holstein cattle. *Can. J. Anim. Sci.* 98:750. doi: <https://doi.org/10.1139/cjas-2018-0009>
- Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto-Neto, L. R., Cristobal, M. S., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 2:e1001258. doi: <https://doi.org/10.1371/journal.pbio.1001258>
- Kijas, J. W., Porto-Neto, L., Dominik, S., Reverter, A., Bunch, R., McCulloch, R., et al. (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Anim. Genet.* 45:754. doi: [10.1111/age.12197](https://doi.org/10.1111/age.12197)
- Kim, S. A., Cho, C. S., Kim, S. R., Bull, S. B., and Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics.* 34:388. doi: [10.1093/bioinformatics/btx609](https://doi.org/10.1093/bioinformatics/btx609)
- Kim, S. A., Brossard, M., Roshandel, D., Paterson, A. D., Bull, S. B., and Yoo, Y. J. (2019). gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics.* 35:4419. doi: <https://doi.org/10.1093/bioinformatics/btz308>
- Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656. doi: <https://doi.org/10.3168/jds.2009-2061>
- Legarra, A., Christensen, O. F., Aguilar, I., Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livest. Sci.* 166:54. doi: <https://doi.org/10.1016/j.livsci.2014.04.029>
- Lenth, R. V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.4. <https://CRAN.R-project.org/package=emmeans>
- Liang, Z., Tan, C., Prakapenka, D., Ma, L., and Da, Y. (2020). Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. *Front. Genet.* 11:1. doi: [doi: 10.3389/fgene.2020.588907](https://doi.org/10.3389/fgene.2020.588907)

- Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., and Misztal, I. (2020). Single-step genomic evaluations from theory to practice: Using SNP chips and sequence data in BLUPF90. *Genes* 11:790. doi: <https://doi.org/10.3390/genes11070790>
- McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet.* 5: e1000686. doi:
- McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, and D. R., Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science.* 104:581. doi: 10.1126/science.1092500
- Makanjuola, B. O., Miglior, F., Abdalla, E. A., Maltecca, C., Schenkel, F. S., and Baes, C. F. (2020). Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. *J. Dairy Sci.* 103:5183. doi: 10.3168/jds.2019-18013
- Meuwissen, T., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 157, 1819-1829.
- Meuwissen, T., Ødegård, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* 46:49. <https://doi.org/10.1186/1297-9686-46-49>
- Misztal, I., Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Aguilar, I., Legarra, A., et al. (2018). Manual for BLUPF90 family programs. University of Georgia. <http://nce.ads.uga.edu/wiki/doku.php?id=documentation>
- Moghaddar, N., Swan, A. A., and Van der Werf, J. H. J. (2017). Genomic prediction from observed and imputed high-density ovine genotypes. *Genet. Sel. Evol.* 49:40. doi: <https://doi.org/10.1186/s12711-017-0315-4>
- Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* 38:2074. doi: [10.1002/sim.8086](https://doi.org/10.1002/sim.8086)
- Moreira, F. F., Oliveira, H. R., Volenec, J. J., Rainey K. M., and Brito, L. F. (2020). Integrating high-throughput phenotyping and statistical genomic methods to

- genetically improve longitudinal traits in crops. *Front. Genet.* 11:681. doi: <https://doi.org/10.3389/fpls.2020.00681>
- Mucha, A., Wierzbicki, H., Kamiński, S., Oleński, K., and Hering, H. (2019). High-frequency marker haplotypes in the genomic selection of dairy cattle. *J. Appl. Genet.* 60:179. doi: <https://doi.org/10.1007/s13353-019-00489-9>
- Ni, G., Caverro, D., Fangmann, A., Erbe, M., and Simianer, H. (2017). Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* 49:8. doi: <https://doi.org/10.1186/s12711-016-0277-y>
- Nicolazzi, E. L., Caprera, A., Nazzicari, N., Cozzi, P., Strozzi, F., Lawley, C., et al. (2015). SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genomics.* 16:283. doi: <https://doi.org/10.1186/s12864-015-1497-1>
- Oliveira, H. O., Brito, L. F., Sargolzaei, M., Silva, F. F., Jamrozik, J., Lourenco, D. A. L., et al. (2019). Impact of including information from bulls and their daughters in the training population of multiple-step genomic evaluations in dairy cattle: A simulation study. *J. Anim. Breed. Genet.* 136:441. doi: <https://doi.org/10.1111/jbg.12407>
- Oliveira, H. R., McEwan, J. C., Jakobsen, J., Blichfeldt, T., Meuwissen, T., Pickering, N., et al. (2020). Genetic connectedness between Norwegian White Sheep and New Zealand Composite Sheep populations with similar development history. *Front. Genet.* 11:371. doi: <https://doi.org/10.3389/fgene.2020.00371>
- Park, L. (2011). Effective population size of current human population. *Genet. Res.* 93:105. doi: [doi:10.1017/S0016672310000558](https://doi.org/10.1017/S0016672310000558)
- Piccoli, M. L., Brito, L. F., Braccinia, J., Oliveira, H. R., Cardoso, F. F., Roso, V. M., et al. (2020). Comparison of genomic prediction methods for evaluation of adaptation and productive efficiency traits in Braford and Hereford cattle. *Livest. Sci.* 230: 103864. doi: <https://doi.org/10.1016/j.livsci.2019.103864>

- Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. R Core Team (2021). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-152, <https://CRAN.R-project.org/package=nlme>.
- Prieur, V., Clarke, S. M., Brito, L. F., McEwan, J. C., Lee, M. A., Brauning, R., et al . (2017). Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. *BMC Genetics*. 18:68. doi: [10.1186/s12863-017-0534-2](https://doi.org/10.1186/s12863-017-0534-2)
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya: Indian J. Stat.* 9:1. doi <https://www.jstor.org/stable/25049339>
- Rasali, D. P., Shrestha, J. N. B., and Crow, G. H. (2006). Development of composite sheep breeds in the world: A review. *Can. J. Anim. Sci.* 86,1-24.
- Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*. 15:478. doi: [10.1186/1471-2164-15-478](https://doi.org/10.1186/1471-2164-15-478).
- Sargolzaei, M., and Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for Livestock. *Bioinformatics*, 25:680. doi: [10.1093/bioinformatics/btp045](https://doi.org/10.1093/bioinformatics/btp045)
- Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51:492. doi: [10.1080/10635150290069913](https://doi.org/10.1080/10635150290069913)
- Stachowicz, K., Brito, L. F., Oliveira, H. R., Miller, S. P., and Schenkel, F. S. (2018). Assessing genetic diversity of various Canadian sheep breeds through pedigree analyses. *Can. J. Anim. Sci.* 98: 741. doi: [dx.doi.org/10.1139/cjas-2017-0187](https://doi.org/10.1139/cjas-2017-0187)
- Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2:125. doi: [10.1016/0040-5809\(71\)90011-6](https://doi.org/10.1016/0040-5809(71)90011-6)
- Teissier, M., Larroque, H., Brito, L. F., Rupp, R., Schenkel, F. S., and Robert-Granié, C. (2020). Genomic predictions based on haplotypes fitted as pseudo-SNPs for milk

production and udder type traits and somatic cell score in French dairy goats. *J. Dairy Sci.* 103:11559. doi: [10.3168/jds.2020-18662](https://doi.org/10.3168/jds.2020-18662)

Vanraden, P. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414. doi: <https://doi.org/10.3168/jds.2007-0980>

Villumsen, T. M., Janss, L., and Lund, M. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126:3. doi: [10.1111/j.1439-0388.2008.00747.x](https://doi.org/10.1111/j.1439-0388.2008.00747.x)

Xu, L., Ga, N., Wang, Z., Xu, L., Li, Y., Chen, Y., et al. (2020). Incorporating genome annotation into genomic prediction for carcass traits in Chinese Simmental beef cattle. *Front. Genet.* 11:481. doi: <https://doi.org/10.3389/fgene.2020.00481>

Won, S., Park, J., Son, J., Lee, S., Park, B. H., Park, M., et al. (2020). Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Front. Genet.* 11:134. doi: [10.3389/fgene.2020.00134](https://doi.org/10.3389/fgene.2020.00134)

Supporting information

Supplementary File 1. Length, number of markers, and quantitative trait loci (QTL) by chromosome in the simulated sheep populations¹. ¹The simulation was done considering only the autosomal chromosomes. ²The size of the autosomal chromosomes were obtained from the reference ovine genome (assembly Oar_v4.0) on the NCBI platform (www.ncbi.nlm.nih.gov/genome?term=ovis%20aries). ³Number of SNPs simulated in the autosomal chromosomes in the 600K SNP panel according to the SNPchiMp v.3 platform (Nicolazzi et al., 2015). ⁴SNP panel randomly sampled from the simulated 600K panel according to the SNPchiMp v.3 platform (Nicolazzi et al., 2015). ⁵The number of QTLs described in the literature for autosomal chromosomes of sheep in the AnimalQTLdb (AnimalQTLdb, 2019), published on December 2019.

Supplementary File 2. Genetic structure of the simulated populations used in the genomic predictions based on SNPs and/or haplotypes.

Supplementary File 3. Values for each replicate, mean and its standard error for the effective population size and the variance components in all studied populations simulated for a trait with moderate heritability ($h^2 = 0.30$).

Supplementary File 4. Values for each replicate, mean and its standard error for the effective population size and the variance components in all studied populations simulated for a trait with low heritability ($h^2 = 0.10$).

Supplementary File 5. Values for each replicate, mean and its standard error for the statistics related to haplotype blocking in all studied populations simulated for a trait with moderate heritability ($h^2 = 0.30$).

Supplementary File 6. Values for each replicate, mean and its standard error for the statistics related to haplotype blocking in all studied populations simulated for a trait with low heritability ($h^2 = 0.10$).

Supplementary File 7. Values for each replicate, average and its standard error for the accuracies and bias of predictions in all studied populations simulated for a trait with moderate heritability ($h^2 = 0.30$).

Supplementary File 8. Properties of the genomic relationship matrix when fitting SNPs or haplotypes in all studied populations simulated for a trait with moderate heritability ($h^2 = 0.30$).

Supplementary File 9. Values for each replicate, average and its standard error for the accuracies and bias of predictions in all studied populations simulated for a trait with low heritability ($h^2 = 0.10$).

Supplementary File 10. Properties of the genomic relationship matrix using SNPs or haplotypes in all studied populations simulated for a trait with low heritability ($h^2 = 0.10$).

IV – CAPÍTULO II

Artigo publicado na revista *Genes*Doi: <https://doi.org/10.3390/genes13010017>

Citation: Araujo, A.C.; Carneiro, P.L.S.; Alvarenga, A.B.; Oliveira, H.R.; Miller, S.P.; Retallick, K.; Brito, L.F. Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle. *Genes* 2022, 13(1),17; <https://doi.org/10.3390/genes13010017>

Academic Editor(s): Samantha A. Brooks and Carissa Wickens

Received: 18 November 2021

Accepted: 18 December 2021

Published: 22 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Article

Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle

Andre C. Araujo ^{1,2}, Paulo L. S. Carneiro ³, Amanda B. Alvarenga ², Hinayah R. Oliveira ^{2,4}, Stephen P. Miller ⁵, Kelli Retallick ⁵ and Luiz F. Brito ^{2,*}

¹ Graduate Program in Animal Sciences, State University of Southwestern Bahia, Itapetinga 45700-000, Brazil; araujoa@purdue.edu

² Department of Animal Science, Purdue University, West Lafayette, IN 47907, USA; alvarena@purdue.edu (A.B.A.); hinayah@gmail.com (H.R.O.)

³ Department of Biology, State University of Southwest Bahia, Jequié 45205-490, Brazil; plscarneiro@uesb.edu.br

⁴ Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON N1G2W1, Canada

⁵ American Angus Association, Angus Genetics Inc., 3201 Frederick Ave, St Joseph, MO 64506, USA; SMiller@angus.org (S.P.M.); kretallick@angus.org (K.R.)

* Correspondence: britol@purdue.edu

Abstract: Behavior is a complex trait and, therefore, understanding its genetic architecture is paramount for the development of effective breeding strategies. The objective of this study was to perform traditional and weighted single-step genome-wide association studies (ssGWAS and WssGWAS, respectively) for yearling temperament (YT) in North American Angus cattle using haplotypes. Approximately 266 K YT records and 70 K animals genotyped using a 50 K single nucleotide polymorphisms (SNP) panel were used. Linkage disequilibrium thresholds (LD) of 0.15, 0.50, and 0.80 were used to create the haploblocks, and the inclusion of non-LD-clustered SNPs (NCSNP) with the haplotypes in the genomic models was also evaluated. WssGWAS did not perform better than ssGWAS. Cattle YT was found to be a highly polygenic trait, with genes and quantitative

trait loci (QTL) broadly distributed across the whole genome. Association studies using LD-based haplotypes should include NCSNPs and different LD thresholds to increase the likelihood of finding the relevant genomic regions affecting the trait of interest. The main candidate genes identified, i.e., *ATXN10*, *ADAM10*, *VAX2*, *ATP6V1B1*, *CRISPLD1*, *CAPRIN1*, *FA2H*, *SPEF2*, *PLXNA1*, and *CACNA2D3*, are involved in important biological processes and metabolic pathways related to behavioral traits, social interactions, and aggressiveness in cattle. Future studies should further investigate the role of these candidate genes.

Keywords: candidate genes; functional analysis; haplotype block; linkage disequilibrium; livestock behavior; pseudo-SNPs; social interaction

1. Introduction

Behavior is a complex trait influenced by multiple factors (e.g., age, health status, life experiences, genetics) and the interaction among group-housed individuals and the environment [1]. Emotional or behavioral responses are actions resultant of feedback from the central nervous system after decodifying an external stimulus, which has been studied for a long time in humans [2]. Prior to domestication, animals presented different behavior characteristics compared to domesticated populations, indicating that behavioral traits can be genetically modified through selective breeding [1]. Livestock behavior is important due to its impact in several other relevant traits for the industry, including production, reproduction, and both animal and handler's welfare and health [3–5]. Docile temperament is a desired behavior in cattle because it facilitates the handling process and it has been proven to be favorably associated with meat quality, productive efficiency, and welfare traits [6]. An indicator of temperament used for selection in North American Angus cattle is yearling temperament (YT). YT is subjectively scored by farmers/handlers when a one-year-old calf is being processed through the chute and should be an observation of how animals enter, exit, and react while being handled [7]. A previous study has shown that YT is heritable (heritability ~ 0.38), suggesting genetic progress can be achieved through direct selection [5]. Additionally, a multi-species systematic review reported 797 genomic regions and 383 candidate genes associated with behavioral traits in cattle [8]. Only six genes (*GRM5*, *MAML3*, *C8B*, *RUSC2*, *POMC*, *MIPOL1*, and *SLC18A2*) were in overlap among trait definitions and

populations [8], suggesting a natural particularity of each population and measurement definition.

Using alternative approaches, such as haplotype-based methods, for detecting genomic regions influencing YT is of great interest to the beef cattle industry. Haplotypes are usually defined as a set of adjacent loci expected to be inherited together with a small probability of recombination [9]. Haplotype blocks (i.e., haploblocks) are sets of adjacent single nucleotide polymorphisms (SNPs) markers expected to be in higher linkage disequilibrium (LD) with the quantitative trait loci (QTL) than single SNPs [10,11]. Furthermore, haplotypes can also capture small epistatic effects within haploblocks [12–14], justifying the advantages of using haplotypes for both genomic prediction of breeding values [12–14] and genome-wide association studies (GWAS) [14–16]. However, due to the more complex implementation of haplotype-based methods, they have been underused compared to SNP-based methods [17,18]. For GWAS purposes, the combination of SNP- and haplotype-based methods are recommended because it might increase the possibility of capturing different types of QTL [15,16], i.e., different sizes (spanning small or large genomic regions), allelic frequency, and LD levels with SNPs due to differential recombination rates.

Both phenotypes and genotypes are required when performing GWAS; however, not all phenotyped individuals in a population are genotyped and vice-versa [19,20]. In this context, Single-step Genomic Best Linear Unbiased Prediction (ssGBLUP) is a method that simultaneously combines information from phenotypes, pedigree, and genotypes when calculating genomic estimated breeding values (GEBV) for both genotyped and non-genotyped individuals [21,22]. Therefore, single-step GWAS (ssGWAS), which uses GEBV from ssGBLUP to derive SNP effects, is an efficient method to perform GWAS because phenotypes from genotyped and ungenotyped individuals are used to more accurately derive SNP effects [20]. However, the infinitesimal model (many loci explaining similar and small proportions of the total additive genetic variance) is an assumption of both ssGBLUP and ssGWAS [19]. As some loci (major genes) can explain major proportions of the additive variance of the traits of interest, weighted ssGBLUP (WssGBLUP) and its GWAS version (WssGWAS) were proposed to prioritize markers potentially explaining larger proportions of the total additive genetic variance [19,23].

Despite its relevance in animal breeding, to the best of our knowledge, no haplotype-based GWAS has been used to investigate the genetic background of temperament in cattle. Additionally, there is a lack of studies implementing haplotypes under ssGWAS and WssGWAS approaches. Therefore, the objective of this study was to perform a haplotype-based ssGWAS for YT in American Angus cattle. Different haplotype-based GWAS approaches were implemented and tested to uncover the potential genomic regions associated with YT, including: (1) different LD thresholds to create the LD-based haplotypes, (2) including or not including the non-LD-clustered SNPs in the association analyses, and (3) ssGWAS and WssGWAS. Understanding the genetic background of behavioral traits is of great interest for the beef cattle industry because it could enable the optimization of genetic selection for more docile animals in which the genetic progress would be permanent and cumulative over generations. Genetic or genomic selection for any trait impacts several biological mechanisms involved in the phenotypic expression of the trait under selection as well as genetically correlated traits. Therefore, it is paramount to understand these underlying biological mechanisms.

2. Materials and Methods

2.1. Phenotypic and Pedigree Data

The American Angus Association (through Angus Genetics Inc.; St Joseph, MO, USA) provided the phenotypic, genotypic, and pedigree datasets. In total, 266,029 animals recorded for YT, born between 2001 and 2018, were available for the analyses. The phenotypic dataset has previously been processed for quality measurements (please see Alvarenga et al. [5] for a complete description of the data). Briefly, yearling temperament is a categorical trait recorded using six scores (from 1 to 6), in which 1 represents docile and 6 represents very aggressive animals. For more details about the codification and criteria to classify the animals, please see [5] and [7]. From the total number of records, 71.9% were classified as docile (score 1), 22.2% as restless (score 2), 5.1% as nervous (score 3), and 0.8% as aggressive (scores 4 to 6). The scores 4, 5, and 6, which represents flighty, aggressive, and very aggressive, respectively, were grouped together as a single category (aggressive) due to their low incidence [5]. The number of animals per management class in the phenotypic data were: 147,671 bulls, 3,332 steers, and 115,026 females. The pedigree data initially had 4,410,551 animals

born from 1836 to 2018, and 578,819 individuals remained to construct the pedigree-based additive genetic relationship matrix (**A**), tracing back ancestors up to four generations.

2.2. Genotypic Data

The genotypic dataset included 69,559 animals genotyped using a 50 K SNP panel (54,609 SNP markers) of an imputed SNP set similar to the Illumina BovineSNP50V2 and Illumina BovineSNP50V3 (Illumina, Inc., San Diego, CA, USA), designed for commercial purposes. Markers with minor allele frequency (MAF) < 0.01, call rate < 0.90, difference between observed and expected heterozygosity > 0.15 (i.e., extreme departure from Hardy-Weinberg equilibrium), and not present in the pseudo-autosomal region (PAR) in the X chromosome were removed from the genotypic data as part of the quality control (QC). PAR was considered the region above BTAX:133,300,518 bp [24]. Additionally, animals with call rates lower than 0.90 were also removed. The QC in the genotypic data was done using the PREGSf90 software from the BLUPf90 family software [25]. After QC, 42,633 markers and 69,437 animals were kept for further analyses.

2.3. Haplotype Block Construction

The haplotype block (haploblock) construction process started with phasing the SNP genotypes in the FImpute software v.3.0 [26]. After phasing, haploblocks were constructed with a variable size approach using the LD values measured by the r^2 metric [27]. The Big-LD method [28] was used to construct the blocks because it is more computationally efficient and accurate in estimating the recombination hotspots than other commonly used algorithms [28]. The “gpart” package [29] implemented in the R software [30] was used to implement the Big-LD method for constructing the haploblocks. As the QTL can have different genetic structures, the LD thresholds of 0.15, 0.50, and 0.80 were used to create the haploblocks to account for different recombination levels within regions, assumed to be high, medium, and low, respectively. In other words, these LD thresholds were used to capture different block structures: bigger blocks with more SNPs in low LD (0.15), intermediary blocks with moderate LD (0.50), and smaller blocks with lower number of SNPs in high LD (0.80). Furthermore, the haploblocks used in this research followed the definition proposed by Gabriel et al. [9], being sizable regions delimited to at least two loci (SNPs).

2.4. Single-Step GWAS with Haplotypes

The unique haplotype alleles within the haploblocks were coded as pseudo-SNPs, which were submitted to the same QC as the SNPs (described in Section 2.2) for performing the ssGWAS with haplotypes. The pseudo-SNPs were coded as 0, 1, or 2 for the absence of both copies, presence of one copy, or presence of two copies of the reference haplotype allele [31]. Thereafter, the THRGIBBS1f90 software [25] was used to predict the GEBV for all individuals considering YT as a categorical trait (threshold model) and the pseudo-SNPs to construct the genomic relationship matrix.

The contemporary groups (CG) and the animal model used to predict the YT GEBV were previously defined by Alvarenga et al. [5], i.e.:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{w} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

where \mathbf{y} is the vector of phenotypic records for YT, \mathbf{b} is the vector of systematic effects (age of dam, conception type, and calf age deviation from 365 days as linear covariate), \mathbf{w} is the random vector of CG effects with $\mathbf{w} \sim N(0, \mathbf{I}\sigma_w^2)$, \mathbf{u} is the random vector of additive genetic effects with $\mathbf{u} \sim N(0, \mathbf{H}\sigma_g^2)$, and \mathbf{e} is the random residual term with $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. CG was formed by concatenating birth month and year, birth herd, birth sex, weaning date, weaning herd, weaning sex, creep feeding offered or not, date of temperament measurement, YT measurement herd, sex at the YT measurement, temperament group age deviation, and presence of ultrasound records (measure of additional human-animal interaction). \mathbf{X} , \mathbf{W} , and \mathbf{Z} are the incidence matrices for the systematic, CG, and additive genetic effects, respectively. A diagonal matrix with large values, $\Sigma_{\mathbf{b}}$, was used to represent a vague prior for the systematic effects. The \mathbf{H} matrix is the matrix that combines the pedigree and genomic relationship matrices [21], and its inverse (\mathbf{H}^{-1}) was directly computed as [22]:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}, \quad (2)$$

where \mathbf{A}^{-1} is the inverse of the \mathbf{A} matrix, \mathbf{G} is the genomic relationship matrix computed using the pseudo-SNPs, and \mathbf{A}_{22}^{-1} is the inverse of the pedigree-based relationship matrix between genotyped individuals. The default value (1.0) was used for the scaling parameters (τ and ω), while 0.90 and 0.10 were used for the weighting parameters α and β , respectively, in the PREGSf90 package [25]. \mathbf{G} was computed as in the first method proposed by VanRaden [32], which had the following structure:

$$\mathbf{G} = \frac{(\mathbf{M} - 2\mathbf{p})\mathbf{D}(\mathbf{M} - 2\mathbf{p})'}{2 \sum p_i(1 - p_i)} \quad (3)$$

where \mathbf{M} is the $n \times m$ (number of individuals by number of markers, respectively) matrix of genotype calls (0, 1, or 2), \mathbf{p} is a vector with the allelic frequencies p_i for the markers, and \mathbf{D} is a $m \times m$ diagonal matrix that corresponds to an identity matrix (\mathbf{I}) when \mathbf{G} is applied with the same weight (i.e., 1) for the markers (i.e., ssGWAS).

The variance components (i.e., σ_w^2 , σ_g^2 , and σ_e^2) for YT in the American Angus population using the above model were previously estimated by Alvarenga et al. [5] and were fixed to predict the GEBVs in this study. The chain parameters used to make the predictions were 10,000 iterations, in which 1,000 were discarded as burn-in, and 10 was used as thin. After the GEBV prediction, the pseudo-SNP effects were back-solved using the POSTGSf90 software [25]. The formula to back-solve the pseudo-SNP effects is [33]:

$$\mathbf{g} = \mathbf{D}(\mathbf{M} - 2\mathbf{p})' \mathbf{G}^{-1} \hat{\mathbf{u}} \quad (4)$$

where \mathbf{g} is the vector of marker effects, \mathbf{G}^{-1} is the inverted \mathbf{G} matrix, $\hat{\mathbf{u}}$ is the vector of predicted GEBV, and all other matrices and vectors were described above. In addition to the marker effects, the POSTGSf90 software [25] was also used to calculate the percentage of the total additive genetic variance explained by each pseudo-SNP (haploblock allele), i.e.:

$$VEM\%_i = \frac{V(g_i)}{\sigma_g^2} \times 100 = \frac{2p_i(1 - p_i)\hat{\alpha}_i^2}{\sigma_g^2} \times 100 \quad (5)$$

where $VEM\%_i$ is the percentage of the total additive genetic variance explained by the i th pseudo-SNP, $V(g_i)$ is the additive genetic variance explained by the i th pseudo-SNP, $\hat{\alpha}_i^2$ is the square of the estimated allelic substitution effect, and the other components of the formula were previously defined. As the pseudo-SNPs were the alleles present in the haploblock loci, the percentage of the variance explained by each haploblock was computed as:

$$VEH\%_j = \sum_{i=1}^{n_j} VEM\%_{ij} \quad (6)$$

where $VEH\%_j$ is the percent of the total additive genetic variance explained by the j th haploblock, n_j is the number of haplotype alleles (pseudo-SNPs) present in the j th haploblock, and $VEM\%_{ij}$ are the variances explained by the i th pseudo-SNPs present within the j th haploblock.

2.5. *Weighted Single-Step GWAS with Haplotypes*

The WssGWAS uses, for simplicity, an interactive process to estimate the weights for the markers. The first step in the WssGWAS is to perform the ssGWAS (i.e., considering equal weights for all markers). The procedure consists of three steps, starting with the prediction of GEBV from ssGBLUP, deriving the weights for the markers by back-solving SNP effects, and including the weights into the \mathbf{D} matrix to construct the \mathbf{G} matrix that is combined with the pedigree-based relationship (resulting in the \mathbf{H}) in the next steps in an iterative process [19]. For details of the full algorithm, please see Wang et al. [19]. The POSTGSf90 software [25] was used for back-solving the GEBV to pseudo-SNP effects and weights, and the non-linear A (NLA) method [32] was used to obtain the weights. The NLA weighting method was used because it has better statistical properties (i.e., convergence of the GEBV accuracies and control over extreme weight values) than the original method proposed by Wang et al. [19], as suggested by Fragomeni et al. [34]. In addition to the first GEBV prediction and association (ssGWAS), two iterations in WssGBLUP were completed to provide high accuracy and low bias of the GEBV used in the WssGWAS [19,34], and the results of these two iterations were compared to ssGWAS. The same genetic model, \mathbf{H} matrix construction, and percentage of the variances explained by each pseudo-SNP and haploblock loci presented in the topic 2.4 were also used for the WssGWAS.

2.6. *Scenarios Evaluated*

In addition to the three initial scenarios regarding the ssGWAS method (i.e., ssGWAS, 2nd iteration WssGWAS, and 3rd iteration WssGWAS), alternative approaches were used when constructing the \mathbf{G} matrix. The scenarios included the construction of haplotypes based on: (1) the LD thresholds of 0.15 (H0.15), 0.50 (H0.50), and 0.80 (H0.80), as mentioned in the topic 2.3.; and (2) considering only haplotypes or the haplotypes and non-LD-clustered SNP (NCSNP) from those same LD thresholds together in the construction of a single \mathbf{G} matrix. The NCSNP were SNP not assigned to any block during the haploblock construction with a determined LD threshold. These scenarios with NCSNP were also evaluated to avoid a possible loss of ability in dissecting the genetic variation by losing the markers outside blocks described by Li et al. [35], and the use of a single \mathbf{G} constructed with NCSNP and haplotypes provides greater GEBV

accuracy and lower bias [18]. Therefore, six scenarios in the context of the **G** construction were investigated: H0.15, H0.50, H0.80, and NCSNP and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80 (NCSNP_H0.15, NCSNP_H0.50, and NCSNP_H0.80, respectively). All scenarios are described in Table 1.

Table 1. Scenarios used to evaluate the traditional and weighted single-step GWAS (ssGWAS and WssGWAS, respectively) using haplotypes for yearling temperament in American Angus cattle.

Method	Marker Information ¹	Scenario Abbreviation
ssGWAS	H0.15	ssGWAS_H0.15
	H0.50	ssGWAS_H0.50
	H0.80	ssGWAS_H0.80
	NCSNP_H0.15	ssGWAS_NCSNP_H0.15
	NCSNP_H0.50	ssGWAS_NCSNP_H0.50
	NCSNP_H0.80	ssGWAS_NCSNP_H0.80
	WssGWAS iteration 2 (WssGWAS_2)	H0.15
H0.50		WssGWAS_2_H0.50
H0.80		WssGWAS_2_H0.80
NCSNP_H0.15		WssGWAS_2_NCSNP_H0.15
NCSNP_H0.50		WssGWAS_2_NCSNP_H0.50
NCSNP_H0.80		WssGWAS_2_NCSNP_H0.80
WssGWAS iteration 3 (WssGWAS_3)		H0.15
	H0.50	WssGWAS_3_H0.50
	H0.80	WssGWAS_3_H0.80
	NCSNP_H0.15	WssGWAS_3_NCSNP_H0.15
	NCSNP_H0.50	WssGWAS_3_NCSNP_H0.50
	NCSNP_H0.80	WssGWAS_3_NCSNP_H0.80

¹ H0.15, H0.50, and H0.80: haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; NCSNP_H0.15, NCSNP_H0.50, and NCSNP_H0.80: non-clustered SNP and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively.

2.7. Empirical Selection of the Candidate Regions for Further Investigation

The percentage of the total additive genetic variance explained by the markers was evaluated to determine the regions to be further investigated. In this context, the moments of the distribution from the percentage of the variance

explained by the markers were estimated first, and then the skewness-kurtosis plot proposed by Cullen and Frey [36] was utilized to select candidate distributions. The skewness-kurtosis plot [36] presented values for the moments of common distributions (normal, uniform, exponential, logistic, Beta, log-normal, Gamma, and Weibull) to select the distribution that better fit the data. The R package “fitdistplus” [37] was used to evaluate the skewness-kurtosis plot using 10,000 bootstrap samples to choose candidate distributions for the percentage of the variance explained by the markers. The Beta or Gamma distributions were chosen based on the skewness-kurtosis plot to be candidate distributions (not the same distribution for all scenarios; Supplementary File 1). Thereafter, the theoretical and empirical probability density function (PDF), cumulative probability function (CDF), and QQ and PP plots for the Beta and Gamma distributions were evaluated. The Beta and Gamma distributions fit the data well and were used to determine the candidate regions to be further investigated. The markers that were further investigated explained the largest percentage of the additive variance and were present in the quantile 0.001% of the fitted distribution, i.e., considered to be the most relevant genomic regions (top 0.001%). To obtain candidate regions for YT, the quantile 0.001% for the top markers that explained most of the additive variance was empirically defined because it is an extreme tail of the distribution. Using greater thresholds, e.g., 0.01 or 0.05%, only increased the number of genes and QTL related to more general functions and biological processes (Supplementary Files 2 and 3).

2.8. Functional Analyses

The top 0.001% genomic regions for YT were used to find genes and overlapping QTL using the Biomart tool from Ensembl (www.ensembl.org/biomart/martview/ad1112a783c0e0ae22e6572189d5bead, accessed on 14 August 2021) and the Animal QTLdb [38] (www.animalgenome.org/cgi-bin/QTLdb/BT/index, accessed on 14 August 2021), respectively. These analyses were done based on the latest ARS-UCD1.2 bovine genome assembly [39,40]. Positional candidate genes overlapping with the top genomic regions were functionally annotated using the DAVID platform (<https://david.ncifcrf.gov/home.jsp>, accessed on 15 August 2021) with focus on the Gene Ontology biological processes (GO_BP) and metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) using the default options.

3. Results

3.1. Statistics from Haplotype Blocking

The number of non-clustered SNPs (i.e., SNPs out of the LD-blocks) ranged from 20,849 to 36,881 when using the LD thresholds of 0.15 and 0.80, respectively (Table 2). The number of clustered SNPs ranged from 5822 to 21,784 with LD thresholds of 0.80 and 0.15, respectively. Similar to the number of clustered SNPs, the number of blocks decreased with higher LD thresholds, varying from 2721 to 9634 when 0.80 and 0.15 were used as LD thresholds, respectively. As previously defined, the minimum number of SNPs in blocks were 2 for all LD thresholds, whereas the maximum number of SNPs in blocks was equal to 9 for the LD threshold of 0.15 and 7 for 0.50 and 0.80. On average, smaller blocks were obtained when the LD threshold was 0.80 (0.030 Mb), and bigger blocks were obtained with an LD threshold of 0.15 (0.035 Mb). The minimum block size was 65 bp with the LD thresholds of 0.15 and 0.50, and 84 bp with the LD threshold of 0.80. The maximum block size ranged from 0.160 Mb to 0.201 Mb with the LD thresholds of 0.80 and 0.15, respectively. The number of pseudo-SNPs (unique haplotype alleles) before QC varied between 12,877 and 56,734 with LD thresholds of 0.80 and 0.15, respectively. After QC, the number of pseudo-SNPs ranged from 11,389 to 44,559, respectively, in the same LD thresholds. The number of NCSNP and pseudo-SNPs before QC, considering them all as genomic markers, ranged between 49,688 and 77,583 with LD thresholds of 0.80 and 0.15, respectively. After QC, the number of NCSNP and pseudo-SNPs ranged from 48,227 to 65,435, respectively, in the same LD thresholds.

3.2. Traditional and Weighted Single-Step GWAS Fitting Only Haplotypes

The number of top 0.001% genomic regions ranged from 5 (WssGWAS_2_H0.80 and WssGWAS_3_H0.80) to 17 (ssGWAS_H0.15) when only haplotypes were used in the ssGWAS (Figure 1). Despite the number of top 0.001% genomic regions identified being slightly lower in WssGWAS compared to ssGWAS scenarios (Figure 1), all top genomic regions highlighted by WssGWAS scenarios (Figures 2–4) were present in the ssGWAS results regardless of the iteration within LD thresholds. For this reason, functional annotation was performed only for ssGWAS results (Supplementary File 2). The top haplotypes for YT were located across 14 chromosomes (BTA1, BTA2, BTA3, BTA4, BTA7, BTA9, BTA11, BTA18, BTA20, BTA22, BTA23, BTA26, BTA27, and

BTA29; Figure 2) for the ssGWAS_H0.15 scenario, which presented top genomic regions more spread out than the other scenarios using only haplotypes. A total of 11, 7, and 5 candidate genes overlapped with the top 0.001% genomic regions identified by ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80, respectively, whereas 67, 55, and 30 QTL overlapped in these same scenarios, respectively. No associations were observed in the X chromosome (PAR region) when fitting only haplotypes.

Table 2. Descriptive statistics of the haplotype blocks with different linkage disequilibrium (LD) thresholds used in each scenario, before and after quality control (QC), in American Angus cattle.

Descriptive	LD_0.15	LD_0.50	LD_0.80
Number of non-clustered SNPs	20,849	30,501	36,811
Number of clustered SNPs	21,784	12,132	5822
Number of blocks	9634	5617	2721
Minimum number of SNP in blocks	2	2	2
Maximum number of SNP in blocks	9	7	7
Average (SD ¹) block size (Mb)	0.035 (0.020)	0.032 (0.014)	0.030 (0.013)
Minimum block size (bp)	65	65	84
Maximum block size (Mb)	0.201	0.161	0.160
Number of pseudo-SNPs ² before QC	56,734	27,324	12,877
Number of pseudo-SNPs after QC	44,559	23,918	11,389
Number of non-clustered and pseudo-SNPs before QC	77,583	57,825	49,688
Number of non-clustered and pseudo-SNPs after QC	65,435	54,444	48,227

¹ Standard deviation. ² Pseudo-SNPs are the unique haplotype alleles from the combination of phased SNPs within haplotype blocks.

3.3. Traditional and Weighted Single-Step GWAS Fitting Haplotype Blocks and Non-Clustered SNP

Including the NCSNP in the association analyses resulted in more top regions being captured by the haploblocks from all LD thresholds and under both ssGWAS or WssGWAS approaches. The number of top 0.001% genomic regions ranged between 36 to 64 in the WssGWAS_3_NCSNP_H0.15 and ssGWAS_NCSNP_H0.50 scenarios, respectively (Figure 1). Similar to what was

observed when using only haplotypes, all top markers (pseudo-SNPs and NCSNP) highlighted by WssGWAS were present in the ssGWAS scenarios (Figures 5–7), which also had more top markers in all LD thresholds (Figure 1). Functional annotation was performed only for ssGWAS in the scenarios including the NCSNP for the same reason previously described (Supplementary File 3). Different from what was observed in the ssGWAS using only haplotypes, the top 0.001% genomic regions were well distributed in the bovine chromosomes, including the PAR region of the X chromosome, with all LD the three thresholds used to create the haploblocks. The additional number of top 0.001% genomic regions using haplotypes and NCSNP together also implied more annotated genes and overlapping QTL (36, 54, and 35 genes and 159, 169, and 157 QTL for the ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80, respectively; Supplementary File 3).

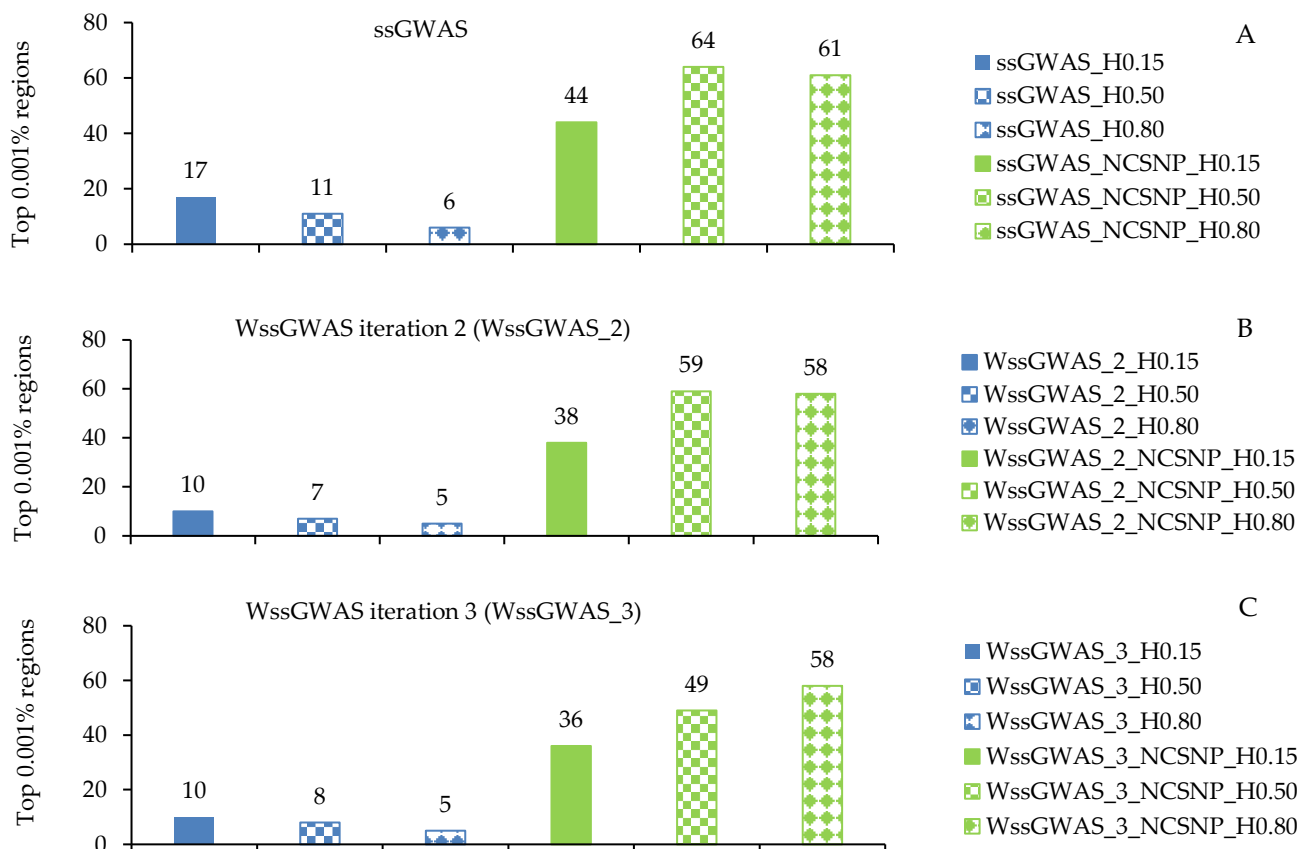


Figure 1. Number of top 0.001% genomic regions for yearling temperament in American Angus cattle found by non-weighted single-step GWAS (ssGWAS) (A) and weighted ssGWAS (WssGWAS) in the second (B) and third (C) iterations. H0.15, H0.50, and H0.80: only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; NCSNP_H0.15, NCSNP_H0.50, and NCSNP_H0.80: non-clustered SNPs (NCSNP) and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively. The column colors highlight not including (blue) or including NCSNP (green). The column filling highlights different LD thresholds (0.15, 0.50, and 0.80 with a solid, square, and diamond filling, respectively).

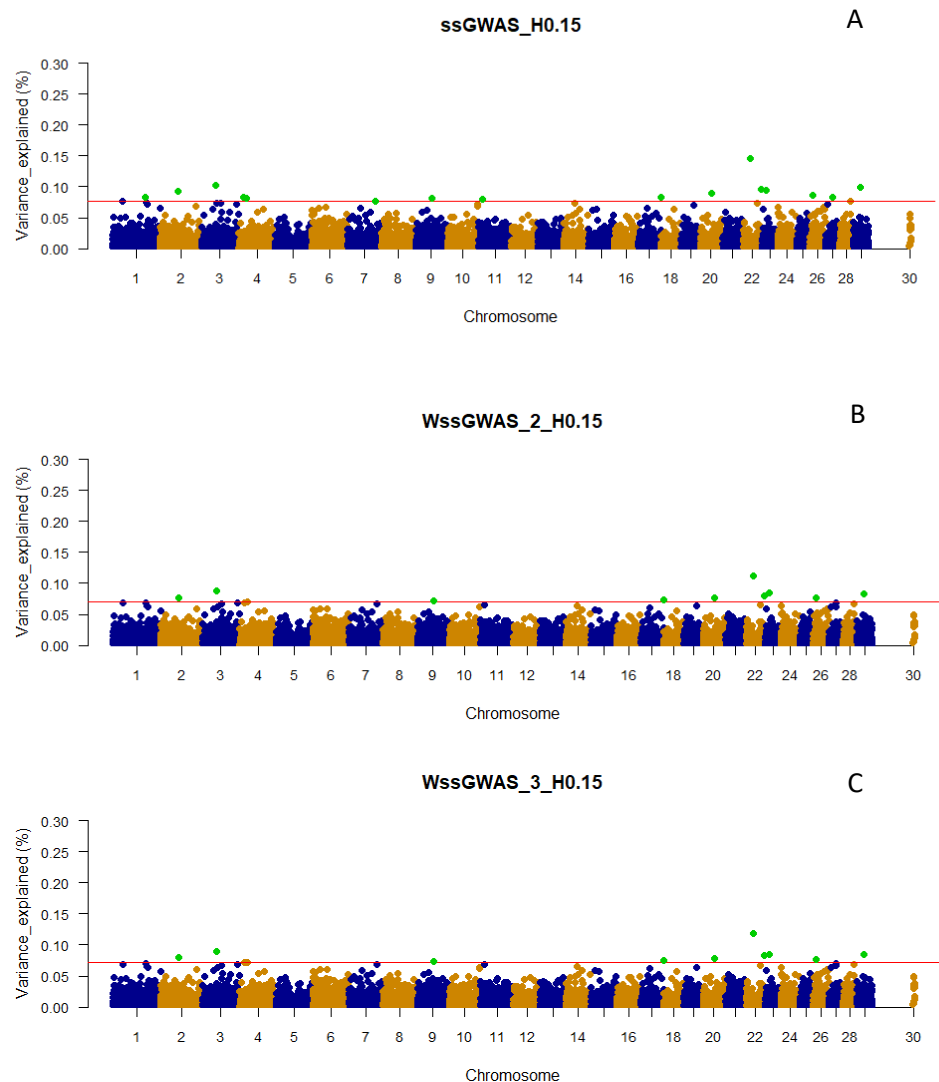


Figure 2. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.15 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.15; **A**) and weighted single-step GWAS in the second (WssGWAS_2_H0.15; **B**) and third iterations (WssGWAS_3_H0.15; **C**). Green points highlighted above the red horizontal line are the top 0.001% of markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30.

3.4. Overlapping Genomic Regions Among Methods and Functional Analyses

3.4.1. Overlapping Markers

Considerable overlap among many of the top markers was found by the different ssGWAS methods. The majority of the top markers identified by ssGWAS using only haplotypes were also present in the scenarios using NCSNP and haplotypes together (Figure 8; Supplementary File 4). Only one top haplotype in the ssGWAS_H0.15 and ssGWAS_H0.50 scenarios and three haplotypes in the

ssGWAS_H0.80 were found exclusively when using haplotypes, i.e., all other markers were captured by their respective scenarios using NCSNP. NCSNP allowed us to identify unique top regions within each LD threshold used to build the haploblocks, with ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80 detecting 28, 53, and 58 additional top regions than their respective scenarios using only haplotypes. When comparing all scenarios fitting only haplotypes, 2 regions were common between ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80, at 89 Mb on the BTA7 and 38 Mb on BTA20. Furthermore, top regions were specifically identified in each scenario with different LD thresholds when only haplotypes were used in the ssGWAS, with 13, 6, and 3 haplotypes identified exclusively in ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80, respectively. A total of 10 markers were found in common between all methods, including NCSNP in ssGWAS. The scenarios with NCSNP and haplotypes together also presented markers exclusively found by specific LD thresholds, with 27, 33, and 32 markers identified by the ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80 scenarios, respectively. The 2 common regions between all ssGWAS scenarios using haplotypes were also present within the 10 common regions when fitting NCSNP and haplotypes together.

3.4.2. Overlapping Genes

The overlapping markers among scenarios were also present in genes shared between ssGWAS strategies using only haplotypes built with different LD thresholds and including the NCSNP. All annotated genes identified based on the ssGWAS_H0.15 scenario were also identified by ssGWAS_NCSNP_H0.15 (Figure 9; Supplementary File 5). Only one (*COMMD10*) and three (*NID2*, *PLXDC1*, and *DOCK1*) annotated genes were identified exclusively by the ssGWAS_H0.50 and ssGWAS_H0.80, respectively, compared to the scenarios including NCSNP. Considering all ssGWAS scenarios using only haplotypes, a unique gene (*SPEF2*) was found by the three scenarios. Seven genes (*5S_rRNA*, *UMAD1*, *PTPRC*, *SPEF2*, *CACNA2D3*, *HMGCLL1*, and *MGMT*) were identified by all three ssGWAS scenarios, including NCSNP, and the unique gene overlapped by all three scenarios, including haplotype-only methods, was also present among them.

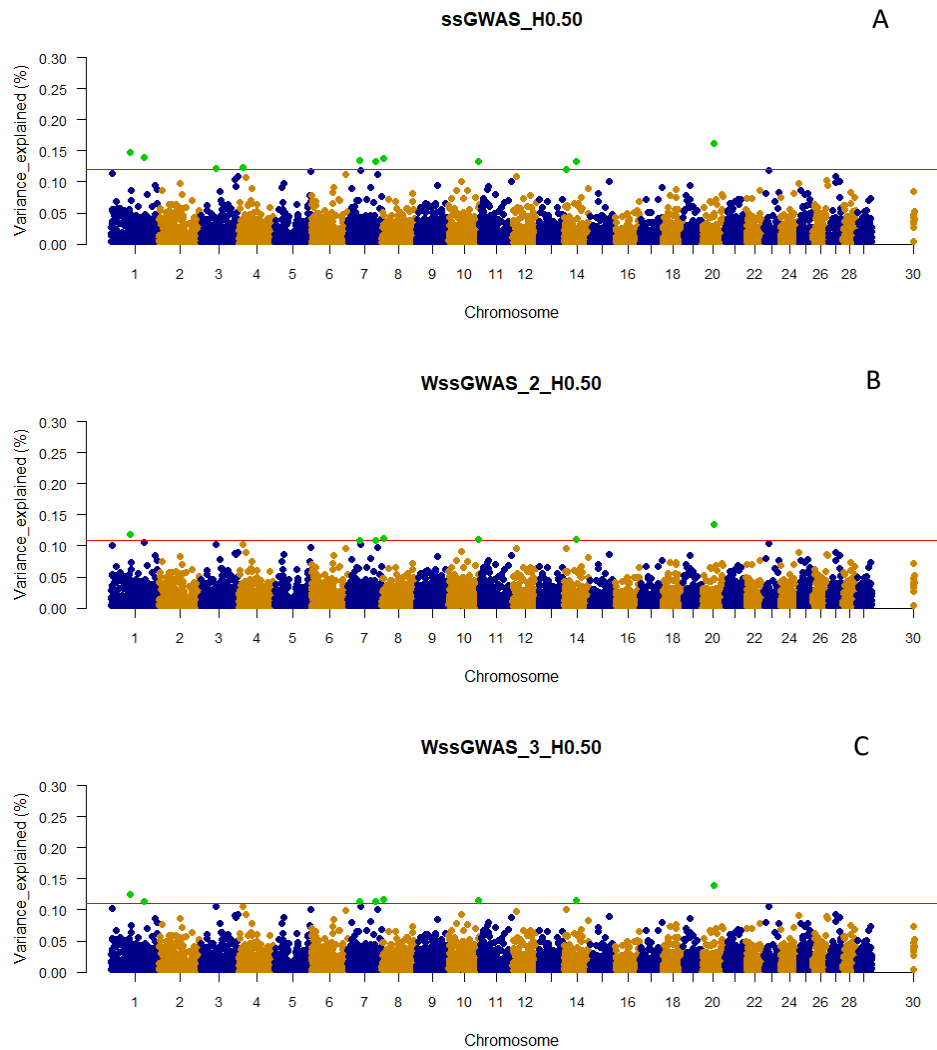


Figure 3. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.50 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.50; **A**) and weighted single-step GWAS in the second (WssGWAS_2_H0.50; **B**) and third iterations (WssGWAS_3_H0.50; **C**). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30.

3.4.3. Functional Analyses

The results from the functional analyses are presented in Table 3 and Supplementary Files 6 and 7. No clusters were significantly enriched using default parameters in the DAVID platform. For simplicity, only the candidate genes, GO_BP, and KEGG metabolic pathways from the Functional Annotation tables provided by DAVID for key candidate genes with direct or indirect implications in behavioral or docility traits such as those related to the nervous system were presented. Details about all the overlapping genes are presented in

the Supplementary Files 2 and 3, and the Functional Annotation tables for all genes are presented in Supplementary Files 6 and 7.

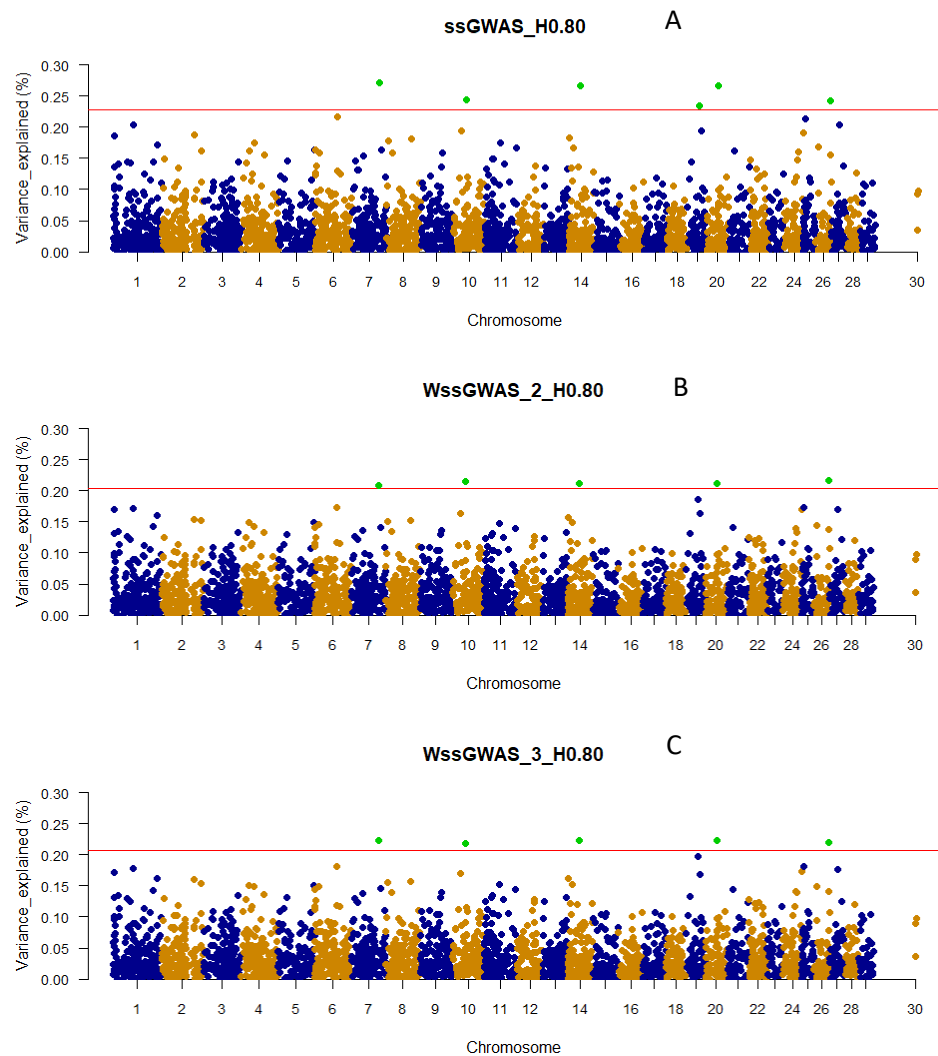


Figure 4. Manhattan plot of the percentage of the total additive genetic variance explained by haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.80 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_H0.80; **A**) and weighted single-step GWAS in the second (WssGWAS_2_H0.80; **B**) and third iterations (WssGWAS_3_H0.80; **C**). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30.

The gene *ATXN10* (position 116 Mb on BTA5) was annotated in the GO:0031175 term, which is associated with neuron projection development. The bta05010 KEGG metabolic pathway was annotated for the gene *ADAM10* (position 51 Mb on the BTA10) and is related to the Alzheimer disease. The gene *VAX2* (position 13 Mb on BTA 11) was annotated in the GO:0007409, GO:0007601, GO:0030900, GO:0048048, and GO:0060041 biological processes,

which are axonogenesis, visual perception, forebrain development, embryonic eye morphogenesis, and retina development in camera-type eye, respectively. The biological processes in GO:00076605 and GO:0042472 included the gene *ATP6V1B1* (position 13 Mb on BTA11) and are related to sensory perception of sound and inner ear morphogenesis, respectively. GO:0060325 is related to face morphogenesis and includes the *CRISPLD1* gene (position 38 Mb on the BTA14). The *CAPRIN1* gene (position 64 Mb on the BTA15) was annotated in GO:0050775 and GO0061003, which are related to the positive regulation of dendrite and dendrite spine morphogenesis, respectively. Two biological processes included the *FA2H* gene (position 2 Mb on BTA18), which are GO:0032286 and GO:0032287, known to be related to central and peripheral nervous system myelin maintenance, respectively. The gene *SPEF2* (position 38 Mb on BTA20) was annotated in GO:0048702, GO:0048854, and GO:0069541, which are related to the embryonic neurocranium, brain morphogenesis, and respiratory system development. Four biological processes related to the nervous system included the *PLXNA1* gene, which are GO:0021785, GO: 0048841, GO:1902287, and GO:1990138, known to be related to branchiomotor neuron axon guidance, regulation of axon extensions involved in guidance, the semaphorin-plexin signaling pathway involved in axon guidance, and neuron projection extension, respectively. The *PLXNA1* gene was also annotated in the bta04360 KEGG pathway, which is related to axon guidance. Finally, the gene *CACNA2D3* was annotated in the bta04921 KEGG pathway and is related to the oxytocin signaling pathway.

3.5. QTL Overlapping with the Top 0.001% Markers for Yearling Temperament

Similar to what was observed with the genes, the overlapping markers across scenarios also implied in QTL found in more than one scenario. All QTL identified using only haplotypes with LD thresholds of 0.15 and 0.50 were captured when the NCSNP were used in the ssGWAS, while only two QTL were found by ssGWAS_H0.80 and not by ssGWAS_NCSNP_H0.80 (Figure 10; Supplementary File 8). Using different LD thresholds to create the haploblocks resulted in specific QTL captured by the different block structures, with 39, 27, and 2 QTL identified exclusively by ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80, respectively. Adding NCSNP in ssGWAS also resulted in QTL captured by specific scenarios regarding the LD threshold to create the haploblocks, with 77, 78, and 69 QTL identified exclusively by

ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80, respectively. A total of 28 QTL were identified by all 3 scenarios using haplotypes only and were also present among the 76 QTL identified by all 3 scenarios including NCSNP and haplotypes.

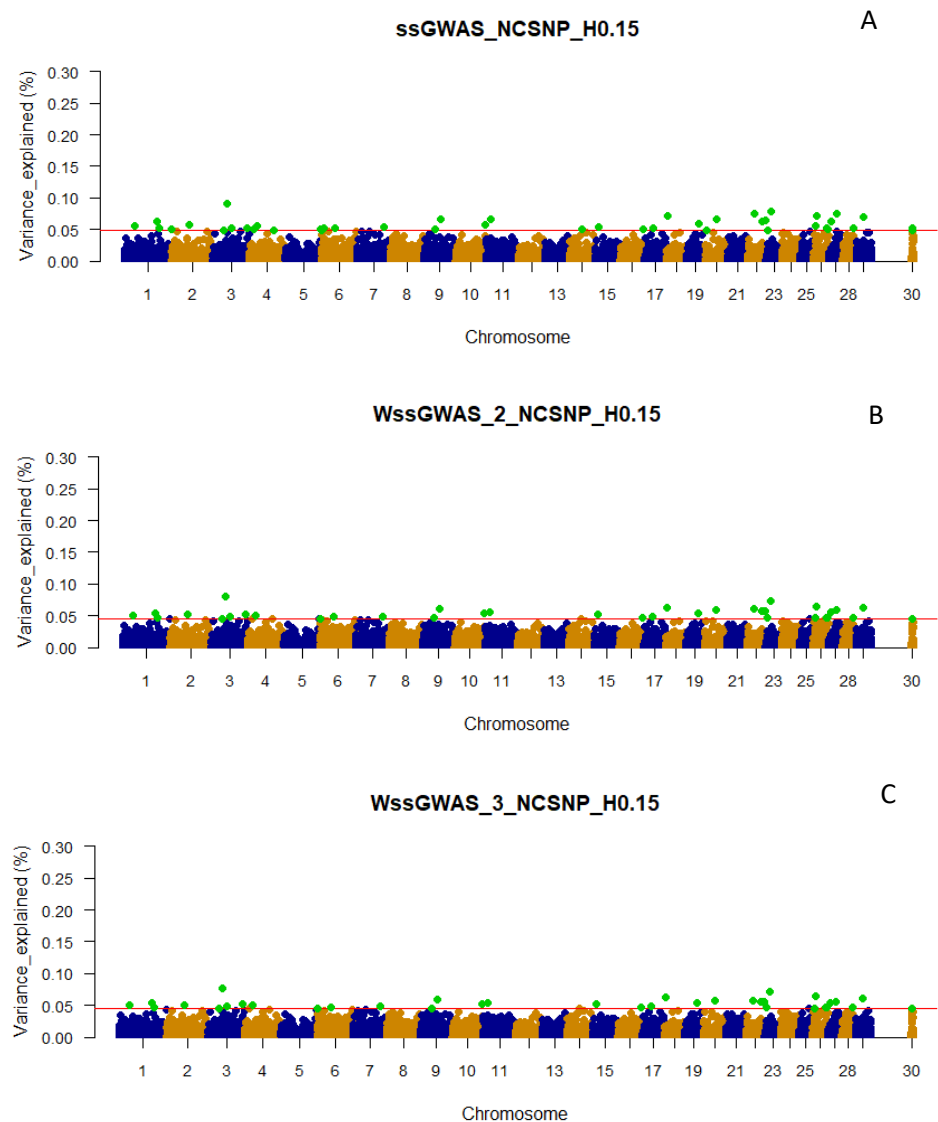


Figure 5. Manhattan plot of the variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.15 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.15; **A**) and weighted single-step GWAS in the second (WssGWAS_2_NCSNP_H0.15; **B**) and third iterations WssGWAS_3_NCSNP_H0.15; **C**). Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome is represented by the chromosome 30.

The QTL identified by the scenarios evaluated in this research belong to the classes “Milk”, “Health”, “Exterior”, “Production”, and “Reproduction” (Figure 11). The majority of the QTL identified in each scenario were related to the class

“Exterior”, except for the ssGWAS_NCSNP_H0.15 scenario (for this, the class “Milk” contained the majority of the QTL). “Milk” was the class most often found among QTL after the class “Exterior”, followed by “Production”, “Reproduction”, and, lastly, “Health”. The QTL from the class “Health” were found only when NCSNP were included in the ssGWAS.

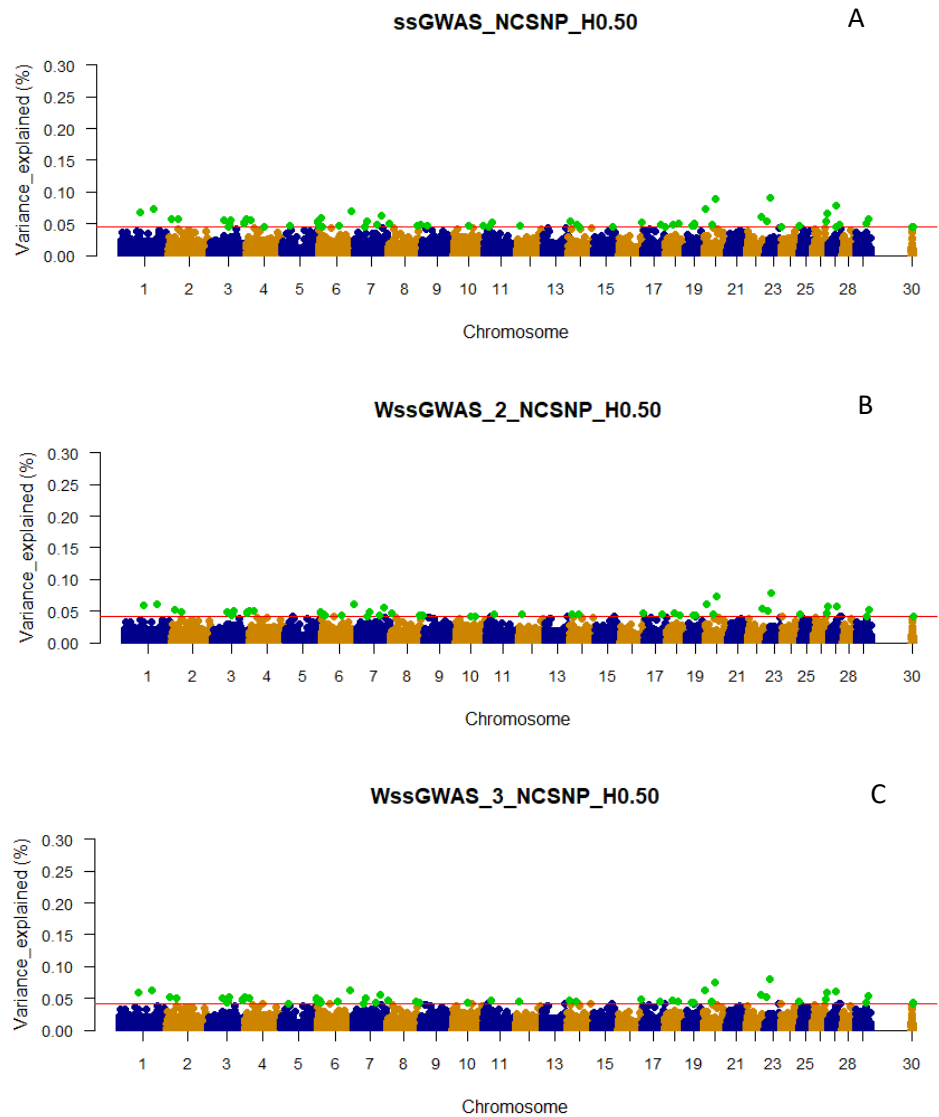


Figure 6. Manhattan plot of the percentage of the total additive genetic variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.50 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.50; A) and weighted single-step GWAS in the second (WssGWAS_2_NCSNP_H0.50; B) and third (WssGWAS_3_NCSNP_H0.50; C) iterations. Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome

4. Discussion

We have investigated the genomic architecture of YT using large phenotypic and genomic datasets from American Angus cattle. Different haplotype structures obtained by three LD thresholds to create blocks were used while including or excluding the NCSNP in order to capture different genes and QTL structures that could affect YT in American Angus.

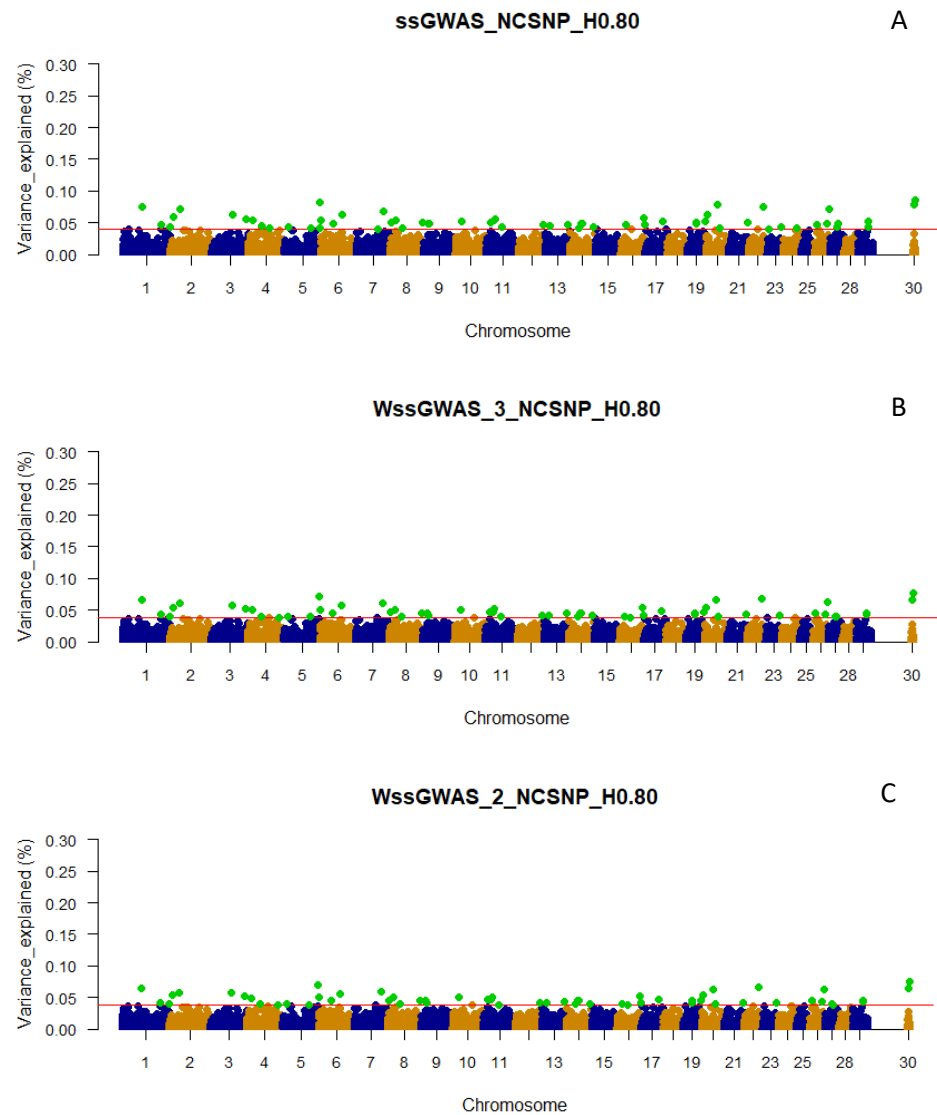


Figure 7. Manhattan plots of the total additive genetic variance explained by non-clustered SNPs and haplotypes from haploblocks built with a linkage disequilibrium threshold of 0.80 for yearling temperament in American Angus cattle using single-step GWAS (ssGWAS_NCSNP_H0.80; **A**) and weighted single-step GWAS in the second (WssGWAS_2_NCSNP_H0.80; **B**) and third (WssGWAS_3_NCSNP_H0.80; **C**) iterations. Green points highlighted above the red horizontal line are the top 0.001% markers that explained greater percentages of the total additive genetic variance for YT. The X-chromosome (PAR region) is represented by the chromosome 30.

4.1. Empirical Selection of the Candidate Genomic Regions

All the steps to obtain the top 0.001% genomic regions for YT presented in the Section 2.7 were done because of the lack of a statistical method to properly test for significance of markers using ssGWAS for threshold traits. The approximated p -values for the markers in the ssGWAS approach was proposed under normality assumptions [41] and are not available for threshold traits.

Only defining a fixed value to select the candidate regions, e.g., 0.50 or 1.00% of the additive variance, as most of the GWAS studies employ, would not be an equivalent comparison with scenarios using different marker densities (e.g., number of pseudo-SNPs from haploblocks with different LD thresholds or including or excluding the NCSNP; Table 2), as the percentage of the additive genetic variance explained by each marker is inversely proportional to the number of markers (Figures 2–7). In addition, it is possible that markers with a percentage of the total additive genetic variance smaller than 0.50% are biologically associated with the traits of interest. Aguilar et al. [41] presented significant p -values for markers that explained ~0.10% (similar to some scenarios in this research) of the additive variance for the birth weights in American Angus using more than one million phenotypes and approximately 1.4 K genotyped sires with phenotyped progeny.

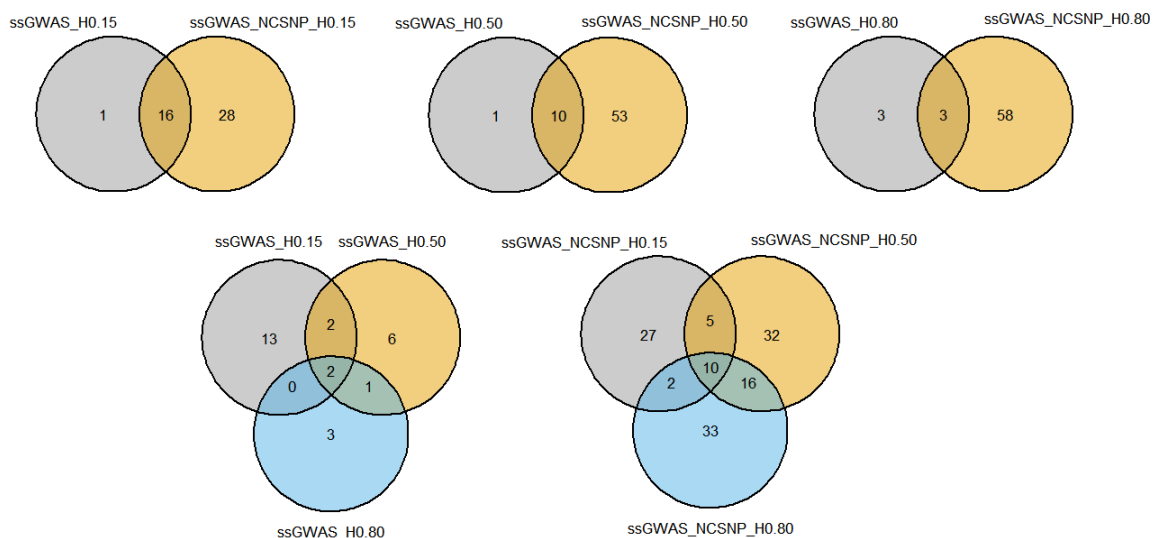


Figure 8. Venn diagrams showing the number of markers overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively.

It is well known that the significance and the additive variance of the markers are dependent on the sample size and population structure, as well as the allelic frequencies, additive values of the QTL tagged by the marker, LD between marker and QTL (assuming that the QTL is not the marker), and the accuracy of the phenotypic information [20,41]. The interaction of these components is complex and makes it difficult to define a threshold for selecting candidate markers based only on the percentage of the total additive genetic variance explained by the markers (or genomic windows). Various key candidate genes associated with GO_BP and KEGG pathways were found. These may play an important role in YT, as well as previously reported QTL. Nevertheless, further studies are needed to evaluate the method to obtain the top regions in this study, as defining false or true positive associations is not straightforward when using real datasets [20].

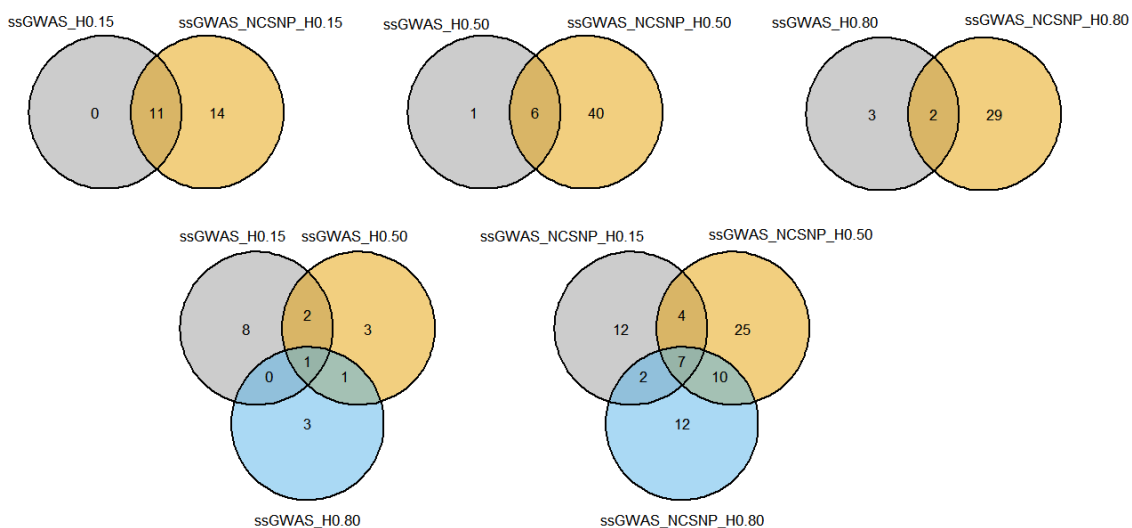


Figure 9. Venn diagrams showing the number of genes overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD threshold of 0.15, 0.50, and 0.80, respectively.

4.2. Additive Genetic Variance Explained by Genomic Regions across Scenarios

It was not surprising to not find regions explaining more than 1% of the total additive genetic variation for YT, since behavioral traits are highly polygenic [4,5,42]. Overall, the variances explained by each unique genomic region (i.e., haplotype or SNP) in all scenarios were small and distributed across the chromosomes (Figures 2–7 and Supplementary Files 2 and 3), highlighting the polygenic nature of YT, which was also reported by Alvarenga et al. [5] using

only SNPs. Alvarenga et al. [5] found 11 genomic windows considering five adjacent SNPs explaining about 3.33% of the total additive genetic variance for YT, and these regions were distributed across the bovine autosome chromosomes, similar to the results in the present study.

Table 3. Gene ontology biological terms (GO_BP) and metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) access of genes overlapped by top 0.001% markers for docility in American Angus cattle ¹.

Chromosome	Gene	S_P ² (Mb)	E_P ³ (Mb)	GO_BP	KEGG
BTA5	<i>ATXN10</i>	116.029	116.169	GO:0031175	-
BTA10	<i>ADAM10</i>	51.536	51.679	-	bta05010
BTA11	<i>VAX2</i>	13.483	13.509	GO:0007409, GO:0007601, GO:0030900, GO:0048048, GO:0060041	-
BTA11	<i>ATP6V1B1</i>	13.454	13.480	GO:00076605, GO:0042472	-
BTA14	<i>CRISPLD1</i>	38.295	38.346	GO:0060325	-
BTA15	<i>CAPRIN1</i>	64.662	64.697	GO:0050775, GO0061003	-
BTA18	<i>FA2H</i>	2.151	2.206	GO:0032286, GO:0032287	-
BTA20	<i>SPEF2</i>	38.369	38.573	GO:0048702, GO:0048854, GO:0069541	-
BTA22	<i>PLXNA1</i>	60.240	60.280	GO:0021785, GO: 0048841, GO:1902287, GO:1990138	bta04360
BTA22	<i>CACNA2D3</i>	45.925	46.819	-	bta04921

¹ Only genes with a GO_BP or metabolic pathway related to a behavior or docility trait are presented. Details for all genes harboring the top markers are presented in Supplementary Files 6 and 7. ² Start position. ³ End position.

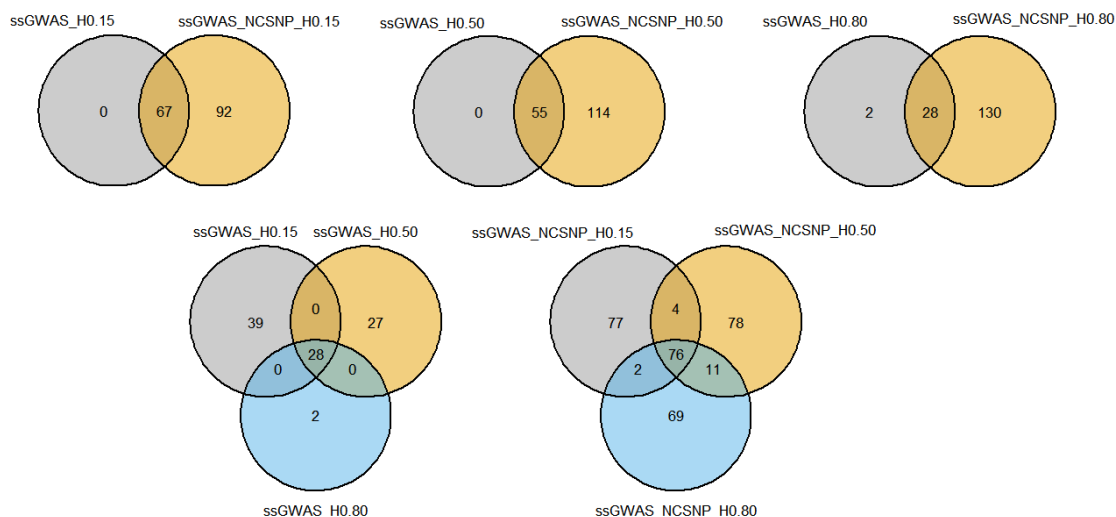


Figure 10. Venn diagrams showing the number of QTL overlapping among different single-step genome-wide association studies (ssGWAS) with haplotypes and non-clustered SNPs. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively.

The fact that WssGWAS did not increase the variance explained by major genomic regions gives further evidence of the polygenic nature of YT (Figures 2–7). Furthermore, a high correlation was observed between the GEBVs from

ssGWAS and WssGWAS methods (greater than 0.98; Supplementary File 10) regardless of the LD threshold used to create the haploblocks (i.e., 0.15, 0.50, 0.80). An assumption in ssGBLUP, and consequently ssGWAS, is that all markers explain similar and small proportion of the total additive genetic variance. Thus, WssGBLUP was developed in order to minimize this effect by giving priorities to some markers with potentially greater effects [19,23]. In this case, one would expect that the genomic regions with a higher impact on the variance of the trait would present higher peaks based on WssGWAS compared to ssGWAS. Substantial changes in the GEBVs, higher accuracies, and lower bias of genomic predictions are also expected with WssGBLUP for those traits with major genes, i.e., regions that should receive greater weight [19,23,43]. Hence, due to the evidence of this polygenic nature, the results from the ssGWAS scenarios were used to identify candidate genes and QTL.

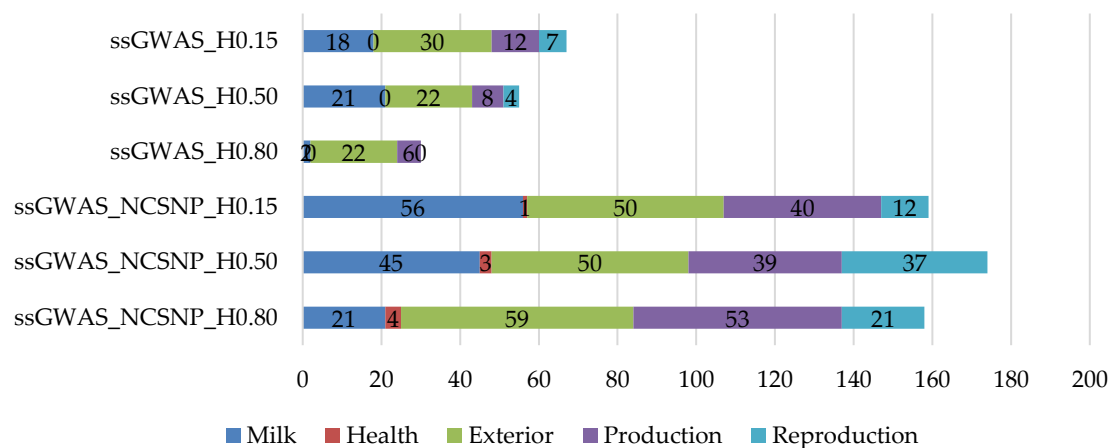


Figure 11. Absolute number of quantitative trait loci (QTL) by class overlapping with the top 0.001% markers for yearling temperament in American Angus cattle using the single-step GWAS fitting only haplotypes or non-clustered SNPs and haplotypes. ssGWAS_H0.15, ssGWAS_H0.50, and ssGWAS_H0.80: ssGWAS using only haplotypes from blocks with linkage disequilibrium (LD) thresholds of 0.15, 0.50, and 0.80, respectively; ssGWAS_NCSNP_H0.15, ssGWAS_NCSNP_H0.50, and ssGWAS_NCSNP_H0.80: ssGWAS using non-clustered SNPs and haplotypes from blocks with LD thresholds of 0.15, 0.50, and 0.80, respectively.

4.3. Weighting Method in the Single-Step GWAS

In early stages in this study, the weighting method proposed by Wang et al. [19] in the WssGWAS beyond the NLA was attempted for comparison purposes. However, problems during the iterative process using the Wang et al. [19] method regarding the inversion of the \mathbf{G} matrix used, and it was not possible to generate results for the second and third iterations in the majority of the scenarios. The NLA method proposed by VanRaden [32] is conservative in the

shrinking process due to the limits in the maximum changes in the weights applied [34]. However, it is relevant to point out that Wang et al. [19]'s assumptions on marker weights rely only on the SNP effect and allelic frequency. The weights considered by Wang et al. [19] in this study had a wider range (from ~0.5 to ~28), and are most likely less accurate, considering the polygenic nature of YT, compared to the NLA approach (~0.9 to ~1.6) after the first iteration of the weighting process (ssGWAS).

This problem using the Wang et al. [19] method could be a result of the haplotypes being more polymorphic than SNPs [10–12,18], so the broader range in the weights could be due to the larger number of alleles in the same region compared to SNP tracking the actual additive genetic effect. As the GEBV accuracy can decline and bias can increase during the iterative process using the Wang et al. [19] method [34], consequently, the SNP effects and variances could be less accurate and more biased. Therefore, it is recommended to use NLA weights for WssGWAS purposes because it results in more robust weighting values.

4.4. Genes and QTL Overlapping the Top Genomic Regions

Some of the genes present in the top regions for YT were previously reported in other studies to be associated with behavioral traits. The genes *7SK*, *U6*, and *5S_rRNA* were associated with behavior in cattle by Alvarenga et al. [5,8]. The *7SK* gene is a small nuclear RNA gene that belongs to a class of subunits spread in the bovine genome and that already had a unit previously associated with fertility traits [44]. The *U6* gene is a gene found in more than one BTA and also belongs to a small nuclear RNA class which was previously related to temperament [45], maternal behavior [46], and sucking reflex [47] in cattle. Beyond small nuclear RNA genes, small nucleolar RNA genes (*SNORD25*, *SNORD26*, and *SNORD27*) and long non-coding RNA genes (*lncRNA*) were found, but no previous functions related to behavior, production, reproduction, or health were found. Beyond the presence in a top region (position 41 Mb on BTA29) for YT, QTL related to milk production and reproduction (two and 11, respectively) were also annotated in the same top region (block_81_chr_29; Supplementary File 3) where those small nuclear and long non-coding RNA genes were found. The RNA genes codify transcriptional factors required for splicing [48] so they can be involved in many different processes affecting gene

expression. The *5S_rRNA* gene is a RNA ribosomal gene that has also been reported to be involved in milk, meat quality, and carcass traits [49].

Despite the small nuclear, nucleolar, and ribosomal RNA genes found, the majority of the genes annotated in the top regions are protein coding (Supplementary Files 2 and 3). Most of the protein-coding genes are related to a broad range of more basal functions (e.g., glucose metabolism, pH regulation, transcription; Supplementary Files 6 and 7) according to the Functional Annotation Table from the DAVID platform, which could explain the absence of significant clusters. The presence of many different biological processes affecting YT is expected, as behavioral traits involve many functions between the perception of a stimulus and the reaction to a specific stress or situation [50,51]. Nevertheless, genes present in GO_PB and KEGG pathways related to nervous system development, mental disorders, stimuli perception, and respiratory development (Table 3) were also found, and these functions are related to behavioral stress responses [1,4,51].

The ataxin 10 gene (*ATXN10*), annotated in the neuron projection development GO_BP, was previously associated with longevity traits in Chinese Holstein cattle [52]. Longevity is a productive trait that is affected to some degree by the cattle's temperament, as aggressiveness is an undesirable trait and is a culling criterion in American Angus [53]. The ADAM metallopeptidase domain 10 (*ADAM10*) gene was already identified as a biomarker for Alzheimer's disease in humans, with functions related to the cleavage amyloid precursors that act during the inflammation process of senile plaques [54]. In cattle, *ADAM10* was associated with tick resistance, with its importance in the inflammation process being the most likely reason [55].

The ventral anterior homeobox 2 (*VAX2*) gene was annotated for biological processes related to visual perception, axonogenesis, and forebrain development, which are very important processes in behavioral responses to environmental stimuli [1]. An association with fertility-related traits was also previously reported for the *VAX2* gene in cattle [56]. Beyond visual perception, auditive-related biological processes (sensory perception of sound and inner ear morphogenesis) were annotated for the gene ATPase H⁺ transporting V1 subunit B1 (*ATP6V1B1*), indicating that vision and auditory senses are among the main functions influencing YT. The animal's perception of the area around it is involved in behavioral responses [1], so that the presence of the handler or other

individuals can affect the animal's response, which can be positive or not depending on the interaction. The *ATP6V1B1* gene was also previously associated with carcass composition traits [57], which makes sense due to other biological processes this gene is involved in (e.g., pH regulation, ossification; Supplementary Files 6 and 7).

The face morphogenesis biological process, annotated for the cysteine-rich secretory protein LCCL domain-containing 1 (*CRISPLD1*) gene, was an interesting result. Neural crest cells are involved in face morphogenesis by generating the craniofacial skeleton, particularly the sensory organs and subsets of cranial sensory receptor neurons, and there are common mechanisms for building faces, brains, peripheral neurons, and central neural circuits that regulate behavioral functions [58]. In addition, the face structure has already been cited as a predictor of aggressiveness in humans, with the specific facial width-to-height ratio highly correlated with aggressiveness in men [59]. However, these associations of face morphogenesis and structure are limited in domestic animals, and the only report for the *CRISPLD1* gene found in cattle was for milk fatty acid traits [60].

Previous studies have reported face hair whorls (FHW) in cattle to be a predictor of cattle temperament [61–63]. In these studies, animals with FHW above the eye line were more agitated than the ones with lower FHW; these animals escaped faster or displayed aggressive behavior. The association of FHW with cattle temperament may also be related to face morphogenesis, which could be associated with early tissue development, as the same embryonic origin is attributed to the epidermis and the nervous system [64]. Recently, genomic regions for FHW in horses were annotated [65], in which, beyond the hair follicle growth, they were related to neurological and behavioral functions. Despite the phenotypic association of cattle temperament with FHW [61–63], no studies underlying the genomic architecture of FHW in cattle were found.

Different dendritic spine densities were previously associated with aggressiveness in rats [66], and the dendrite and dendrite spine morphogenesis GO_BP was annotated for the cell cycle-associated protein 1 (*CAPRIN1*) gene. The *CAPRIN1* gene was previously enriched for bovine respiratory disease [67]; however, studies reporting associations for this gene in cattle are scarce. The haploblock that overlapped the *CAPRIN1* gene (block_135 position 64 Mb on BTA15) also overlapped with 2 QTL for milk fat yield. The fatty acid 2-

hydroxylase (*FA2H*) gene was previously associated with carcass traits [68] and was functionally annotated to the central and peripheral nervous system myelin maintenance GO_BP in our study. Differences in the myelination were found in mice that presented social avoidance behavior (susceptibility to behavioral stress), with less social mice having thinner myelin compared to the controls [69]. Fat is one of the main components of the myelin layers surrounding the nerves and has an important role in the electric transmission across nerve cells [70], and the presence of a fatty acid gene may suggest the involvement of fatty acid metabolism in its formation and maintenance.

The sperm flagellar 2 (*SPEF2*) gene is a well-known gene in cattle because of its association with important traits, such as adaptation to heat stress [71], fertility [72], and milk production and composition [60]; however, no reports related to behavioral traits for this gene were found. The *SPEF2* gene was found for all scenarios investigated, and beyond the GO_BP related to immune system and fertility, the embryonic neurocranium, brain morphogenesis, and respiratory system development GO_BP were also functionally annotated. Respiration is changed during flight or fight response in animals [73] and consistent respiratory alterations were observed in highly aggressive rats, with elevated basal respiratory rates denoted for the highly aggressive animals compared to controls [74]. Knowledge about alteration of the respiratory rate as a function of behavior is limited in cattle.

The GO_BP and metabolic pathway that the plexin A1 (*PLXNA1*) gene are involved were mainly related to axon guidance and neuron projection. Specific neuron projections during aggression were previously described in mice, with some periaqueductal gray (PAG) neurons being selective for attack action [75]. No behavior or neuron studies reporting associations of the *PLXNA1* gene were found in cattle; however, this gene was associated to fertility [76] and growth traits [77] in cattle. As the PAG brain region is conserved across species [78] and plays roles in survival behavior [75], further behavioral studies considering this gene in cattle or other species are recommended.

The social behavior response is affected by the oxytocin signaling pathway (OSP) [79], and this KEGG metabolic pathway was annotated for the calcium voltage-gated channel auxiliary subunit α 2 delta 3 (*CACNA2D3*) gene. Reports related to OSP and social interactions in cattle are scarce. In cattle, the *CACNA2D3* gene was previously associated with intramuscular fat [80], which

is in accordance with the fact that the MAPK signaling pathway was also annotated for this gene and is associated with marbling [81], but no studies related to behavior indicators were found.

The vast diversity of functions found by the genes present in the top regions for YT and the previous associations with different sorts of traits may suggest pleiotropic effects. This can also be supported by the range of trait classes presented by QTL overlapped by the top regions (Figure 11), which were related to more than 50 different traits linked to milk production, health, exterior, production, and reproduction traits (Supplementary Files 2 and 3). Three QTL for milking speed (position 43 Mb on BTA7, and positions 30 and 41 Mb on BTA14), an indicator of workability in cattle [82], were overlapped by the top regions when NCSNP and haplotypes were fitted in ssGWAS (Supplementary File 3). Additionally, for another indicator of behavior in cattle, 5 QTL for length of the productive life were found (position 71 Mb on BTA6, positions 21 and 43 Mb on BTA18, position 47 on BTA20, and position 19 on BTA 24).

4.5. Use of Different Linkage Disequilibrium Thresholds and Non-Clustered SNPs in the ssGWAS

Many of the genes and QTL would not have been identified without using different LD thresholds to create the haplotype blocks or the inclusion of the NCSNP. Haplotype-based GWAS using overlapping sliding windows was suggested as more powerful than SNP-based or LD-based haplotype GWAS (considering low recombination rates within haploblocks, i.e., high LD levels), because these would be more efficient for regions with low LD and high recombination [83]. Exclusive genes and QTL were found when different LD thresholds (0.15, 0.50, and 0.80, which are low, moderate, and high, respectively) were used to create the haploblocks, suggesting the LD levels in the regions that affect YT are not consistent. Considering the complexity of the genomic organization, QTL sizes, and genetic factors experienced by the populations, it is unlikely that QTL affecting any trait would follow a specific LD pattern. Using haplotypes and NCSNP under the ssGWAS framework with a low, moderate, and high LD to create the haploblocks allowed us to find genes and biological processes involved in YT that were not reported in previous studies, which used only SNPs, for a set of behavioral traits in cattle.

The LD-block approach considering high LD levels was previously suggested to be inefficient for association studies because many individual SNPs

would be placed out of the haploblocks [84], and thus could not contribute to dissecting the genetic architecture of the trait [35]. Additionally, results from genomic predictions using haplotypes not including the NCSNP were worse regardless of the level of genetic diversity and heritability of the trait, indicating that important genomic regions were not considered without the NCSNP [18]. Our results show important genes and QTL for YT would not be considered without the inclusion of NCSNP, and using both haplotypes and NCSNP accounts for most relevant genes and QTL found based on haplotypes only. Top genomic regions in the X chromosome (PAR region) were found only when NCSNP were included in the analyses. Two genes (*MXRA5* and *CD99* in the position 138 Mb; Supplementary File 3), three QTL related to metabolic body weight (position 136 Mb), and other regions that did not have genes or QTL previously annotated (positions 134 and 135 Mb) were found in the X chromosome (PAR region) when fitting NCSNP and haplotypes together. This finding indicates that further studies including additional markers located in the X chromosome are needed, as also suggested by Alvarenga et al. [5]. Thus, association studies using LD haplotypes should also include the NCSNP.

It is important to highlight that the density of the panel used to make the haplotype analyses can affect the results, as the amount of QTL variance explained tend to be higher with denser haploblocks [85-86]. The precision in the estimation of the recombination hotspots also tend to be higher with denser SNP panels [87], which can affect the accuracy of haplotype phasing [88]. The 50 K SNP panel used in this research was designed to be similar to the Illumina BovineSNP50V2 and Illumina BovineSNP50V3 SNP panels, with SNPs 50.6 kb apart on average (www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp50.pdf, accessed on 11 December 2021), presenting SNPs as further apart compared to the high-density panel available for cattle (~777 K SNPs). However, as the haplotype blocks in cattle are ~70 kb [89], the 50 K panel provides reasonable resolution for capturing the extent of LD in the population investigated. Nevertheless, further studies using denser SNP panels are also recommended to investigate ssGWAS using haplotypes for YT, as well as other economically important traits and alternative indicators of cattle temperament.

4.6. Future Studies

Additional research should be conducted next to further explore the results obtained in this current study. For instance, it would be valuable to repeat these analyses based on whole-genome sequence data in North American Angus cattle as well as in other worldwide Angus cattle populations. The key candidate genes should also be validated in vitro or based on gene editing and gene knock-out experiments. Furthermore, additional polymorphisms located in the candidate genes identified in this study could be added to existing SNP panels to increase the accuracy of genomic predictions for docility. From a practical point of view, these results obtained could be incorporated in current genomic prediction models for YT in North American Angus. We are also evaluating the genetic trends of docility using the traditional and genomic estimated breeding values for YT and correlated responses in other important traits. Another area of research in our research group is the definition of novel indicators of cattle temperament, especially based on data derived from sensors and other precision technologies.

5. Conclusions

Yearling temperament in cattle is a highly polygenic trait, with genes and QTL broadly distributed across the whole genome. Association studies using LD-based haplotypes should include the non-LD-clustered SNPs, as well as different thresholds to increase the likelihood of finding the genomic regions affecting the phenotype of interest. The key candidate genes *ATXN10*, *ADAM10*, *VAX2*, *ATP6V1B1*, *CRISPLD1*, *CAPRIN1*, *FA2H*, *SPEF2*, *PLXNA1*, and *CACNA2D3* are involved in important biological processes and metabolic pathways related to behavioral traits, social interactions, and aggressiveness. Further studies investigating the role of these genes in behavioral traits are recommended.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13010017/s1>, File S1: Distributions of the percentage of the additive variance explained by makers from single-step GWAS using haplotypes for yearling temperament in American Angus cattle. File S2: Genes and QTL overlapped with top regions using haplotypes in single-step GWAS for yearling temperament in American Angus cattle. File S3: Genes and QTL overlapped with top regions using haplotypes and non-clustered SNPs in single-step GWAS for yearling

temperament in American Angus cattle. File S4: Overlapped top regions among scenarios using haplotypes or non-clustered SNPs in single-step GWAS for yearling temperament in American Angus cattle. File S5: Overlapped top genes among scenarios using haplotypes or non-clustered SNPs in single-step GWAS for yearling temperament in American Angus cattle. File S6: Functional annotation tables for genes overlapped by top regions using haplotypes in single-step GWAS for yearling temperament in American Angus cattle. File S7: Functional annotation tables for genes overlapped by top regions using haplotypes and non-clustered SNPs in single-step GWAS for yearling temperament in American Angus cattle. File S8: Overlapped QTL among scenarios using haplotypes or non-clustered SNPs in single-step GWAS for yearling temperament in American Angus cattle. File S9: Correlations between the GEBV used in the ssGWAS and WssGWAS with haplotypes and non-clustered SNPs for yearling temperament in American Angus.

Author Contributions: A.C.A. and L.F.B.: conception of the work. A.B.A.: phenotypic and pedigree quality control and model definition. A.C.A.: haplotype-based GWAS analyses. A.C.A. and L.F.B.: results interpretation. A.C.A. and L.F.B.: drafting the manuscript. A.C.A., P.L.S.C., H.R.O., A.B.A., S.P.M., K.R., and L.F.B.: critical revision of the manuscript. A.C.A., P.L.S.C., H.R.O., A.B.A., S.P.M., K.R. and L.F.B.: final approval of the version to be published. All authors contributed to the article and approved the submitted version.

Funding: This research was funded by Purdue University, State University of Southwest Bahia and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES) Finance Code 001.

Institutional Review Board Statement: Ethical review and approval were waived for this study, as all datasets used were provided by commercial breeding operations.

Informed Consent Statement: Not applicable.

Data Availability Statement: The phenotypic and genomic data used in this study are the property of the industry partner that contributed to the study and therefore are not readily available due to its commercial sensitivity. Requests to access the datasets should be directed to the American Angus Association. The computing pipelines used in this research are available by request to the corresponding authors.

Acknowledgments: We acknowledge the American Angus Association for providing the datasets used in this research. We also thank the members of Brito's lab for providing scientific support to develop this research. Lastly, we thank Purdue University and the State University of Southwest Bahia for providing academic and financial support to the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grandin, T.; Deesing, M.J. Behavioral genetics and animal science. In *Genetics and the Behavior of Domestic Animals*, 2nd ed.; Academic Press: San Diego, CA, USA, 2014; pp. 1–40.
2. Steimer, T. The biology of fear- and anxiety-related behaviors. *Dialogues Clin. Neurosci.* **2002**, *4*, 231–249.
3. Valente, T.S.; Baldi, F.; Sant'Anna, A.C.; Albuquerque, L.G.; Paranhos da Costa, M.J. Genome-wide association study between single nucleotide polymorphisms and flight speed in Nelore Cattle. *PLoS ONE.* **2016**, *11*, e0156956.
4. Costilla, R.; Kemper, K.E.; Byrne, E.M.; Porto-Neto, L.R.; Carvalheiro, R.; Purfield, D.C.; Doyle, J.L.; Berry, D.P.; Moore, S.S.; Wray, N.R.; et al. Genetic control of temperament traits across species: Association of autism spectrum disorder risk genes with cattle temperament. *Genet. Sel. Evol.* **2020**, *52*, 1–14.
5. Alvarenga, A.B.; Oliveira, H.R.; Miller, S.P.; Silva, F.F.; Brito, L.F. Genetic modeling and genomic analysis of yearling temperament in American Angus Cattle and its relationship with productive efficiency and resilience traits. *Front. Genet.* under review.
6. Cooke, R.F.; Moriel, P.; Cappellozza, B.I.; Miranda, V.F.B.; Batista, L.F.D.; Colombo, E.A.; Ferreira, V.S.M.; Miranda, M.F.; Marques, R.S.; Vasconcelos, J.L.M. Effects of temperament on growth, plasma cortisol concentrations and puberty attainment in Nelore beef heifers. *Animal* **2019**, *13*, 1208–1213.
7. By the Numbers: Docility Genetic Evaluation Research. Available online: <http://www.angus.org/nce/documents/bythenumbersdocility.pdf> (accessed on 12 August 2021).

8. Alvarenga, A.B.; Oliveira, H.R.; Chen, S.Y.; Miller, S.P.; Marchant-Forde, J.N.; Grigoletto, L.; Brito, L.F. A systematic review of genomic regions and candidate genes underlying behavioral traits in farmed mammals and their link with human disorders. *Animals* **2021**, *11*, 1–27.
9. Gabriel, S.B.; Schaffner, S.F.; Nguyen, H.; Moore, J.M.; Roy, J.; Blumenstiel, B.; Higgins, J.; DeFelice, M.; Lochner, A.; Faggart, M.; et al. The structure of haplotype blocks in the human genome. *Science* **2002**, *296*, 2225–2229.
10. Calus, M.P.L.; Meuwissen, T.H.E.; de Roos, A.P.W.; Veerkamp, R.F. Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **2008**, *178*, 553–561.
11. Villumsen, T.M.; Janss, L.; Lund, M.S. The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* **2009**, *126*, 3–13.
12. Hess, M.; Druet, T.; Hess, A.; Garrick, D. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* **2017**, *49*, 54.
13. Jiang, Y.; Schmidt, R.H.; Reif, J.C. Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3* **2018**, *8*, 1687–1699.
14. Liang, Z.; Tan, C.; Prakapenka, D.; Ma, L.; Da, Y. Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. *Front. Genet.* **2020**, *11*, 1–20.
15. Braz, C.U.; Taylor, J.F.; Bresolin, T.; Espigolan, R.; Feitosa, F.L.B.; Carvalheiro, R.; Baldi, F.; Albuquerque, L.G.; Oliveira, H.N. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. *BMC Genet.* **2019**, *20*, 1–12.
16. Bovo, S.; Ballan, M.; Schiavo, G.; Ribani, A.; Tinarelli, S.; Utzeri, V.J.; Dall'Olio, S.; Gallo, M.; Fontanesi, L. Single-marker and haplotype-based genome-wide association studies for the number of teats in two heavy pig breeds. *Anim. Genet.* **2021**, *52*, 440–450.
17. Martin, E.R.; Lai, E.H.; Gilbert, J.R.; Rogala, A.R.; Afshari, A.J.; Riley, J.; Finch, K.L.; Stevens, J.F.; Livak, K.J.; Slotterbeck, B.D.; et al. SNPing away at complex diseases: Analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **2000**, *67*, 383–394.
18. Araujo, A.C.; Carneiro, P.L.S.; Oliveira, H.R.; Schenkel, F.S.; Veroneze, R.; Lourenco, D.A.L.; Brito, L.F. A comprehensive comparison of haplotype-based single-step genomic predictions in livestock populations with different genetic diversity levels: A simulation study. *Front. Genet.* **2021**, *12*, 1–17.
19. Wang, H.; Misztal, I.; Aguilar, I.; Legarra, A.; Muir, W.M. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* **2012**, *94*, 73–83.
20. Mancin, E.; Lourenco, D.; Bermann, M.; Mantovani, R.; Misztal, I. Accounting for population structure and phenotypes from relatives in association mapping for farm animals: A simulation study. *Front. Genet.* **2021**, *12*, 1–14.
21. Legarra, A.; Aguilar, I.; Misztal, I. A Relationship Matrix Including Full Pedigree and Genomic Information. *J. Dairy Sci.* **2009**, *92*, 4656–4663.

22. Aguilar, I.; Misztal, I.; Johnson, D.L.; Legarra, A.; Tsuruta, S.; Lawlor, T.J. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* **2010**, *93*, 743–752.
23. Zhang, X.; Lourenco, D.; Aguilar, I.; Legarra, A.; Misztal, I. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* **2016**, *7*, 1–14.
24. Johnson, T.; Keehan, M.; Harland, C.; Lopdell, T.; Spelman, R.J.; Davis, S.R.; Rosen, B.D.; Smith, T.P.L.; Couldrey, C. Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. *J. Dairy Sci.* **2019**, *102*, 3254–3258.
25. Misztal, I.; Tsuruta, S.; Lourenco, D.A.L.; Masuda, Y.; Aguilar, I.; Legarra, A.; Vitezica, Z. *Manual for BLUPF90 Family Programs*; University of Georgia, Athens, GA, USA: 2018. Available online: <http://nce.ads.uga.edu/897/wiki/doku.php?id=documentation> (accessed on 12 June 2021).
26. Sargolzaei, M.; Chesnais, J.P.; Schenkel, F.S. A new approach for efficient genotype imputation using information from relatives. *BMC Genom.* **2014**, *15*, 478.
27. Hill, W.G.; Robertson, A. Linkage disequilibrium in finite populations. *Theoret. Appl. Genet.* **1968**, *38*, 226–231.
28. Kim, S.A.; Cho, C.-S.; Kim, S.-R.; Bull, S.B.; Yoo, Y.J. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics* **2018**, *34*, 388–397.
29. Kim, S.A.; Brossard, M.; Roshandel, D.; Paterson, A.D.; Bull, S.B.; Yoo, Y.J. gpart: Human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics* **2019**, *35*, 4419–4421.
30. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation For Statistical Computing. 2020. Available online: www.R-project.org/ (accessed on 10 June 2021).
31. Teissier, M.; Larroque, H.; Brito, L.F.; Rupp, R.; Schenkel, F.S.; Robert-Granié, C. Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. *J. Dairy Sci.* **2020**, *103*, 11559–11573.
32. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423.
33. Strandén, I.; Garrick, D.J. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* **2009**, *92*, 2971–2975.
34. Fragomeni, B.O.; Lourenco, D.A.L.; Legarra, A.; VanRaden, P.; Misztal, I. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. *J. Dairy Sci.* **2019**, *102*, 10012–10019.
35. Li, Y.; Sung, W.K.; Liu, J.J. Association mapping via regularized regression analysis of single-nucleotide polymorphism haplotypes in variable-sized sliding windows. *Am. J. Hum. Genet.* **2007**, *80*, 705–715.

36. Cullen, A.; Frey, H. *Probabilistic Techniques in Exposure Assessment*, 1st ed.; Plenum Publishing Co, Springer: New York, NY, USA, 1999.
37. Delignette-Muller, M.L.; Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*, 1–34.
38. Hu, Z.L.; Park, C.A.; Reecy, J.M. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* **2019**, *47*, D701–D710.
39. Medrano, J.F. The new bovine reference assembly and its value for genomic research. *Proc. Assoc. Advmt. Anim. Breed. Genet.* **2017**, *22*, 161–166.
40. Rosen, B.D.; Bickhart, D.M.; Schnabel, R.D.; Koren, S.; Elsik, C.G.; Zimin, A.; Dreischer, C.; Schultheiss, S.; Hall, R.; Schroeder, S.G.; et al. Modernizing the bovine reference genome assembly. *Proc. World Congr. Genet. Appl. Livest Prod.* **2018**, *3*, 802.
41. Aguilar, I.; Legarra, A.; Cardoso, F.; Masuda, Y.; Lourenco, D.; Misztal, I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. *Genet. Sel. Evol.* **2019**, *51*, 28.
42. Chen, S.Y.; Oliveira, H.O.; Schenkel, F.S.; Pedrosa, V.B.; Melka, M.G.; Brito, L.F. Using imputed whole-genome sequence variants to uncover candidate mutations and genes affecting milking speed and temperament in Holstein cattle. *J. Dairy Sci.* **2020**, *103*, 10383–10398.
43. Mehrban, H.; Naserkheil, M.; Lee, D.H.; Cho, C.; Choi, T.; Park, M.; Ibáñez-Escriche, N. Genomic prediction using alternative strategies of weighted single-step genomic BLUP for yearling weight and carcass traits in Hanwoo beef cattle. *Genes* **2021**, *12*, 266.
44. Suchocki, T.; Szyda, J. Genome-wide association study for semen production traits in Holstein-Friesian bulls. *J. Dairy Sci.* **2015**, *98*, 5774–5780.
45. Riley, D.G.; Gill, C.A.; Boldt, C.R.; Funkhouser, R.R.; Herring, A.D.; Riggs, P.K.; Sawyer J.E.; Lunt, D.K.; Sanders, J.O. Crossbred Bos indicus steer temperament as yearlings and whole genome association of steer temperament as yearlings and calf temperament post-weaning. *J. Anim. Sci.* **2016**, *94*, 1408–1414.
46. Michenet, A.; Saintilan, R.; Venot, E.; Phocas, F. Insights into the genetic variation of maternal 1187 behavior and suckling performance of continental beef cows. *Genet. Sel. Evol.* **2016**, *48*, 1–12.
47. Dreher, C.; Wellmann, R.; Stratz, P.; Schmid, M.; Preuß, S.; Hamann, H.; Bennewitz, J. Genomic analysis of perinatal sucking reflex in German Brown Swiss calves. *J. Dairy Sci.* **2019**, *102*, 6296–6305.
48. Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2001**, *2*, 919–929.
49. Taye, M.; Lee, W.; Jeo, S.; Yoon, J.; Dessie, T.; Hanotte, O.; Mwai, O.A.; Kemp, S.; Cho, S.; Oh, S.J.; et al. Exploring evidence of positive selection signatures in cattle breeds selected for different traits. *Mamm. Genome* **2017**, *28*, 528–541.

50. Brito, L.F.; Oliveira, H.R.; McConn, B.R.; Schinckel, A.P.; Arrazola, A.; Marchant-Forde, J.N.; Johnson, J.S. Large-scale phenotyping of livestock welfare in commercial production systems: A new frontier in animal breeding. *Front. Genet.* **2020**, *11*, 793.
51. Cheng, H.W. Breeding of tomorrow's chickens to improve well-being. *Poult. Sci.* **2010**, *89*, 805–813.
52. Zhang, H.; Liu, A.; Wang, Y.; Luo, H.; Yan, X.; Guo, X.; Li, X.; Liu, L.; Su, G. Genetic parameters and genome-wide association studies of eight longevity traits representing either full or partial lifespan in Chinese Holsteins. *Front. Genet.* **2021**, *12*, 634986.
53. Oliveira, H.R.; Brito, L.F.; Miller, S.P.; Schenkel, F.S. Using random regression models to genetically evaluate functional longevity traits in North American angus cattle. *Animals* **2020**, *10*, 1–30.
54. Pereira Vatanabe, I.; Peron, R.; Mantellatto Grigoli, M.; Pelucchi, S.; De Cesare, G.; Magalhães, T.; Manzine, P.R.; Figueredo Balthazar, M.L.; Di Luca, M.; Marcello, E.; et al. ADAM10 plasma and CSF levels are increased in mild Alzheimer's disease. *Int. J. Mol. Sci.* **2021**, *22*, 2416.
55. Sollero, B.P.; Junqueira, V.S.; Gomes, C.C.G.; Caetano, A.R.; Cardoso, F.F. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. *Genet. Sel. Evol.* **2017**, *49*, 49.
56. Kasarapu, P.; Porto-Neto, L.R.; Fortes, M.R.S.; Lehnert, S.A.; Mudadu, M.A.; Coutinho, L.; Regitano, L.; George, A.; Reverter, A. The *Bos taurus-Bos indicus* balance in fertility and milk related genes. *PLoS ONE* **2017**, *12*, e0181930.
57. Silva, R.P.; Berton, M.P.; Grigoletto, L.; Carvalho, F.E.; Silva, R.M.O.; Peripolli, E.; Castro, L.M.; Ferraz, J.B.S.; Eler, J.P.; Lobo, R.B.; et al. Genomic regions and enrichment analyses associated with carcass composition indicator traits in Nellore cattle. *J. Anim. Breed. Genet.* **2018**, *136*, 1–16.
58. LaMantia, A.-S. Why does the face predict the brain? Neural crest induction, craniofacial morphogenesis, and neural circuit development. *Front. Physiol.* **2020**, *11*, 610970.
59. Carre, J.M.; McCormick, C.M.; Mondloch, C.J. Facial structure is a reliable cue of aggressive behavior. *Psychol. Sci.* **2009**, *20*, 1194–1198.
60. Li, C.; Sun, D.; Zhang, S.; Wang, S.; Wu, X.; Zhang, Q.; Liu, L.; Li, Y.; Qiao, L. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. *PLoS ONE* **2014**, *9*, e96186.
61. Grandin, T.; Deesing, M.J.; Struthers, J.J.; Swinker, A.M. Cattle with hair whorl patterns above the eyes are more behaviorally agitated during restraint. *Appl. Anim. Behav. Sci.* **1995**, *46*, 117–123.
62. Lanier, J.L.; Grandin, T.; Green, R.D.; Avery, D.; Mcgee, K. Cattle hair whorl position and temperament in auction houses. *J. Anim. Sci.* **1999**, *77*, 147.
63. Lanier, J.L.; Grandin, T.; Green, R.D.; Avery, D.; Mcgee, K. A note on hair whorl position and cattle temperament in the auction ring. *Appl. Anim. Behav. Sci.* **2001**, *73*, 93–101.

64. Furdon, S.A.; Clark, D.A. Scalp hair characteristics in the newborn infant. *Adv. Neonatal Care* **2003**, *3*, 286–296.
65. Lima, D.F.P.d.A.; da Cruz, V.A.R.; Pereira, G.L.; Curi, R.A.; Costa, R.B.; de Camargo, G.M.F. Genomic Regions Associated with the Position and Number of Hair Whorls in Horses. *Animals* **2021**, *11*, 2925.
66. Anilkumar, S.; Patel, D.; Boer, S.F.; Chattarji, S.; Buwalda, B. Decreased dendritic spine density in poster dorsal medial amygdala neurons of proactive coping rats. *Behav. Brain Res.* **2021**, *397*, 112940.
67. Neupane, M.; Kiser, J.N.; The Bovine Respiratory Disease Complex Coordinated Agricultural Project Research Team; Neibergs, H.L. Gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease. *Anim. Genet.* **2018**, *49*, 527–538.
68. Hay, E.L.; Roberts, A. Genome-wide association study for carcass traits in a composite beef cattle breed. *Livest. Sci.* **2018**, *213*, 35–43.
69. Bonnefil, V.; Dietz, K.; Amatruda, M.; Wentling, M.; Aubry, A.V.; Dupree, J.L.; Temple, G.; Park, H.J.; Burghardt, N.S.; Casaccia, P. Region-specific myelin differences define behavioral consequences of chronic social defeat stress in mice. *eLife* **2019**, *8*, e40855.
70. Hartline, D.K. What is myelin? *Neuron Glia Biol.* **2008**, *4*, 153–163.
71. Huson, H.J.; Kim, E.S.; Godfrey, R.W.; Olson, T.A.; McClure, M.C.; Chase, C.C.; Rizzi, R.; O'Brien, A.M.P.; VanTassell, C.P.; Garcia, J.F. Genome-wide association study and ancestral origins of the slick-hair coat in tropically adapted cattle. *Front. Genet.* **2014**, *5*, 1–12.
72. Sweett, H.; Fonseca, P.A.S.; Suárez-Vega, A.; Livernois, A.; Miglior, F.; Cánovas, A. Genome-wide association study to identify genomic regions and positional candidate genes associated with male fertility in beef cattle. *Sci. Rep.* **2020**, *10*, 20102.
73. Manuck, S.B.; Schaefer, D.C. Stability of individual differences in cardiovascular reactivity. *Physiol. Behav.* **1978**, *21*, 675–678.
74. Carnevali, L.; Nalivaiko, E.; Sgoifo, A. Respiratory patterns reflect different levels of aggressiveness and emotionality in Wild-type Groningen rats. *Respir. Physiol. Neurobiol.* **2014**, *204*, 28–35.
75. Falkner, A.L.; Wei, D.; Song, A.; Watsek, L.W.; Chen, I.; Chen, P.; Feng, J.; Lin, D. Hierarchical Representations of Aggression in a Hypothalamic-Midbrain Circuit. *Neuron* **2020**, *106*, 637–648.
76. Yin, H.; Zhou, C.; Shi, S.; Fang, L.; Liu, J.; Sun, D.; Jiang, L.; Zhang, S. Weighted single-step genome-wide association study of semen traits in Holstein bulls of China. *Front. Genet.* **2019**, *10*, 1053.
77. Imumorin, I.G.; Kim, E.H.; Lee, Y.M.; De Koning, D.J.; van Arendonk, J.A.; Donato, M.D.; Taylor, J.F.; Kim, J.J. Genome scan for parent-of-origin QTL effects on bovine growth and carcass traits. *Front. Genet.* **2011**, *2*, 44.
78. Bandler, R.; Keay, K.A. Columnar organization in the midbrain periaqueductal gray and the integration of emotional expression. *Prog. Brain Res.* **1996**, *107*, 285–300.
79. Fineberg, S.K.; Ross, D.A. Oxytocin and the Social Brain. *Biol. Psychiatry* **2017**, *81*, e19–e21.

80. Hudson, N.J.; Reverter, A.; Greenwood, P.L.; Guo, B.; Café, L.M.; Dalrymple, B.P. Longitudinal muscle gene expression patterns associated with differential intramuscular fat in cattle. *Animal* **2015**, *9*, 650–659.
81. Roudbari, Z.; Coort, S.L.; Kutmon, M.; Eijssen, L.; Melius, J.; Sadkowski, T.; Evelo, C.T. Identification of biological pathways contributing to marbling in skeletal muscle to improve beef cattle breeding. *Front. Genet.* **2020**, *10*, 1370.
82. Sewalem, A.; Miglior, F.; Kistemaker, G.J. Short communication: Genetic parameters of milking temperament and milking speed in Canadian Holsteins. *J. Dairy Sci.* **2011**, *94*, 512–516.
83. Guo, Y.; Li, J.; Bonham, A.J.; Wang, Y.; Deng, H. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: A comparison of association-mapping strategies. *Eur. J. Hum. Genet.* **2009**, *17*, 785–792.
84. Zhao, H.G.; Pfeiffer, R.; Gail, M.H. Haplotype analysis in population genetics and association studies. *Pharmacogenomics* **2003**, *4*, 171–178.
85. Hayes, B.J.; Chamberlain, A.J.; McPartlan, H.; Macleod, I.; Sethuraman, L.; Goddard, M.E. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet. Res.* **2007**, *89*, 215–220.
86. Calus, M.P.; Meuwissen, T.H.; Windig, J.J.; Knol, E.F.; Schrooten, C.; Vereijken, A.L.; Veerkamp, R.F. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.* **2009**, *41*, 11.
87. Weng, Z.; Wolc, A.; Su, H.; Fernando, R.L.; Dekkers, J.C.M.; Arango, J.; Settar, P.; Fulton, J.E.; O'Sullivan, N.P.; Garrick, D.J. Identification of recombination hotspots and quantitative trait loci for recombination rate in layer chickens. *J. Anim. Sci. Biotechnol.* **2019**, *10*, 20.
88. Weng, Z.-Q.; Saatchi, M.; Schnabel, R.D.; Taylor, J.F.; Garrick, D.J. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet. Sel. Evol.* **2014**, *46*, 34.
89. Khatkar, M.S.; Zenger, K.R.; Hobbs, M.; Hawken, R.J.; Cavanagh, J.A.L.; Barris, W.; McClintock, A.E.; McClintock, S.; Thomson, P.T.; Tier, B.; et al. A Primary Assembly of a Bovine Haplotype Block Map Based on a 15,036-Single-Nucleotide Polymorphism Panel Genotyped in Holstein–Friesian Cattle. *Genetics* **2007**, *176*, 763–772.

V – CAPÍTULO III

Artigo no formato da revista *Journal of Animal Breeding and Genetics*

**Haplotype-based single-step genomic predictions for body weight, wool, and
reproductive traits in North American Rambouillet sheep**

Andre C. Araujo^{1,2}, Paulo L. S. Carneiro³, Hinayah R. Oliveira², Ronald M. Lewis^{4,a},

Luiz F. Brito^{2, a, *}

¹Graduate Program in Animal Sciences, State University of Southwestern Bahia, Itapetinga, BA, Brazil

²Department of Animal Sciences, Purdue University, West Lafayette, IN, USA

³Department of Biology, State University of Southwestern Bahia, Jequié, BA, Brazil

⁴Department of Animal Sciences. University of Nebraska-Lincoln, Lincoln, NE, USA

^aThese authors share the last authorship

***Corresponding author:** Luiz F. Brito

Address: Department of Animal Sciences, Purdue University, West Lafayette, 47907, IN, USA

E-mail: britol@purdue.edu

Phone number: +1 765 586 2515

Abstract

Rambouillet sheep are commonly raised in extensive grazing systems in the US, mainly for wool and meat production. Genomic evaluations in US sheep breeds, including Rambouillet, are still incipient. Therefore, we aimed to evaluate the feasibility of performing genomic prediction of breeding values for various traits in Rambouillet sheep based on single nucleotide polymorphisms (SNP) or haplotypes (fitted as pseudo-SNP) under a single-step GBLUP approach. Approximately 5K to 28K records for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB), were used for this study. A total of 741 individuals were genotyped using a moderate (50K; $n = 677$) or high (600K; $n = 64$) density SNP panel, in which 32K SNP in common between the two SNP panels (after genotypic quality control) were used for further analyses. Single-step genomic predictions using SNP (H-BLUP) or haplotypes (HAP-BLUP) from blocks with different linkage disequilibrium (LD) thresholds (0.15, 0.35, 0.50, 0.65, and 0.80) were evaluated. We also considered genomic weights (alpha parameter) equal to 0.95 or 0.50 when constructing the genomic relationship matrix used to predict the genomic enhanced estimated breeding values (GEBV). The GEBV were compared to the estimated breeding values (EBV) obtained from traditional pedigree-based evaluations (A-BLUP). The mean theoretical accuracy ranged from 0.499 (A-BLUP for PWT) to 0.795 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.35 and alpha equal to 0.95 for YFD). The prediction accuracies ranged from 0.143 (A-BLUP for PWT) to 0.330 (A-BLUP for YGFW) while the prediction bias ranged from -0.104 (H-BLUP for PWT) to 0.087 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.15 and alpha equal to 0.95 for YGFW). The GEBV dispersion ranged from 0.428 (A-BLUP for PWT) to 1.035 (A-BLUP for YGDW). Using genomic information from SNP or haplotypes provided similar or higher prediction and theoretical accuracies and reduced the dispersion of the GEBV for body weight,

wool, and reproductive traits in Rambouillet sheep. However, there was no clear improvements in the prediction bias when compared to pedigree-based predictions. The next step will be to enlarge the training populations for this breed to increase the benefits of genomic predictions.

Keywords: BLUP, genomically enhanced estimated breeding values, haplotype prediction, linkage disequilibrium, small ruminants, ssGBLUP

1. Introduction

The global demand for products from small ruminants is increasing. Further scientific innovation, with its greater application through increased education and training, is needed to meet this demand (Mazinani & Rude, 2020). Rambouillet sheep are commonly used in extensive grazing systems. They produce heavy fleeces with fine fiber diameter (Thorne et al., 2021). Yet, as a dual-purpose breed, body weight and reproductive traits are also of economic importance in this breed (Thorne et al., 2021). Estimated breeding values (EBV) for these sets of traits have been generated and shared with US sheep producers through the National Sheep Improvement Program (NSIP; Notter, 1998). However, no genomic estimated breeding values (GEBV) have been calculated for Rambouillet sheep yet.

The original application of the mixed model equations (MME) in animal genetics was to obtain solutions for fixed and random effects while avoiding the need to invert the full covariance matrix of the data (Henderson, 1950). In this context, the MME were used to obtain Best Linear Unbiased Prediction (BLUP) of EBV for selection candidates, based on the relatedness among individuals from pedigree (Henderson, 1984). However, with the availability of large-scale genomic information, the pedigree relationship matrix (**A**) can be replaced or combined with the genomic relationship matrix (**G**) to predict GEBV (Aguilar et al., 2010). The GEBV can be more accurate than EBV especially for young animals (not yet recorded for the

traits of interest), and for lowly heritable and sex-limited traits (Meuwissen et al., 2001). Furthermore, GEBV can provide advantages for the evaluation of difficult- or expensive-to-measure traits (Brito et al., 2020; Thorne et al., 2021).

The single-step genomic BLUP (ssGBLUP; Legarra et al., 2009; Christensen & Lund, 2010) is a method that simultaneously includes both genotyped and non-genotyped individuals in the analysis to obtain GEBV for all individuals by combining the genomic and pedigree information. The ssGBLUP is more compatible with current breeding programs (where not all breeding individuals are genotyped) and provides similar or better results than other methods (Legarra et al., 2014; Guarini et al., 2018). However, an important consideration when implementing the ssGBLUP is how to weight the genomic and pedigree information (McMillan & Swan, 2017; Meyer et al., 2018). This conundrum arises because as \mathbf{G} computes the relationships at the genomic marker level, it can be difficult to invert, may not be on the same scale as the \mathbf{A} , and may not account for residual polygenic effects (Meyer et al., 2018).

To help during the inversion process, and to account for residual polygenic effects, two parameters, α and β (with $\alpha = 0, \dots, 1$ and $\beta = 1 - \alpha$), are commonly used to include a proportion of \mathbf{A} in the \mathbf{G} that is used in the genomic evaluation (Meyer et al., 2018). Values between 0.95 to 0.99 are common choices to weight \mathbf{G} (McMillan & Swan, 2017). However, some authors (McMillan & Swan, 2017; Gao et al., 2012) showed that different α can affect the accuracy and bias of the single-step genomic predictions. McMillan and Swan (2017) used $\alpha = 0.50$ to place equal emphasis on the pedigree and genomic relationships for animals when both were recorded. Defining the appropriate value for these parameters is therefore important as they may differ even for different traits in the same population (Gao et al., 2012).

The \mathbf{G} matrix used in the ssGBLUP can also be computed based on different methods. Fitting single nucleotide polymorphisms (SNP) has been the standard method used in genomic analyses; however, haplotypes can also be used for both genomic prediction (Teissier et al.,

2020; Feitosa et al., 2020; Araujo et al., 2021) and genome-wide association (Bovo et al., 2021; Feitosa et al., 2021; Araujo et al., 2022) studies. Haplotypes are the alleles from a set of adjacent loci (sizable regions called haplotype blocks or haploblocks) expected to be inherited together due to lower recombination (Gabriel et al., 2002). Haplotypes are also expected to be in higher linkage disequilibrium (LD) with the quantitative trait loci (QTL) than the single SNP (Calus et al., 2008) and capture epistatic effects (Hess et al., 2017; Jiang et al., 2018), which could result in higher accuracies and lower bias in the genomic predictions (Calus et al., 2008; Araujo et al., 2021).

Araujo et al. (2021) hypothesized that fitting haplotypes in genomic predictions could outperform the use of SNP in populations with high effective population size (N_e) because it would better capture the complex interactions within haploblocks; however, those authors did not simulate epistasis and recommended new studies in such populations. Sheep is a species in which moderate to high N_e are common in some commercial populations (Kijas et al., 2012; Brito et al., 2017a) with predictions of GEBV based on haplotypes scarce (Araujo et al., 2021). Therefore, we aimed to evaluate the GEBV accuracies, bias, dispersion, and individual theoretical accuracies (TA) using ssGBLUP fitting SNP or haplotypes for body weight, wool, and reproductive traits in Rambouillet sheep. We also evaluated the effect of constructing the haplotypes with different LD thresholds and α values when forming the **G** matrix. Finally, recommendations for future steps for the implementation of genomic evaluations in Rambouillet sheep were also provided.

2. Material and Methods

No ethical review and approval were needed for this study because all datasets used were provided by commercial breeding operations.

2.1 Phenotypic and pedigree data

The phenotypic datasets were provided by the NSIP (Ames, IA, USA), which included three body weight traits [birth weight (BWT), post-weaning weight (PWT), and yearling weight (YWT)], two wool traits [yearling fiber diameter (YFD) and yearling greasy fleece weight (YGFW)] and one reproductive trait [number of lambs born (NLB)] as described in Table 1. The BWT was the lamb weight recorded within 24 h after birth, while PWT and YWT were the body weights recorded at five to 10 (151 to 304 days) and 10 to 14 months of age, respectively. The wool traits were measured at yearling age (10 to 14 months). The pedigree dataset had 36,297 individuals born from 1985 to 2021, spanning up to 15 generations from animals with phenotypic records.

The phenotypic datasets used to make the genetic evaluation for the body weight and wool traits were processed previously by the NSIP, which provided pre-adjusted phenotypes (<http://nsip.org/wp-content/uploads/2015/04/Lambplan-TC-Report-Notter.pdf>). Briefly, the pre-adjustment considered birth and rearing type (fixed levels as a multiplicative adjustment), and age of dam at recording (fixed covariates as quadratic and fixed regressions, respectively). No pre-adjustments were done for NLB. The data underwent quality control (QC) with observations deviating more than three standard deviations from the mean removed from further analyses.

Contemporary groups (CG) were created by concatenating flock, year, season, management group, sex, recording date, and 70-day age groups to split lambing (birth) dates into 70 days periods, for PWT, YWT, YFD, and YGFW. For BWT, the CG included all the effects previously mentioned for body weight and wool traits, excluding recording date; 35- rather than 70-day age group were also used. The CG for the NLB were created considering ewe's flock, birth year, season, management group, and parturition number (e.g., ewe's first, second or third lambing). The pre-adjusted phenotypes for body weight and wool traits and the

NLB phenotypes were then adjusted for the CG effect, so that the phenotypes analyzed—henceforth referred to as corrected phenotypes—accounted for all systematic environmental effects considered in the NSIP genetic evaluations. As a final QC step, CG with less than three animals and with no phenotypic variability within CG were removed.

Table 1. Description of the datasets used for the genetic and genomic predictions of birth weight (BWT), post-weaning body weight (PWT), yearling body weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep.

Dataset	Variable ¹	Trait					
		BWT (kg)	PWT (kg)	YWT (kg)	YFD (μm)	YGFW (kg)	NLB (count)
Complete	Average	4.86	35.36	59.99	18.87	3.04	1.71
	SD	1.02	7.78	15.28	1.67	0.81	0.59
	Individuals (n)	28,317	22,781	5,653	9,586	11,542	6,846
	Records (n)	28,317	22,781	5,653	9,586	11,542	15,904
	CG (n)	427	461	110	195	210	445
	Genotypes (n)	587	632	442	529	502	242
Partial	Average	4.77	35.01	58.03	18.92	2.99	1.79
	SD	0.98	7.12	13.54	1.57	0.72	0.59
	Individuals (n)	22,115	17,118	3,657	6,808	9,055	5,407
	Records (n)	22,115	17,118	3,657	6,808	9,055	13,790
	CG (n)	404	402	101	175	198	411
	Genotypes (n)	469	456	341	426	402	138
	n focal	118	176	101	103	100	104

¹Standard deviation (SD); number of phenotyped individuals (Individuals), records (Records), contemporary groups (CG), and genotypes (Genotypes) included in the whole and partial datasets (after quality control); and number of focal individuals (n focal). All genotyped and focal animals had own phenotypes or progeny with phenotypes. The whole data set contained all corrected phenotypes after quality control and the partial data set was a subset of the whole data truncated by the birth year of the focal individuals (young selection candidates used to compare the methods evaluated).

2.2 Genotypic data

Samples from 677 and 64 animals were genotyped using the GeneSeek Genomic Profiler Ovine 50K array (Neogen Corporation, Lansing, MI, USA) (52,260 SNP) and OvineHD BeadChips (Illumina Inc., San Diego, CA, USA) (606,006 SNP) SNP panels by Neogen (GeneSeek, A Neogen Company, Lincoln, NE, USA). These individuals were chosen to be genotyped based on pedigree-based relatedness to try to capture the genetic diversity as much possible in animals with DNA samples, coming from nine representative NSIP Rambouillet flocks. As an additional criterion, animals with phenotypic information, either on themselves or on their progeny, for most traits analyzed were prioritized for genotyping.

Approximately 35K (35,105) autosomal SNP were in common between the two panels and were used in the QC. The QC for the genotypic data was done using the PLINK 1.9 software (Purcell et al., 2007), and markers with MAF < 0.05, call rate < 0.90, extreme departure from Hardy-Weinberg equilibrium ($P < 10^{-8}$), located on non-autosomal chromosomes, and duplicated SNP were removed. Samples with call rate < 0.90 were also removed. A total of 32,584 SNP and 722 samples remained for further analyses.

2.3 Haplotype construction

The SNP genotypes for all samples were phased using the FImpute v.3.0 software (Sargolzaei et al., 2014) to infer the parental inheritance (i.e., which allele came from which parent), before creating the haplotype blocks. LD haploblocks were constructed using the r^2 metric (Hill and Robertson, 1968) with the thresholds of 0.15, 0.35, 0.50, 0.65, and 0.80 using the Big-LD approach (Kim et al., 2018). The “gpart” package (Kim et al., 2019) implemented in R (R Core Team, 2020) was used to build the haploblocks.

2.4 Genetic evaluation

2.4.1 Pedigree-based predictions

Three linear mixed models for the pedigree based BLUP (A-BLUP) were used in this study, which are defined as follow:

$$\mathbf{y}_1 = \mathbf{1}'\boldsymbol{\mu} + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (1)$$

$$\mathbf{y}_2 = \mathbf{1}'\boldsymbol{\mu} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e} \quad (2)$$

$$\mathbf{y}_3 = \mathbf{1}'\boldsymbol{\mu} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{S}\mathbf{q} + \mathbf{e} \quad (3)$$

where the model (1) is an additive genetic model with \mathbf{y}_1 representing a vector of single corrected phenotypic records, $\boldsymbol{\mu}$ is the overall mean, \mathbf{u} is the random direct additive genetic effect, and \mathbf{e} is the random residual. The model (2) is a repeatability model, in which \mathbf{p} is the random permanent environment effect, \mathbf{y}_2 contains the repeated corrected phenotypic records, and the other vectors are the same as in model (1). The model (3) also includes the random maternal additive genetic and maternal permanent environment effects, \mathbf{m} and \mathbf{q} , respectively. The $\mathbf{1}'$ is a vector of ones used to calculate the overall mean and \mathbf{Z} , \mathbf{W} , \mathbf{Z}_2 , and \mathbf{S} are the incidence matrices that relates the corrected phenotypic records to the random direct additive genetic, permanent environment, maternal additive genetic, and maternal permanent environment effects, respectively. The random effects for the above models were assumed to be normally distributed with (co)variance structures as follows:

$$\text{Model (1): } \text{Var} \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_u^2 & 0 \\ 0 & \mathbf{I}\sigma_e^2 \end{bmatrix} \quad (4)$$

$$\text{Model (2): } \text{Var} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_u^2 & 0 & 0 \\ 0 & \mathbf{I}\sigma_p^2 & 0 \\ 0 & 0 & \mathbf{I}\sigma_e^2 \end{bmatrix} \quad (5)$$

$$\text{Model (3): } \text{Var} \begin{bmatrix} \mathbf{u} \\ \mathbf{m} \\ \mathbf{q} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_u^2 & 0 & 0 & 0 \\ 0 & \mathbf{A}\sigma_m^2 & 0 & 0 \\ 0 & 0 & \mathbf{I}\sigma_q^2 & 0 \\ 0 & 0 & 0 & \mathbf{I}\sigma_e^2 \end{bmatrix} \quad (6)$$

where model (1) was used to make the EBV prediction for the YFD, model (2) for NLB, and model (3) for BWT, PWT, YWT, and YGFW.

The BLUPf90 software (Misztal et al., 2018) was used to predict EBV assuming the variance components were known (Table 2). To be consistent with the national genetic evaluation underway in Rambouillet sheep, the models fitted and the variance components used to predict the EBV were provided by NSIP.

2.4.2 Single-step genomic BLUP using SNP

The corrected phenotypes, models, and variance components used to predict the GEBV under the single-step genomic BLUP using SNP (H-BLUP) approach were similar to the ones used in A-BLUP, except for the inclusion of genomic relationships from the genotyped samples. In the assumptions of the H-BLUP, the \mathbf{y} vector had corrected phenotypes for genotyped and non-genotyped animals and $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$. \mathbf{H} is the matrix that combines the pedigree and the genomic relationship matrices (Legarra et al., 2009), with its inverse computed as follows (Aguilar et al., 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix} \quad (7)$$

Table 2. Variance components and genetic parameters used to predict the estimated breeding values for birth weight (BWT), post-weaning body weight (PWT), yearling body weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep.

Parameter ¹	BWT	PWT	YWT	YFD	YGFW	NLB
$\sigma_{\mathbf{u}}^2$	0.085	3.211	15.402	1.311	0.122	0.025
$\sigma_{\mathbf{p}}^2$	-	-	-	-	-	0.009
$\sigma_{\mathbf{m}}^2$	0.091	1.926	1.777	-	0.013	-
$\sigma_{\mathbf{q}}^2$	0.061	1.926	1.777	-	0.013	-
$\sigma_{\mathbf{e}}^2$	0.372	25.046	40.283	0.989	0.181	0.250
$\sigma_{\mathbf{p}}^2$	0.610	32.110	59.240	2.300	0.330	0.284
h^2	0.140	0.100	0.260	0.570	0.370	0.090
p^2	-	-	-	-	-	0.030
h_m^2	0.150	0.060	0.030	-	0.040	-
c^2	0.100	0.060	0.030	-	0.040	-

¹ $\sigma_{\mathbf{u}}^2$ = additive genetic variance, $\sigma_{\mathbf{p}}^2$ = permanent environment variance associated with repeated records, $\sigma_{\mathbf{m}}^2$ = maternal additive genetic variance, $\sigma_{\mathbf{q}}^2$ = maternal permanent environment variance, $\sigma_{\mathbf{e}}^2$ = residual variance, $\sigma_{\mathbf{p}}^2$ = phenotypic variance, h^2 = heritability for the direct additive genetic effect, p^2 = repeatability, h_m^2 = heritability for the maternal additive genetic effect, c^2 = fraction of the phenotypic variance explained by the maternal permanent environment effect.

where \mathbf{A}^{-1} is the inverse of the pedigree relationship matrix, \mathbf{A}_{22} and \mathbf{A}_{22}^{-1} are the pedigree relationship matrix for the genotyped animals and its inverse, respectively, and \mathbf{G} is the genomic relationship matrix calculated as proposed by VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{2 \sum p_i(1-p_i)} \quad (8)$$

where \mathbf{M} has the dimension of n genotyped animals by m SNP markers (coded as 0, 1, or 2 for the absence, presence of one copy, or presence of two copies of the reference allele, respectively) and is centered based on twice of the allelic frequencies (p_i ; $1 - p_i$). The PREGSf90 software (Misztal et al., 2018) was used to create the \mathbf{H}^{-1} matrix, with τ and ω parameters assumed as the default values (1.0). Different values of α were used to evaluate the impact of increasing the proportion of \mathbf{A}_{22} on \mathbf{G} , with $\alpha = 0.95$ and 0.50 ($\beta = 0.05$ and 0.50 , respectively), in which the former is the default value in the PREGSf90 software and the later was suggested as the choice in Terminal Sire sheep populations (McMillan & Swan, 2017).

2.4.3 Single-step genomic BLUP using haplotypes

The model and assumptions used in the HAP-BLUP approach were similar to those described for H-BLUP. However, the \mathbf{G} used in the $(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1}$ component was constructed using non-LD-clustered SNP (NCSNP) and pseudo-SNP (ps-SNP). A ps-SNP corresponds to one of the unique haplotype alleles present within a haploblock, coded as 0, 1, or 2 to account for the number of copies of the reference haplotype allele, similar to Araujo et al. (2021). As a haploblock can be multi allelic, several ps-SNP can be created from a haploblock. The ps-SNP were subjected to the same QC criteria as the SNP before their use for genomic prediction. The number of NCSNP plus ps-SNP before QC ranged between 33,922 and 44,695 with the LD thresholds of 0.80 and 0.15, respectively, while the number of NCSNP plus ps-SNP after QC (markers used in the haplotype predictions) ranged from 32,649 to 39,787 with these same LD thresholds (Supplementary File 1). All the scenarios regarding the different combinations of α and β parameters described for H-GBLUP were also tested in the HAP-BLUP method.

2.5 Comparing genetic and genomic predictions

The whole and partial datasets used to compare the genetic and genomic predictions (Legarra & Reverter, 2018) for each trait (Table 1) were defined separately based on time thresholds considering the birth date of the genotyped animals as the reference. The whole datasets included all corrected phenotypic records and genotyped individuals with corrected phenotypes on itself or in its progeny. As the number of genotyped individuals was small, the division into partial datasets considered the following two criteria: 1) at least 100 genotyped individuals with average EBV accuracy higher than 0.50 as focal individuals (selection candidates with masked corrected phenotypes in itself and in its progeny) were kept; and 2) at least 20% of the genotyped individuals as focal individuals were kept.

The performance of genetic and genomic predictions was evaluated using the linear regression (LR) method as described by Legarra and Reverter (2018). The LR method provides a series of statistics derived from the comparison of genetic evaluations using the whole and partial datasets, resulting in easy-to-use methods to evaluate the reliability of the predictions (Legarra & Reverter, 2018). The LR statistics obtained were:

$$Accuracy = \sqrt{\frac{cov((G)EBV_W, (G)EBV_P)}{(1-\bar{F})\sigma_{u_d}^2}} \quad (9)$$

$$Bias = ave(\widehat{\mathbf{u}}_P) - ave(\widehat{\mathbf{u}}_W) \quad (10)$$

$$Dispersion = \left(\frac{cov((G)EBV_W, (G)EBV_P)}{var((G)EBV_P)} \right) - 1 \quad (11)$$

where $cov((G)EBV_W, (G)EBV_P)$ is the covariance between the GEBV or EBV in whole ($(G)EBV_W$) and partial ($(G)EBV_P$) datasets, \bar{F} is the average inbreeding, $ave()$ represent the arithmetic average function, $\widehat{\mathbf{u}}_W$ and $\widehat{\mathbf{u}}_P$ are the predicted GEBV or EBV in the whole and partial datasets, respectively, and $var((G)EBV_P)$ is the variance of the GEBV or EBV. The other components were previously described.

In addition to the LR statistics, the individual theoretical accuracies (TA) were calculated for the focal individuals according to Van Vleck (1993):

$$TA = \sqrt{1 - \frac{se_i^2}{(1+f_i)\sigma_u^2}} \quad (12)$$

where TA is the individual theoretical accuracies, se_i^2 is the square of the GEBV or EBV standard error for the individual i , f_i is the inbreeding coefficient for the individual i , and the other variables were previously described.

2.6 Evaluated scenarios

The scenarios consisted of combinations of 1) A-BLUP, 2) H-BLUP with $\alpha = 0.95$ and 0.50 to construct \mathbf{G} , and 3) HAP-BLUP using ps-SNP from different LD thresholds (0.15, 0.35, 0.50, 0.65, and 0.80) also with $\alpha = 0.95$ and 0.50 to construct \mathbf{G} . In total, 13 scenarios were evaluated for each of the six traits, resulting in 78 analyses.

3. Results

3.1 Accuracies

The prediction accuracies for body weight, wool, and NLB traits ranged between 0.143 (A-BLUP for PWT) to 0.330 (A-BLUP for YGFW). The lowest and highest prediction accuracies were observed for NLB and wool (both YFD and YGFW) traits, respectively. Similar prediction accuracies were observed for the HAP-BLUP across different LD thresholds, regardless of the α value and trait evaluated. We, therefore, only present the results for the HAP-BLUP considering the LD threshold of 0.50 (HAP-BLUP-LD_0.50) to compare the predictions between pedigree, SNP, and haplotype-based methods for all traits (Figure 1).

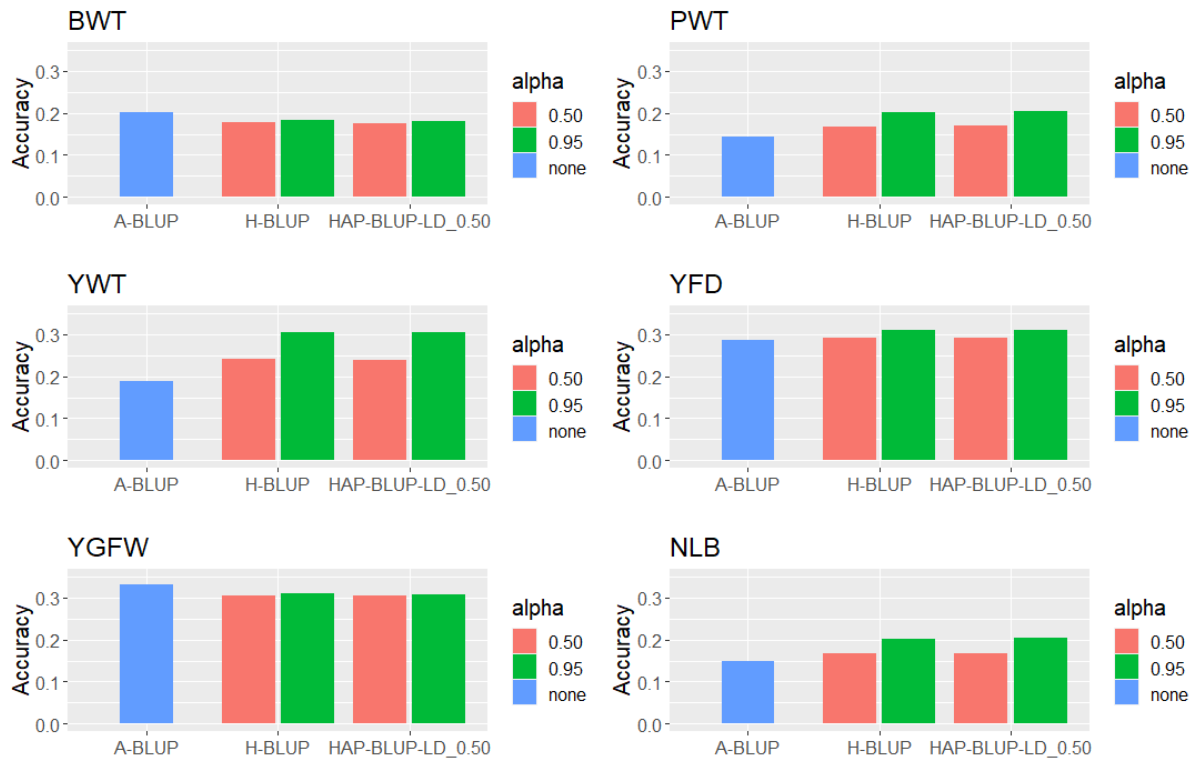


Figure 1. Prediction accuracies for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLB) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP fitting pseudo-haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different α values (0.95 or 0.50) were used to create the genomic relationship matrices.

Using genomic information provided similar or higher GEBV prediction accuracies compared to EBVs. An increase of ~41% (~0.06), ~62% (~0.12), ~8% (~0.02), and ~37% (~0.05) in the GEBV prediction accuracies was observed for the PWT, YWT, YFD, and NLB, respectively, when using α equal to 0.95. Using an α of 0.50 generally resulted in half of the increase in the prediction accuracy compared to 0.95. No gains in GEBV accuracy were observed for BWT and YGFW by using genomic information. The increase in the accuracy

using the SNP- and haplotype-based models were similar, with differences smaller than 1% for all traits.

3.2 Bias

The prediction bias ranged between -0.104 (H-BLUP for PWT) and 0.087 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.15 and α of 0.95 for YGFW (Supplementary File 2). Different from what was observed for the GEBV accuracies, the predictions for NLB were in general less biased than the other traits while those for PWT were the most biased. The prediction bias for the haplotype-based methods was similar across LD thresholds (used to create the haploblocks) and, thus, only the HAP-BLUP-LD_0.50 were presented for comparison purposes (Figure 2).

Incorporating genomic information in the analyses resulted in similar or more bias when compared to the pedigree-based prediction. Alpha equal to 0.95 tended to reduce the bias for BWT and YWT, while the opposite was observed for the other traits (i.e., using $\alpha = 0.50$ reduced the prediction bias for the other traits).

3.3 Dispersion

The GEBVs dispersion ranged from -0.572 (A-BLUP for PWT) to 0.035 (A-BLUP for YGDW) (Supplementary File 2). The dispersion was closer to zero (expected value for this statistic under no dispersion) for YGFW while it was more distant from and typical below one for BWT and PWT indicating GEBV were overestimated. GEBV predictions using haplotypes from blocks with different LD thresholds resulted in similar dispersion of the GEBV. Therefore, the HAP-BLUP-LD_0.50 scenario was also used to represent the haplotype-based methods to compare with A-BLUP and H-BLUP (Figure 3).



Figure 2. Prediction bias for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different α values (0.95 or 0.50) were used to create the genomic relationship matrices.

A dispersion of -0.29, -0.16, and -0.35 was observed for PWT, YWT, and NLB, respectively, using H-BLUP and HAP-BLUP-LD_0.50 with α of 0.95. Those values were closer to zero than when using A-BLUP (-0.57, -0.35, and -0.39, respectively), showing reduced dispersion for genomic based models. Pedigree-based models presented similar or lower dispersion for BWT and wool traits. The dispersion with H-BLUP and HAP-BLUP-LD_0.50 showed similar results regardless of the α values. Alpha equal to 0.95 tended to present better

dispersion for PWT, YWT, and NLB compared to 0.50, while the opposite was observed for the other traits.

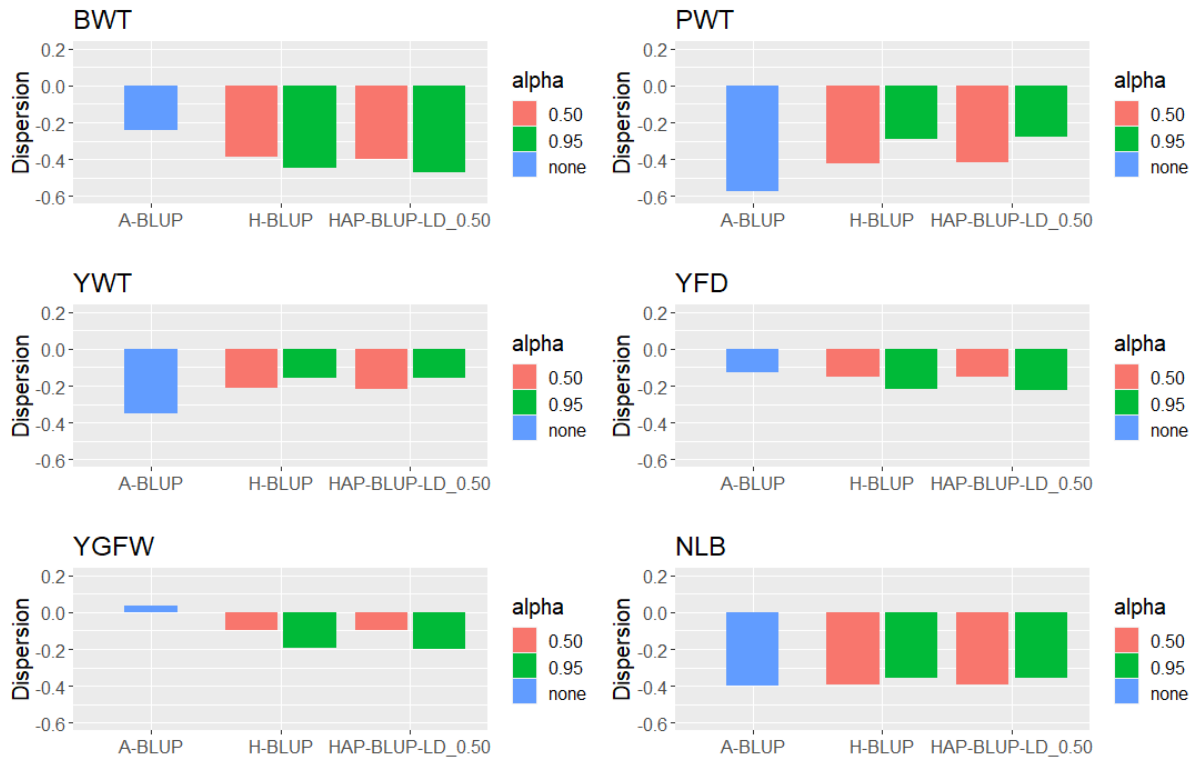


Figure 3. Dispersion of the GEBVs for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFYW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different α values (0.95 or 0.50) were used to create the genomic relationship matrices.

3.4 Theoretical accuracy

The mean TA ranged from 0.499 (A-BLUP for PWT) to 0.795 (HAP-BLUP using haplotypes from blocks with LD threshold of 0.35 and alpha equal to 0.95 for YFD) (Supplementary File 2). Considering all traits, the mean TA was 0.631 (0.085) and TA values

were higher for YFD and lower for PWT. Results from the haplotype-based methods had similar mean TA regardless of the LD threshold used to construct the haploblocks for all traits. The HAP-BLUP-LD_0.50 was, therefore, again used to represent the HAP-BLUP methods (Figure 4).

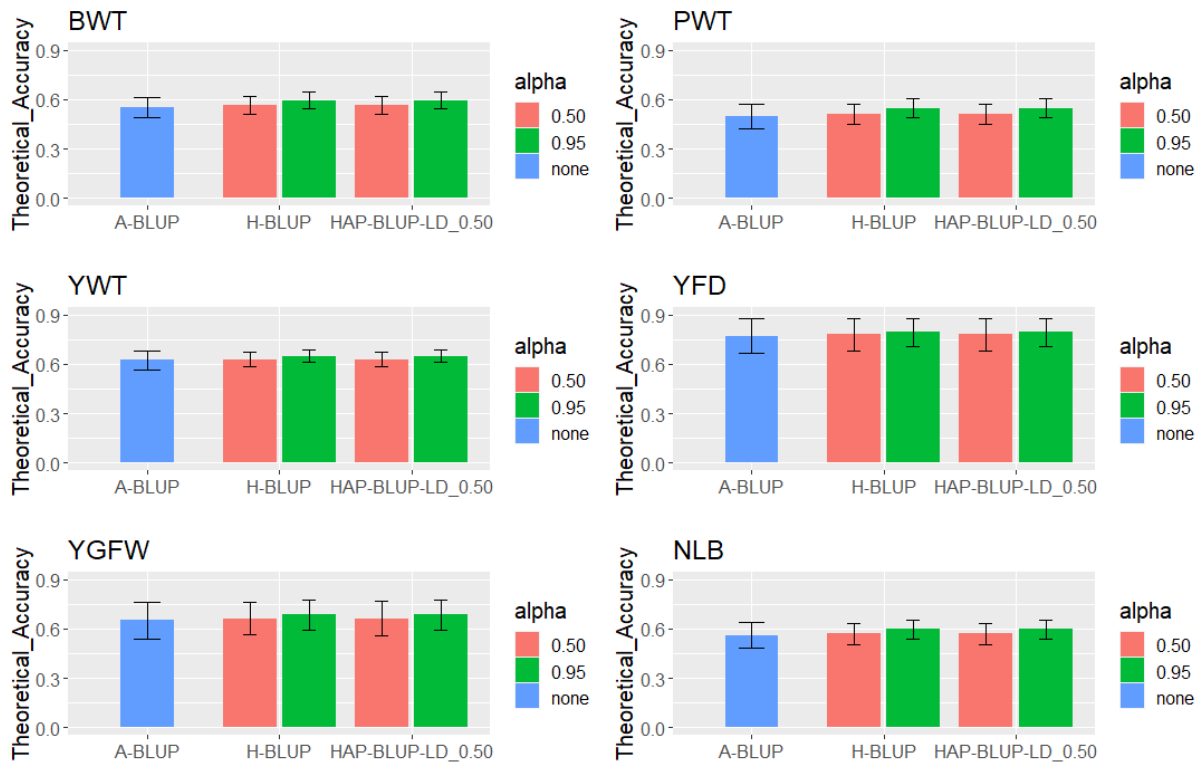


Figure 4. Mean theoretical accuracies for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR) in Rambouillet sheep using pedigree BLUP (A-BLUP), SNP-based single-step GBLUP (H-BLUP), and haplotype-based single-step GBLUP using haplotypes from blocks with LD threshold of 0.50 (HAP-BLUP-LD_0.50). Different α values (0.95 or 0.50) were used to create the genomic relationship matrices.

The genomic information tended to improve the mean TA for all traits, with increases up to ~7% (~0.04), ~9% (~0.05), ~4% (~0.03), ~3% (~0.02), ~5% (~0.03), and ~6% (~0.04)

for BWT, PWT, YWT, YFD, YGFW, and NLB, respectively, using H-BLUP and HAP-BLUP-LD_0.50 with α of 0.95. Negligible difference (less than 1%) was observed in the increase of the mean TA between H-BLUP and HAP-BLUP-LD_0.50 with α of 0.95. Using α equal to 0.50 resulted in the smallest increase in the mean TA (less than 2%) with both SNP- and haplotype-based methods for all traits. At the individual level, the TA using H-BLUP with α of 0.95 were higher compared to A-BLUP for the younger individuals and those with no phenotypic information (sires and dams with genotyped progeny) in the partial datasets for all traits (Figure 5).

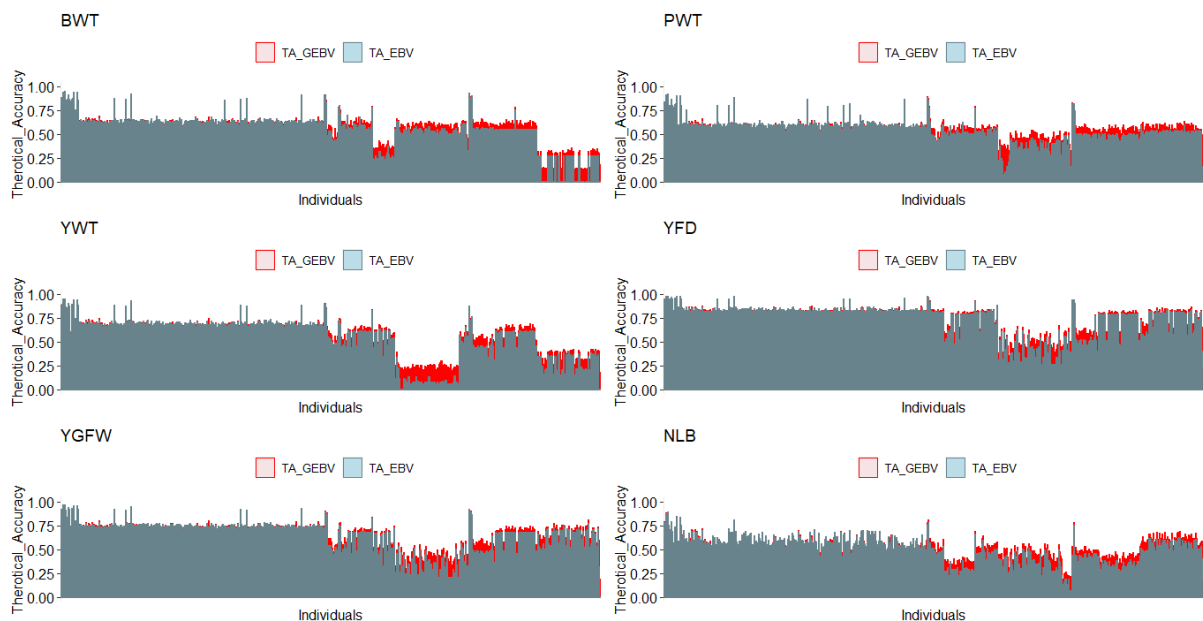


Figure 5. Theoretical accuracies for the genomic estimated breeding values using SNP (TA_GEBV) and estimated breeding values (TA_EBV) per genotyped individuals for birth weight (BWT), post-weaning weight (PWT), yearling weight (YWT), yearling fiber diameter (YFD), yearling greasy fleece weight (YGFW), and number of lambs born (NLR). The TA_GEBV and TA_EBV were obtained using SNP in the single-step GBLUP (H-BLUP) with alpha equal to 0.95 and pedigree-based BLUP (A-BLUP), respectively. The individuals were sorted by birth date, so that the youngest individuals are in the right side of each plot.

4. Discussion

Genomic selection is the state-of-the-art in modern sheep breeding programs. Here we present the first results of genomic predictions for body weight, wool, and reproductive traits in North American Rambouillet sheep. We performed single-step genomic predictions fitting SNP or haplotypes to create the genomic relationship matrices used to compute GEBV. Despite the small dataset, promising results were obtained, which can be used as a starting point for the implementation of genomic selection in Rambouillet sheep as well as in other sheep populations.

4.1 Genetic and genomic prediction results

The accuracy of genomic predictions relies mainly on the trait heritability, LD between SNP and QTL (Meuwissen et al., 2001), population structure, and genetic diversity of the population (Daetwyler et al., 2012). For Rambouillet sheep, the prediction accuracies for the pedigree- and genomic-based models followed the expected pattern in being higher for more heritable traits (Table 1; Figure 1). However, the smallest (~8%) differences between the pedigree- and genomic-based prediction accuracies were observed for YFD, which was the trait with the highest heritability (0.57), in comparison to NLB (~37%), which was the trait with lowest heritability (0.09).

Despite the expectation of the theoretical accuracies of genomic predictions to be higher for traits with higher heritability (Meuwissen et al., 2001), benefits of using GEBV are expected to be higher for traits with low heritability, sex limited, hard-to-measure, and recorded late in life, especially in sheep (Brito et al., 2017b; Brown et al., 2018). In this context, given that the assumptions for the MME (Henderson, 1984) are met (e.g., polygenic architecture, deep and accurate pedigree data, large number of phenotypic records, no preselection), the EBV are

expected to be BLUP and predict the unknown true breeding values well. In other words, the prediction accuracies using A-BLUP for highly heritable traits where individuals and/or their progeny have phenotypes are expected to be high.

As the genotyped cohort used in this study is the first official attempt to create a training population for genomic evaluations of Rambouillet sheep, the genotyped individuals included key ancestors and other selected animals with phenotypes or representative progeny with phenotypes for the traits evaluated. In this case, the EBV for those individuals can be well estimated, especially for YFD and YGFW (higher heritabilities; Table 2), which showed small and negligible increase in the prediction accuracy for the genomic- compared to pedigree-based models. For such highly heritable traits, using GEBV would still be more important to select breeding candidates at a younger age, i.e., measured only at yearling age. The substantial increase in the prediction accuracy for PWT, YWT, and NLB (higher than ~37%) shows that greater genetic gains can be achieved for these traits by including genomic information, as the accuracy is one of the main components of the selection response (Falconer & Mackay, 1996).

The genomic prediction accuracies observed in our study were within the range for most of the economic traits in sheep, which is between 0.20 and 0.50 according to Brown et al. (2018), especially when using α of 0.95 to construct the **G** matrix. Oliveira et al. (2021) observed prediction accuracies for BWT ranging from 0.06 to 0.13 using H-BLUP for Norwegian White and New Zealand Composite sheep populations. Unlike what was observed in the current study, Moghaddar et al. (2019) showed accuracies for genomic predictions ranging between 0.40 to 0.60 for PWT, 0.30 to 0.40 for yearling clean fleece weight, and 0.30 to 0.50 for YFD using BayesR and GBLUP for purebred Merino and crosses between Merino and Border Leicester. Genomic prediction accuracies of 0.24 and 0.28 were observed for YGFW using GBLUP and BayesR, respectively, and 0.31 and 0.35 for YFD for the same methods, respectively, in Merino and crossed Merino (Bolormaa et al., 2017a). For NLB,

Bolormaa et al. (2017b) reported genomic prediction accuracies ranging from 0.15 to 0.56 considering different validation strategies and prediction based on GBLUP and BayesR under cross-validation approaches for Merino sheep. Those differences in the genomic prediction accuracies for the same traits are also related to the statistical model and validation method used in the evaluations, beyond the other factors previously mentioned (e.g., heritability, population structure, and genetic diversity). It is also important to point out that GEBV accuracies were not calculated in the same way across all the studies, but the LR method used in this study is currently considered as the gold-standard approach.

The regression coefficient of the adjusted (or corrected) phenotypes or EBV on the GEBV is usually used to measure the “bias” of GEBV (Brown et al., 2018; Gao et al., 2012; Moghaddar et al., 2019; Oliveira et al., 2021). This measure was accessed as dispersion in our study, as it represents how the GEBV were deflated (over or under-estimated). Prediction bias, as a property of the method and the population under evaluation, is the expectation of the difference between average true and predicted breeding value; bias is zero under ideal conditions and can be approximated by the difference between the average (G)EBV in the whole and partial data sets (Legarra & Reverter, 2018). The fact that GEBV and EBV for most of the scenarios across traits were over-estimated in the Rambouillet sheep was similar to that observed in other studies (Brown et al., 2018; Moghaddar et al., 2019; Oliveira et al., 2021). Our conclusions regarding the benefit of including genomic information in prediction based on dispersion followed the same pattern observed for the GEBV prediction accuracies; this was likely because they are affected by similar factors.

Reports of prediction bias in sheep, as described by Legarra and Reverter (2018), are scarce, and were found only in dairy sheep (Macedo et al., 2020; 2022). The same is true for TA, although this is an important metric when reporting breeding values back to producers. Brito et al. (2017b) observed TA values ranging from 0.25 to 0.49 across a range of growth,

carcass, and meat quality traits, which are smaller than the values observed in the current study. The substantial increase in the TA (Figure 4) especially for the young individuals using genomic information (Figure 5) is promising because these are the focal individuals that need to be ranked for selection purposes.

Selective genotyping can result in maximum genetic response (Bologn et al., 2012), which could explain the improvements in the prediction results for most of the traits analyzed using genomic information even based on a small number of individuals (242 to 632 for NLB and BWT, respectively) genotyped using a moderate SNP density panel (~32 K SNP). Most of the prediction accuracy using genomic information is due to population structure, as described by Daetwyler et al. (2012). Those authors showed that up to 86% of the prediction accuracy can be achieved by using only one chromosome in a multibreed sheep population. Although one chromosome was enough to capture the population structure, it was unlikely to contain all the QTL affecting a trait (Daetwyler et al., 2012). Nevertheless, the recommendation is to increase the SNP panel density, through genotyping and imputation, for genomic predictions so that both population structure and LD between marker and QTL are fully explored (Daetwyler et al., 2012). Exploring and using weighted single-step genomic predictions and genome-wide associations has also been encouraged (Wang et al., 2012) as there may be important genomic regions that explain more of the total additive genetic variance for the traits of interest.

Selective genotyping can increase bias in the variance component estimation and therefore is not recommended for the breeding programs that only use phenotypes and pedigree relationships to drive selection decisions (Wang et al., 2020). The potential bias from selective genotyping was a reason for using the variance components provided by the NSIP, derived using solely phenotypic and pedigree information. Using random selection to choose the samples to be genotyped, as well as increasing the training population size, are recommended

to avoid bias in both variance component estimation and GEBV, as more biased predictions were observed using genomic information for most of the traits (Figure 2).

4.2 Using different alpha values to construct the genomic relationship matrices

In general, appropriate α and β parameters have more impact in GEBV bias reduction (Gao et al., 2012). Despite greater GEBV accuracies for a higher (0.95) α value, 0.50 is the choice to create \mathbf{G} in single-step evaluations for a range of carcass traits in terminal sire sheep breeds in Australia (McMillan & Swan, 2017). According to these authors, an α equal to 0.50 was chosen because: 1) when increasing α , accuracies increased until reaching an asymptote at around 0.50, which was not the case in the current study; 2) the GEBV using α between 0.50 and 0.95 were highly correlated; and 3) less variation was observed in the GEBV of genotyped individuals without phenotypes with α equal to 0.50; 4) with higher α values, GEBV bias (over-prediction) increased.

In this study, using α of 0.50 showed only half of the increase in the accuracies compared to 0.95. No clear advantage in GEBV bias, dispersion, or TA was observed with one α value compared to the other. Therefore, we recommend α of 0.95 for single-step genomic evaluations in U.S. Rambouillet sheep. However, it is important to highlight that we had a smaller number of genotyped individuals as compared to McMillan and Swan (2017) and we used the LR method (Legarra & Reverter, 2018) to derive the GEBV accuracies, bias, and dispersion; they instead used cross-validation to test their predictions with random assignment of individuals to groups.

4.3 Haplotype-based single-step genomic predictions

The HAP-BLUP-LD_0.50 scenario was chosen to represent the haplotype-based methods because the LD of 0.50 was the level most likely to estimate the recombination

hotspots properly, which are the specific points in the genome with higher probability of recombination (Kim et al., 2018). The LD threshold of 0.50 to create the haploblocks also tended to provide better results between the haplotype-based prediction for many of the traits, but in other scenarios, similar results were observed.

Using haplotype-based methods did not improve the prediction results for any of the traits analyzed (accuracies, bias, dispersion, and TA) compared to fitting SNP in a real sheep dataset. Improved predictions had been hypothesized by Araujo et al. (2021). According to this previous simulation-based study, haplotype-based genomic predictions could outperform SNP-based models in real sheep datasets because the former can capture epistasis and these populations could have more complex interactions within haplotype blocks due higher effective population size. Liang et al. (2020) showed that epistasis was the main reason for higher GEBV accuracies when using haplotypes instead of SNP in seven traits in humans, which is a highly genetically diverse population (Park, 2011). However, in this study, the small number of genotyped individuals (722) as well as the density of the SNP panel used (~32K SNP) could have affected both SNP and haplotype predictions.

The algorithm to create the LD-based haploblocks and the method to code the haplotypes during the creation of the relationship matrix could also have affected the prediction results. There are several algorithms to create LD-haploblocks, such as MATILDE (Pattaro et al., 2008), confidence interval (Gabriel et al., 2002), four gamete test (Wang et al., 2002), solid spine (Barret et al., 2005), MIG++ (Taliun et al., 2014), S-MIG++ (Taliun et al., 2016), and Big-LD (Kim et al., 2018). We have used the Big-LD algorithm to construct the haploblocks because the LD-blocks produced by this method agree better with the true recombination hotspots (determined experimentally in the major histocompatibility complex region from semen of north-European British donors) and is more computationally efficient than the previously mentioned algorithms (Kim et al., 2018). However, as the haplotype diversity index

and true discovery ratio of the recombination hotspots can be lower using Big-LD (Kim et al., 2018), evaluating different algorithms to create the LD-haplotype blocks is also recommended. New methods based on clustering algorithms (Won et al., 2020) and machine learning methods (Lim et al., 2022) have also recently been proposed to create and select the best haplotypes to be used, respectively. This type of study is scarce not only in sheep but also in other livestock species.

The haplotypes can be multiallelic markers (Gabriel et al., 2002; Calus et al., 2008). However, we used the unique multiallelic haplotype alleles coded as ps-SNP to perform genomic predictions under the ssGBLUP framework. The ps-SNP derived from the LD-haploblocks were then merged with the NCSNP to create the \mathbf{G} matrix, similar to Araujo et al. (2021). This strategy enables using haplotypes to perform genomic predictions using software developed for fitting individual SNP (Teissier et al., 2020), including or excluding non-genotyped individuals (ssGBLUP and GBLUP, respectively). Teissier et al. (2020) considered both NCSNP and unique multiallelic haplotype alleles as ps-SNP and observed up to 22% increase in GEBV prediction accuracy using different LD- or fixed-SNP-length-based haploblocks for milk production traits in dairy goats using ssGBLUP. Milk production traits are known to be affected by a major gene (*DGATI*). In general, GEBV prediction results for haplotype-based methods are scarce in small ruminants and additional studies are needed.

The GVCHAP is a computing pipeline that allows multiallelic haplotypes to be used directly to create a genomic additive (and dominance) relationship matrix for both genomic prediction and variance component estimation using haplotypes or SNP (Prakapenca et al., 2020). GVCHAP is based in the multiallelic haplotype model proposed by Da (2015), which uses the quantitative genetic theory to derive a general multiallelic partition of genotypic values with factorization to define the genomic relationships. However, the GVCHAP is based on GREML and GBLUP and, thus, only considers genotyped individuals with phenotypes.

Considering the different algorithms and methods to create the genomic relationship matrix including haplotypes, there are still alternatives to evaluate the feasibility of including haplotypes in genomic predictions. Future studies in sheep should also consider the possibility of creating haplotypes based in functional information (e.g., gene regions) to make haplotype predictions (Da et al., 2015; Prakapenca et al., 2020).

Despite the hypothesis that haplotypes could outperform SNP and provide high accuracies and lower bias in genomic predictions, practical results show that this does not usually happen. As summarized by Araujo et al. (2021), the benefits of using haplotype-based methods for genomic prediction are equivocal. Improvements occur mainly in the evaluation of traits with major genes, as shown by Teissier et al. (2020). Nevertheless, as stated before, there are haplotype blocking and selection methods that should be further investigated.

Marker density can also affect the accuracy of SNP phasing (Weng et al., 2014) and the precision in which the recombination hotspots are determined (Weng et al., 2019) which, respectively, are the first steps for the haplotype prediction and the basis of the LD-based haploblocks. In addition, epistasis, which is the component that might contribute the most to improvements in accuracy with haplotype predictions (Liang et al., 2020), is a complex effect and requires a substantial number of individuals and markers/bins to be properly estimated (Zhang et al., 2016). Therefore, larger reference populations and denser SNP panels are recommended to evaluate genomic predictions using haplotypes in sheep populations.

5. Conclusions

Fitting SNP or haplotypes (as pseudo-SNP) provided similar or higher GEBV prediction and theoretical accuracies and reduced the dispersion of the GEBV for body weight, wool, and reproductive traits in Rambouillet sheep, while the prediction bias showed no clear improvements by adding genomic information. Alpha value equal to 0.95 is recommended to

weight the genomic relationships to model the covariances between individuals. The use of haplotypes showed no advantage compared to SNP at the current reference population size and SNP panel density used, regardless of the LD threshold used to create the haploblocks. Efforts to increase the number of genotyped individuals are paramount to take full advantage of genomic information to accelerate genetic progress in the U.S. Rambouillet sheep breeding program.

6. Conflict of Interest

The authors declare no conflict of interest.

7. Data availability

The phenotypic, pedigree and genomic data used in this study are the property of the industry partner that contributed to the study and therefore are not readily available due to its commercial sensitivity. Requests to access the datasets should be directed to the National Sheep Improvement Program (NSIP). The computing pipelines used in this research are available by request to the corresponding authors.

8. Author Contribution Statement

ACA., PLSC, RML, and LFB: conception of the work. RML and LFB: data acquisition. ACA: SNP and haplotype-based single-step genomic prediction analyses. ACA, RML, and LFB: results interpretation. ACA: drafting the manuscript. ACA, PLSC, HRO, RML and LFB: critical revision of the manuscript. ACA, PLSC, HRO, RML, and LFB: final approval of the version to be published. All authors contributed to the article and approved the submitted version. All authors have read and agreed to the published version of the manuscript.

9. Funding

This research was funded by the National Sheep Industry Improvement Center (NSIIC-USDA), the American Rambouillet Sheep Breeders Association, and the American Sheep Industry Association Let's Grow program. The State University of Southwest Bahia and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil), Financial Code 001, provided the scholarship of the first author.

10. Acknowledgements

We acknowledge the NSIP for providing the datasets used in this research, and the nine NSIP Rambouillet flocks that provided the DNA samples genotyped. We also thank Dr. Jessica Petersen (University of Nebraska-Lincoln) for updating the marker positions to the most recent ovine reference genome (ARS-UI Ramb v2.0) and merging the SNP panels used, Dr. Mathew Spangler (University of Nebraska-Lincoln) for providing useful comments on the final version of the manuscript, and other members of Brito's lab for providing scientific support to develop this research. Lastly, we thank Purdue University and the State University of Southwest Bahia for providing academic and financial support to the authors. We also acknowledge the National Development Council Scientific Technological (CNPq, Brazil) for the fellowship.

11. References

- Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot Topic: A Unified Approach to Utilize Phenotypic, Full Pedigree, and Genomic Information for Genetic Evaluation of Holstein Final. *Journal of Dairy Science*, 93(2):743–752. doi:10.3168/jds.2009-2730
- Araujo, A. C., Carneiro, P. L. S., Oliveira, H. R., Schenkel, F. S., Veroneze, R., Lourenco, D. A. L., & Brito, L. F. (2021). A comprehensive comparison of haplotype-based single-

- step genomic predictions in livestock populations with different genetic diversity levels: A simulation study. *Frontiers in Genetics*, 12(729867). doi: 10.3389/fgene.2021.729867
- Araujo, A. C., Carneiro, P. L. S., Alvarenga, A. B., Oliveira, H. R., Miller, S. P., Retallick, K., & Brito, L. F. (2022). Haplotype-Based Single-Step GWAS for Yearling Temperament in American Angus Cattle. *Genes*, 13(17). doi: <https://doi.org/10.3390/genes13010017>
- Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, 21(2):263–265. doi: <https://doi.org/10.1093/bioinformatics/bth457>
- Bohmanova, J., Sargolzaei, M., & Schenkel, F. S. (2010). Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics*, 11(421). doi: <https://doi.org/10.1186/1471-2164-11-421>
- Boligon A. A., Long, N., Albuquerque, L. G., Weigel, K. A., Gianola, D., & Rosa, G. J. M. (2012). Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *Journal of Animal Science*, 90(13):4716–4722. doi: 10.2527/jas.2012-4857
- Bolormaa, S., Brown, D. J., Swan, A. A., van der Werf, J. H. J., Hayes B. J., & Daetwyler, H. D. (2017a). Genomic prediction of reproduction traits for Merino sheep. *Animal Genetics*, 48(3):338-348. doi: 10.1111/age.12541
- Bolormaa, S., Swan, A. A., Brown, D. J., Hatcher, S., Moghaddar, N., van der Werf, J. H. J., Goddard, M. E., & Daetwyler, H. D. (2017b). Multiple-trait QTL mapping and genomic prediction for wool traits in sheep. *Genetics Selection Evolution*, 49(62). doi: 10.1186/s12711-017-0337-y
- Bovo, S., Ballan, M., Schiavo, G., Ribani, A., Tinarelli, S., Utzeri, V. J., Dall'Olio, S., Gallo, M., & Fontanesi, L. (2021). Single-marker and haplotype-based genome-wide

- association studies for the number of teats in two heavy pig breeds. *Animal Genetics*, 52(4):440–450. doi: <https://doi.org/10.1111/age.13095>
- Brito, L. F., Mcewan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., Schenkel, F. S., & Clarke, S. M. (2017a). Genetic Diversity of a New Zealand Multi-Breed Sheep Population and Composite Breed's History Revealed by a High-Density SNP Chip. *BMC Genetics*, 18(25). doi:10.1186/s12863-017-0492-8
- Brito, L. F., Clarke, S. M., Mcewan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., Sargolzaei, M., & Schenkel, F. S. (2017b). Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. *BMC Genetics*, 18(7). doi: <https://doi.org/10.1186/s12863-017-0476-8>
- Brito L. F., Oliveira, H. R., McConn, B. R., Schinckel, A. P., Arrazola, A., Marchant-Forde, J. N., & Johnson J. S. (2020). Large-Scale Phenotyping of Livestock Welfare in Commercial Production Systems: A New Frontier in Animal Breeding. *Frontiers in Genetics*, 11(793). doi: 10.3389/fgene.2020.00793
- Brown, D. J., Swan, A. A., Boerner, V., Li, L., Gurman, P. M., McMillan, A.J., van der Werf, J. H. J., Chandler, H. R., Tier, B., & Banks, R.G. (2018). Single-Step Genetic Evaluations in the Australian Sheep Industry. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, 11. 460.
- Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., & Veerkamp, R. F. (2008). Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics*, 178(1):553–561. doi:10.1534/genetics.107.080838
- Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, 42(1). doi: <https://doi.org/10.1186/1297-9686-42-2>

- Da Y. (2015). Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genetics*, 16(144). doi: <https://doi.org/10.1186/s12863-015-0301-1>
- Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., & Hayes, B. J. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal Animal Science*, 90(10):3375–3384. doi: 10.2527/jas2011-4557
- Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Essex, UK: Longman, 4.
- Feitosa, F. L. B., Pereira, A. S. C., Amorim, S. T., Peripolli, E., Silva, R. M. D. O., Braz, C. U., Ferrinho, A. M., Schenkel, F. S., Brito, L. F., Espigolan, R., Albuquerque, L. G., & Baldi, F. (2020). Comparison between haplotype-based and individual SNP-based genomic predictions for beef fatty acid profile in Nelore cattle. *Journal of Animal Breeding and Genetics*, 137(5):468-476. doi: <https://doi.org/10.1111/jbg.12463>
- Feitosa, F. L. B., Pereira, A. S. C., Mueller, L. F., de Souza Fonseca, P. A., Braz, C. U., Amarin, S., Espigolan, R., Lemos, M. A., de Albuquerque, L. G., Schenkel, F. S. Brito, L. F., Stafuzza, N. B., & Baldi, F. (2021). Genome-wide association study for beef fatty acid profile using haplotypes in Nelore cattle. *Livestock Science*, 245(104396). doi: <https://doi.org/10.1016/j.livsci.2021.104396>
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordeiro, S. N., Rotini, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. L., Daly, M. J., & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229. doi: 10.1126/science.1069424
- Gao, H., Christensen, O. F., Madsen, P., Nielsen, U. S., Zhang, Y., Lund, M. S., & Su, G. (2012). Comparison on genomic predictions using three GBLUP methods and two

- single-step blending methods in the Nordic Holstein population. *Genetics Selection Evolution*, 44(8). doi: <http://www.gsejournal.org/content/44/1/8>
- Guarini, A. R., Lourenco, D. A. L., Brito, L. F., Sargolzaei, M., Baes, C. F., Miglior, F., Misztal, I., & Schenkel, F. S. (2018). Comparison of Genomic Predictions for Lowly Heritable Traits Using Multi-step and Single-step Genomic Best Linear Unbiased Predictor in Holstein Cattle. *Journal of Dairy Science*, 101(9):8076–8086. doi:10.3168/jds.2017-14193
- Henderson, C. R. (1950) Estimation of Genetic Parameters. *Annals of Mathematical Statistics*, 21:309-310.
- Henderson, C. R. (1984). *Application of linear models in animal breeding*. Guelph: University of Guelph.
- Hess, M., Druet, T., Hess, A., & Garrick, D. (2017). Fixed-length Haplotypes Can Improve Genomic Prediction Accuracy in an Admixed Dairy Cattle Population. *Genetics Selection Evolution*, 49(54). doi: 10.1186/s12711-017-0329-y
- Hill, W. G., & Robertson, A. Linkage disequilibrium in finite populations. (1968). *Theoretical and Applied Genetics*, 38(226). doi: <https://doi.org/10.1007/BF01245622>
- Jiang, Y., Schmidt, R. H., & Reif, J. C. (2018). Haplotype-based Genome-wide Prediction Models Exploit Local Epistatic Interactions Among Markers. *Genes Genomes Genetics*, 8(5):1687–1699. doi:10.1534/g3.117.300548
- Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto-Neto, L. R., Cristobal, M. S., Servin, B., McCulloch, R., Whan, V., Gietzen, K., Paiva, S., Barendse, W., Ciani, E, Raadsma, H., McEwan, J., Dalrymple, & other members of the International Sheep Genomics Consortium. (2012). Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *Plos Biology*, 10:e1001258. doi:10.1371/journal.pbio.1001258

- Kim, S. A., Brossard, M., Roshandel, D., Paterson, A. D., Bull, S. B., & Yoo, Y. J. (2019). gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics*, 35(4419). doi: <https://doi.org/10.1093/bioinformatics/btz308>
- Kim, S. A., Cho, C. S., Kim, S. R., Bull, S. B., & Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, 34(388). doi: [10.1093/bioinformatics/btx609](https://doi.org/10.1093/bioinformatics/btx609)
- Legarra, A., Aguilar, I., & Misztal, I. (2009). A Relationship Matrix Including Full Pedigree and Genomic Information. *Journal of Dairy Science*, 92(9):4656–4663. doi:10.3168/jds.2009-2061
- Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livestock Science*, 166:54-65. Doi: <https://doi.org/10.1016/j.livsci.2014.04.029>
- Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, 50(53). doi: [10.1186/s12711-018-0426-6](https://doi.org/10.1186/s12711-018-0426-6)
- Liang, Z., Tan, C., Prakapenka, D., Ma, L., & Da, Y. (2020). Haplotype Analysis of Genomic Prediction Using Structural and Functional Genomic Information for Seven Human Phenotypes. *Frontiers in Genetics*, 11(1). doi: [10.3389/fgene.2020.588907](https://doi.org/10.3389/fgene.2020.588907)
- Lim, A. J. W., Lim, L. J., Ooi, B. N. S., Koh, E. T., Tan, J. W. L., Chong, S. S., Khor, C. C., Tucker-Kellogg, L., Leong, K. P., Lee, C. G. (2022). Functional coding haplotypes and machine-learning feature elimination identifies predictors of Methotrexate Response in Rheumatoid Arthritis patients. *eBioMedicine*, 75(103800). doi: <https://doi.org/10.1016/j.ebiom.2021.103800>.

- Macedo, F. L., Christensen, O. L., Astruc, J.-M., Aguilar, I., Masuda, Y., & Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genetics Selection Evolution*, 52(47). doi: <https://doi.org/10.1186/s12711-020-00567-1>
- Macedo, F L., Astruc, J. M., Meuwissen, T. H. E., & Legarra, A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. *Journal of Dairy Science*, 105(3):2439-2452. doi: <https://doi.org/10.3168/jds.2021-20860>
- Mazinani, M., & Rude, B. (2020). Population, World Production and Quality of Sheep and Goat Products. *American Journal of Animal and Veterinary Sciences*, 15(4):291-299. doi: [10.3844/ajavsp.2020.291.299](https://doi.org/10.3844/ajavsp.2020.291.299)
- Meyer, K., Tier, B., & Swan, A. (2018). Estimates of genetic trend for single-step genomic evaluations. *Genetics Selection Evolution*, 50(39). doi: <https://doi.org/10.1186/s12711-018-0410-1>
- McMillan A. J. & A. A. Swan, 2017. Weighting of genomic and pedigree relationships in single step evaluation of carcass traits in Australian sheep. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*, 22(1).
- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-wide Dense Marker Maps. *Genetics*, 157(4):1819–1829. doi: [10.1093/genetics/157.4.1819](https://doi.org/10.1093/genetics/157.4.1819)
- Misztal, I., Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Aguilar, I., Legarra, A., & Vitezica, Z. (2018). *Manual for BLUPF90 family programs*. University of Georgia. <http://nce.ads.uga.edu/wiki/doku.php?id=documentation>
- Moghaddar, N., Khansefid, M., van der Werf, J. H. J., Bolormaa, S., Duijvesteijn, N., Clark, S. A., Swan, A. A., Daetwyler, H. D., & MacLeod, I. M. Genomic prediction based

- on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genetics Selection Evolution*, 51(72). doi: <https://doi.org/10.1186/s12711-019-0514-2>
- Notter, D. R. 1998. The U.S. National Sheep Improvement Program: across-flock genetic evaluations and new trait development. *Journal of Animal Science*, 76(9):2324–2330. <https://doi.org/10.2527/1998.7692324x>.
- Oliveira, H. R., McEwan, J. C., Jakobsen, J. H., Blichfeldt, T., Meuwissen, T. H. E., Pickering, N. K., Clarke, S. M., & Brito, L. F. (2021). Across-country genomic predictions in Norwegian and New Zealand Composite sheep populations with similar development history. *Journal of Animal Breeding and Genetics*, 139(1):1-12. doi: <https://doi.org/10.1111/jbg.12642>
- Park, L. (2011). Effective Population Size of Current Human Population. *Genetics Research*, 93(2):105–114. doi:10.1017/S0016672310000558
- Pattaro, C., Ruczinski, I., Fallin, D. M., & Parmigiani, G. (2008). Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. *BMC Genomics*, (9)405. doi: <https://doi.org/10.1186/1471-2164-9-405>
- Prakapenka, D., Wang, C., Liang, Z., Bian, C., Tan, C., & Da, Y. (2020). GVCHAP: A Computing Pipeline for Genomic Prediction and Variance Component Estimation Using Haplotypes and SNP Markers. *Frontiers in Genetics*, 11(282). doi: <https://doi.org/10.3389/fgene.2020.00282>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81:559-575. doi: 10.1086/519795

- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, 15(478). doi: 10.1186/1471-2164-15-478.
- Taliun, D., Gamper, J., & Pattaro, C. (2014). Efficient haplotype block recognition of very long and dense genetic sequences. *BMC Bioinformatics*, 15(10). doi: <https://doi.org/10.1186/1471-2105-15-10>
- Taliun, D., Gamper, J., Leser, U., & Pattaro, C. (2016). Fast Sampling-Based Whole-Genome Haplotype Block Recognition. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(2):315–325. doi: <https://doi.org/10.1109/TCBB.2015.2456897>
- Teissier, M., Larroque, H., Brito, L. F., Rupp, R., Schenkel, F. S., & Robert-Granié, C. (2020). Genomic predictions based on haplotypes fitted as pseudoSNPs for milk production and udder type traits and somatic cell score in French dairy goats. *Journal of Dairy Science*, 103(12):11559-11573. doi: 10.3168/jds.2020-18662
- Thorne, J. W., Murdoch, B. M., Freking, B. A., Redden, R. R., Murphy, T. W., Taylor, J. B., & Blackburn, H. D. (2021). Evolution of the sheep industry and genetic research in the United States: opportunities for convergence in the twenty-first century. *Animal Genetics*, 52(4):395–408. doi: 10.1111/age.13067
- Vanraden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91(11):4414–4423. doi: 10.3168/jds.2007-0980
- Van Vleck, L. D. (1993). Variance of prediction error with mixed model equations when relationships are ignored. *Theoretical and Applied Genetics*, 85(5):545-549. doi: <https://doi.org/10.1007/BF00220912>.

- Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research*, 94(2):73–83. doi: 10.1017/S0016672312000274
- Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., & Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American journal of human genetics*, 71(5): 1227–1234. doi: <https://doi.org/10.1086/344398>
- Wang, L., Janss, L. L., Madsen, P., Henshall, J., Huang, C.-H., Marois, D., Alemu, S., Sørensen, A. C., & Jensen, J. (2020). Effect of genomic selection and genotyping strategy on estimation of variance components in animal models using different relationship matrices. *Genetics Selection Evolution*, 52(31). doi: <https://doi.org/10.1186/s12711-020-00550-w>
- Weng, Z., Wolc, A., Su, H., Fernando, R. L., Dekkers, J. C. M., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., & Garrick, D. J. (2019). Identification of recombination hotspots and quantitative trait loci for recombination rate in layer chickens. *Journal of Animal Science and Biotechnology*, 10(20). doi: <https://doi.org/10.1186/s40104-019-0332-y>
- Weng, Z.-Q., Saatchi, M., Schnabel, R. D., Taylor, J. F., & Garrick, D. J. (2014). Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genetics Selection Evolution*, 46(34). doi: <https://doi.org/10.1186/1297-9686-46-34>
- Won, S., Park, J.-E., Son, J.-H., Lee, S.-H., Park, B. H., Park, M., Park, W.-C., Chai, H.-H., Kim, H., Lee, J., & Lim, D. (2020). Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium. *Frontiers in Genetics*, 11(134). doi: 10.3389/fgene.2020.00134

Zhang, W., Dai, X., Wang, Q., Xu, S., & Zhao, P. X. (2016). PEPIS: A Pipeline for Estimating Epistatic Effects in Quantitative Trait Locus Mapping and Genome-Wide Association Studies. *PLoS Computational Biology*, 12(5):e1004925. doi: 10.1371/journal.pcbi.1004925

VI – CONSIDERAÇÕES FINAIS

O uso de haplótipos nas análises genômicas em ruminantes de interesse zootécnico é promissor. As acurácias, viés e dispersão dos valores genéticos genômicos dos indivíduos não apresentaram diferenças substanciais usando os haplótipos em comparação com os SNPs em estudo de simulação ou em dados reais de ovinos. No entanto, deve-se notar que a epistasia, um importante componente para a melhora dos resultados da predição com haplótipos, não foi considerado na simulação, sendo recomendado considerá-la em estudos futuros. No caso da predição genômica em dados reais de ovinos, deve-se observar que a população de referência utilizada foi relativamente pequena devido a limitações no número de animais genotipados, o que pode afetar negativamente a estimação dos efeitos dos haplótipos, sendo recomendado o aumento da população de referência em estudos futuros. No entanto, apesar das limitações, o uso de informação genômica (SNPs ou haplótipos) na construção do parentesco proporcionou melhores resultados na predição dos valores genéticos para a maioria das características de crescimento, lã e reprodutivas comparado ao pedigree; sendo recomendado a implementação de seleção genômica para a raça Rambouillet nos Estados Unidos. O incremento no tempo de análise com haplótipos quando comparado aos SNPs é evidente, no entanto, depende do número de indivíduos genotipados e nível de desequilíbrio de ligação. A presença de haplótipos nos estudos de associação genômica foram essenciais para a descoberta de genes candidatos novos, permitindo também identificar genes previamente descritos no comportamento de bovinos. Recomendamos o uso de haplótipos considerando diferentes níveis de desequilíbrio ligação junto com os SNPs para a prospecção de genes e *loci* de características quantitativas em animais de produção.

VII – ANEXOS

7.1 Normas da revista *Frontiers in Genetics*

Manuscript Formatting Guidelines

- **1. General standards**
 - 1.1. Article Type
 - 1.2. Templates
 - 1.3. Manuscript Length
 - 1.4. Language Editing
 - 1.5. Language Style
 - 1.6. Search Engine Optimization (SEO)
 - 1.7. CrossMark Policy
 - 1.8. Title
 - 1.9. Authors and Affiliations
 - 1.10. Consortium/Group and Collaborative Authors
 - 1.11. Abstract
 - 1.12. Keywords
 - 1.13. Text
 - 1.14. Nomenclature
 - 1.15. Sections
 - 1.16. Acknowledgments
 - 1.17. Contribution to the Field Statement
- **2. Figure and Table Guidelines**
 - 2.1. CC-BY Licence
 - 2.2. Figure Requirements and Style Guidelines
 - 2.2.1. Captions
 - 2.2.2. Image Size and Resolution Requirements
 - 2.2.3. Format and Color Image Mode
 - 2.2.4. Chemical Structures
 - 2.3. Table Requirements and Style Guidelines
 - 2.4. Accessibility
- **3. Supplementary Material**
- **4. References**
 - 4.1. Harvard Reference Style (Author-Date)
 - 4.1.1. In-text Citations
 - 4.1.2. Reference List

- **4.2. Vancouver Reference Style (Numbered)**

- 4.2.1. In-text Citations
- 4.2.2. Reference List

1. General standards

1.1. Article Type

Frontiers requires authors to carefully select the appropriate article type for their manuscript and to comply with the article type descriptions defined in the journal's "Article Types" page, which can be seen from the "For Authors" menu on any Frontiers journal page. Please pay close attention to the word count limits.

1.2. Templates

If working with Word please use our [Frontiers Word templates](#). If you wish to submit your article as LaTeX, we recommend our [Frontiers LaTeX templates](#).

For LaTeX files, please ensure all relevant manuscript files are uploaded: .tex file, PDF, and .bib file (if the bibliography is not already included in the .tex file).

During the [Interactive Review](#), authors are encouraged to upload versions using "Track Changes." Editors and reviewers can only download the PDF file of the submitted manuscript.

1.3. Manuscript Length

Frontiers encourages the authors to closely follow the article word count lengths given in the "Article Types" page of the journals. The manuscript length includes only the main body of the text, footnotes, and all citations within it, and excludes the abstract, section titles, figure and table captions, funding statement, acknowledgments, and references in the bibliography. Please indicate the number of words and the number of figures and tables included in your manuscript on the first page.

1.4. Language Editing

Frontiers requires manuscripts submitted to meet international English language standards to be considered for publication.

For authors who would like their manuscript to receive language editing or proofreading to improve the clarity of the manuscript and help highlight their research, Frontiers recommends the language-editing services provided by the following external partners:

Editage

Frontiers is pleased to recommend the language-editing service provided by our external partner Editage to authors who believe their manuscripts would benefit from professional editing. These services may be particularly useful for researchers for whom English is not the primary language. They can help to improve the grammar, syntax, and flow of your manuscript prior to submission. Frontiers authors will receive a 10% discount by visiting the following link: <https://editage.com/frontiers/>.

The Charlesworth Group

Frontiers recommends the Charlesworth Group's author services, who has a long-standing track record in language editing and proofreading. This is a third-party service for which Frontiers authors will receive a 10% discount by visiting the following link: <https://www.cwauthors.com/frontiers/>.

Frontiers **推荐您使用在英语语言编辑和校对领域具有悠久历史和良好口碑的查尔斯沃思作者服务**。此项服务由第三方为您提供。Frontiers **中国作者**通过此链接提交稿件时可获得**10%的特别优惠**: www.cwauthors.com.cn/frontiers/.

Note that sending your manuscript for language editing does not imply or guarantee that it will be accepted for publication by a Frontiers journal. Editorial decisions on the scientific content of

a manuscript are independent of whether it has received language editing or proofreading by the partner services, or other services.

1.5. Language Style

The default language style at Frontiers is American English. If you prefer your article to be formatted in British English, please specify this on the first page of your manuscript. For any questions regarding style, Frontiers recommends authors to consult the Chicago Manual of Style.

1.6. Search Engine Optimization (SEO)

There are a few simple ways to maximize your article's discoverability. Follow the steps below to improve search results of your article:

- include a few of your article's keywords in the title of the article;
- do not use long article titles;
- pick 5 to 8 keywords using a mix of generic and more specific terms on the article subject(s);
- use the maximum amount of keywords in the first 2 sentences of the abstract;
- use some of the keywords in level 1 headings.

1.7. CrossMark Policy

CrossMark is a multi-publisher initiative to provide a standard way for readers to locate the current version of a piece of content. By applying the CrossMark logo Frontiers is committed to maintaining the content it publishes and to alerting readers to changes if and when they occur. Clicking on the CrossMark logo will tell you the current status of a document and may also give you additional publication record information about the document.

1.8. Title

The title should be concise, omitting terms that are implicit and, where possible, be a statement of the main result or conclusion presented in the manuscript. Abbreviations should be avoided within the title.

Witty or creative titles are welcome, but only if relevant and within measure. Consider if a title meant to be thought-provoking might be misinterpreted as offensive or alarming. In extreme cases, the editorial office may veto a title and propose an alternative.

Authors should try to avoid, if possible:

- titles that are a mere question without giving the answer;
- unambitious titles, for example starting with "Towards," "A description of," "A characterization of," "Preliminary study on;"
- vague titles, for example starting with "Role of...," "Link between...," "Effect of..." that do not specify the role, link, or effect;
- include terms that are out of place, for example the taxonomic affiliation apart from species name.

For Corrigenda, General Commentaries, and Editorials, the title of your manuscript should have the following format:

- "Corrigendum: Title of Original Article"
- General Commentaries
 - "Commentary: Title of Original Article"
 - "Response: Commentary: Title of Original Article"
- "Editorial: Title of Research Topic"

The running title should be a maximum of 5 words in length.

1.9. Authors and Affiliations

All names are listed together and separated by commas. Provide exact and correct author names as these will be indexed in official archives. Affiliations should be keyed to the author's name with superscript numbers and be listed as follows: Laboratory, Institute, Department, Organization, City, State abbreviation (only for United States, Canada, and Australia), and Country (without detailed address information such as city zip codes or street names).

Example: Max Maximus¹

¹ Department of Excellence, International University of Science, New York, NY, United States.

Correspondence:

The Corresponding Author(s) should be marked with an asterisk in the author list. Provide the exact contact email address of the corresponding author(s) in a separate section.

Example: Max Maximus

maximus@iuscience.edu

If any authors wish to include a change of address, list the present address(es) below the correspondence details using a unique superscript symbol keyed to the author(s) in the author list.

Equal contributions:

The authors who have contributed equally should be marked with a symbol (†) in the author list of the doc/latex and pdf files of the manuscript uploaded at submission.

Standard statements to include in the author list:

Equal contribution	These authors have contributed equally to this work
First authorship	These authors share first authorship
Senior authorship	These authors share senior authorship
Last authorship	These authors share last authorship
Equal contribution & First authorship	These authors have contributed equally to this work and share first authorship
Equal contribution & Senior authorship	These authors have contributed equally to this work and share senior authorship
Equal contribution & Last authorship	These authors have contributed equally to this work and share last authorship

Example: Max Maximus 1[†], John Smith2[†] and Barbara Smith1

[†]These authors have contributed equally to this work and share first authorship

1.10. Consortium/Group and Collaborative Authors

Consortium/group authorship should be listed in the manuscript with the other author(s).

In cases where authorship is retained by the consortium/group, the consortium/group should be listed as an author separated by “,” or “and,”. The consortium/group name will appear in the author list, in the citation, and in the copyright. If provided, the consortium/group members will be listed in a separate section at the end of the article.

For the collaborators of the consortium/group to be indexed in PubMed, they do not have to be inserted in the Frontiers submission system individually. However, in the manuscript itself, provide a section with the name of the consortium/group as the heading followed by the list of collaborators, so they can be tagged accordingly and indexed properly.

Example: John Smith, Barbara Smith and The Collaborative Working Group.

In cases where work is presented by the author(s) on behalf of a consortium/group, it should be included in the author list separated with the wording “for” or “on behalf of.” The consortium/group will not retain authorship and will only appear in the author list.

Example: John Smith and Barbara Smith on behalf of The Collaborative Working Group.

1.11. Abstract

As a primary goal, the abstract should render the general significance and conceptual advance of the work clearly accessible to a broad readership. In the abstract, minimize the use of abbreviations and do not cite references, figures or tables.

For Clinical Trial articles, please include the Unique Identifier and the URL of the publicly accessible website on which the trial is registered.

1.12. Keywords

All article types require a minimum of 5 and a maximum of 8 keywords.

1.13. Text

The entire document should be single-spaced and must contain page and line numbers in order to facilitate the review process. The manuscript should be written using either Word or LaTeX. For templates, see [1.2. Templates](#).

1.14. Nomenclature

- The use of abbreviations should be kept to a minimum. Non-standard abbreviations should be avoided unless they appear at least four times, and defined upon first use in the main text. Consider also giving a list of non-standard abbreviations at the end, immediately before the Acknowledgments.
- Equations should be inserted in editable format from the equation editor.
- Italicize gene symbols and use the approved gene nomenclature where it is available. For human genes, please refer to the HUGO Gene Nomenclature Committee ([HGNC](#)). New gene symbols should be submitted [here](#). Common alternative gene aliases may also be reported, but should not be used alone in place of the HGNC symbol. Nomenclature committees for other species are listed [here](#). Protein products are not italicized.
- We encourage the use of Standard International Units in all manuscripts.
- Chemical compounds and biomolecules should be referred to using systematic nomenclature, preferably using the recommendations by IUPAC.
- Astronomical objects should be referred to using the nomenclature given by the International Astronomical Union provided [here](#).
- Life Science Identifiers (LSIDs) for ZOOBANK registered names or nomenclatural acts should be listed in the manuscript before the keywords. An LSID is represented as a uniform resource name (URN) with the following format:
urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]

For more information on LSIDs please see the [Code](#) section.

1.15. Sections

The manuscript is organized by headings and subheadings. The section headings should be those appropriate for your field and the research itself. You may insert up to 5 heading levels into your manuscript (i.e.,: 3.2.2.1.2 Heading Title).

For Original Research articles, it is recommended to organize your manuscript in the following sections or their equivalents for your field:

INTRODUCTION

Succinct, with no subheadings.

MATERIALS AND METHODS

This section may be divided by subheadings and should contain sufficient detail so that when read in conjunction with cited references, all procedures can be repeated. For experiments reporting results on animal or human subject research, an ethics approval statement should be included in this section (for further information, see the [Bioethics](#) section.)

RESULTS

This section may be divided by subheadings. Footnotes should not be used and must be transferred to the main text.

DISCUSSION

This section may be divided by subheadings. Discussions should cover the key findings of the study: discuss any prior research related to the subject to place the novelty of the discovery in the appropriate context, discuss the potential shortcomings and limitations on their interpretations, discuss their integration into the current understanding of the problem and how this advances the current views, speculate on the future direction of the research, and freely postulate theories that could be tested in the future.

For further information, please check the descriptions defined in the journal's "Article Types" page, which can be seen from the "For Authors" menu on any Frontiers journal page.

1.16. Acknowledgments

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors. Should the content of the manuscript have previously appeared online, such as in a thesis or preprint, this should be mentioned here, in addition to listing the source within the reference list.

1.17. Contribution to the Field Statement

When you submit your manuscript, you will be required to briefly summarize in 200 words your manuscript's contribution to, and position in, the existing literature in your field. This should be written avoiding any technical language or non-standard acronyms. The aim should be to convey the meaning and importance of this research to a non-expert. While Frontiers evaluates articles using objective criteria, rather than impact or novelty, your statement should frame the question(s) you have addressed in your work in the context of the current body of knowledge, providing evidence that the findings—whether positive or negative—contribute to progress in your research discipline. This will assist the Chief Editors to determine whether your manuscript fits within the scope of a specialty as defined in its mission statement; a detailed statement will also facilitate the identification of the editors and reviewers most appropriate to evaluate your work, ultimately expediting your manuscript's initial consideration.

Example Statement on: Markram K and Markram H (2010) The Intense World Theory – a unifying theory of the neurobiology of autism. *Front. Hum. Neurosci.* 4:224. doi: 10.3389/fnhum.2010.00224

Autism spectrum disorders are a group of neurodevelopmental disorders that affect up to 1 in 100 individuals. People with autism display an array of symptoms encompassing emotional processing, sociability, perception and memory, and present as uniquely as the individual. No theory has suggested a single underlying neuropathology to account for these diverse symptoms. The Intense World Theory, proposed here, describes a unifying pathology producing the wide spectrum of manifestations observed in autists. This theory focuses on the neocortex, fundamental for higher cognitive functions, and the limbic system, key for processing emotions and social signals. Drawing on discoveries in animal models and neuroimaging studies in individuals with autism, we propose how a combination of genetics, toxin exposure and/or environmental stress could produce hyper-reactivity and hyper-plasticity in the microcircuits involved with perception, attention, memory and emotionality. These hyper-functioning circuits will eventually come to dominate their neighbors, leading to hyper-sensitivity to incoming stimuli, over-specialization in tasks and a hyper-preference syndrome. We make the case that this theory of enhanced brain function in autism explains many of the varied past results and resolves conflicting findings and views and makes some testable experimental predictions.

2. Figure and Table Guidelines

2.1. CC-BY Licence

All figures, tables, and images will be published under a **Creative Commons CC-BY licence**, and permission must be obtained for use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

For additional information, please see the **Image Manipulation** section.

2.2. Figure Requirements and Style Guidelines

- Frontiers requires figures to be submitted individually, in the same order as they are referred to in the manuscript; the figures will then be automatically embedded at the end of the submitted manuscript. Kindly ensure that each figure is mentioned in the text and in numerical order.
- For figures with more than one panel, panels should be clearly indicated using labels (A), (B), (C), (D), etc. However, do not embed the part labels over any part of the image, these labels will be replaced during typesetting according to Frontiers' journal style. For graphs, there must be a self-explanatory label (including units) along each axis.
- For LaTeX files, figures should be included in the provided PDF. In case of acceptance, our Production Office might require high-resolution files of the figures included in the manuscript in EPS, JPEG or TIF/TIFF format.
- In order to be able to upload more than one figure at a time, save the figures (labeled in order of appearance in the manuscript) in a zip file and upload them as 'Supplementary Material Presentation.'

Please note that figures not in accordance with the guidelines will cause substantial delay during the production process.

2.2.1. Captions

Captions should be preceded by the appropriate label, for example "Figure 1." Figure captions should be placed at the end of the manuscript. Figure panels are referred to by bold capital letters in brackets: (A), (B), (C), (D), etc.

2.2.2. Image Size and Resolution Requirements

Figures should be prepared with the PDF layout in mind. Individual figures should not be longer than one page and with a width that corresponds to 1 column (85 mm) or 2 columns (180 mm).

All images must have a resolution of 300 dpi at final size. Check the resolution of your figure by enlarging it to 150%. If the image appears blurry, jagged or has a stair-stepped effect, the resolution is too low.

- The text should be legible and of high quality. The smallest visible text should be no less than 8 points in height when viewed at actual size.
- Solid lines should not be broken up. Any lines in the graphic should be no smaller than 2 points wide.

Please note that saving a figure directly as an image file (JPEG, TIF) can greatly affect the resolution of your image. To avoid this, one option is to export the file as PDF, then convert into TIFF or EPS using a graphics software.

2.2.3. Format and Color Image Mode

- The following formats are accepted: TIF/TIFF (.tif/.tiff), JPEG (.jpg), and EPS (.eps) (upon acceptance).
- Images must be submitted in the color mode RGB.

2.2.4. Chemical Structures

Chemical structures should be prepared using ChemDraw or a similar program. If working with ChemDraw please use our [Frontiers ChemDraw template](#). If working with another program please follow the guidelines given below:

- Drawing settings: chain angle, 120° bond spacing, 18% width; fixed length, 14.4 pt; bold width, 2.0 pt; line width, 0.6 pt; margin width, 1.6 pt; hash spacing, 2.5 pt. Scale 100% Atom Label settings: font, Arial; size, 8 pt.
- Assign all chemical compounds a bold, Arabic numeral in the order in which the compounds are presented in the manuscript text.

2.3. Table Requirements and Style Guidelines

- Tables should be inserted at the end of the manuscript in an editable format. If you use a word processor, build your table in Word. If you use a LaTeX processor, build your table in LaTeX. An empty line should be left before and after the table.
- Table captions must be placed immediately before the table. Captions should be preceded by the appropriate label, for example "Table 1." Please use only a single paragraph for the caption.
- Kindly ensure that each table is mentioned in the text and in numerical order.
- Please note that large tables covering several pages cannot be included in the final PDF for formatting reasons. These tables will be published as supplementary material.

Please note that tables which are not according to the guidelines will cause substantial delay during the production process.

2.4. Accessibility

Frontiers encourages authors to make the figures and visual elements of their articles accessible for the visually impaired. An effective use of color can help people with low visual acuity, or color blindness, understand all the content of an article.

These guidelines are easy to implement and are in accordance with the [W3C Web Content Accessibility Guidelines \(WCAG 2.1\)](#), the standard for web accessibility best practices.

A. Ensure sufficient contrast between text and its background

People who have low visual acuity or color blindness could find it difficult to read text with low contrast background color. Try using colors that provide maximum contrast.

WC3 recommends the following contrast ratio levels:

- Level AA, contrast ratio of at least 4.5:1
- Level AAA, contrast ratio of at least 7:1

Level AA
Contrast ratio 4.6:1

Level AA
Contrast ratio 9.5:1

You can verify the contrast ratio of your palette with these online ratio checkers:

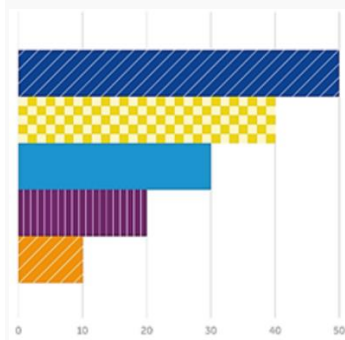
- [WebAIM](#)
- [Color Safe](#)

B. Avoid using red or green indicators

More than 99% of color-blind people have a red-green color vision deficiency.

C. Avoid using only color to communicate information

Elements with complex information like charts and graphs can be hard to read when only color is used to distinguish the data. Try to use other visual aspects to communicate information, such as shape, labels, and size. Incorporating patterns into the shape fills also make differences clearer; for an example please see below:



3. Supplementary Material

Data that are not of primary importance to the text, or which cannot be included in the article because they are too large or the current format does not permit it (such as videos, raw data traces, powerpoint presentations, etc.), can be uploaded as Supplementary Material during the submission procedure and will be displayed along with the published article. All supplementary files are deposited to Figshare for permanent storage and receive a DOI.

Supplementary Material is not typeset, so please ensure that all information is clearly presented without tracked changes/highlighted text/line numbers, and the appropriate caption is included in the file. To avoid discrepancies between the published article and the supplementary material, please do not add the title, author list, affiliations or correspondence in the supplementary files.

The Supplementary Material can be uploaded as Data Sheet (Word, Excel, CSV, CDX, FASTA, PDF or Zip files), Presentation (PowerPoint, PDF or Zip files), Image (CDX, EPS, JPEG, PDF, PNG or TIF/TIFF), Table (Word, Excel, CSV or PDF), Audio (MP3, WAV or WMA) or Video (AVI, DIVX, FLV, MOV, MP4, MPEG, MPG or WMV).

Technical requirements for Supplementary Images:

- 300 DPIs
- RGB color mode

For Supplementary Material templates (LaTeX and Word), see our [Supplementary Material templates](#).

4. References

Frontiers journals use one of two reference styles, either Harvard (Author-Date) or Vancouver (Numbered). Please check [this page](#) to find the correct style for your target journal.

- All citations in the text, figures or tables must be in the reference list and vice-versa.
- The names of the first six authors followed by et al. and the DOI (when available) should be provided.
- Given names of authors should be abbreviated to initials (e.g., Smith, J., Lewis, C.S., etc.)
- The reference list should only include articles that are published or accepted.
- Unpublished data, submitted manuscripts or personal communications should be cited within the text only, for the article types that allow such inclusions.
- For accepted but unpublished works use "in press" instead of page numbers.
- Data sets that have been deposited to an online repository should be included in the reference list. Include the version and unique identifier when available.
- Personal communications should be documented by a letter of permission.

- Website URLs should be included as footnotes.
- Any inclusion of verbatim text must be contained in quotation marks and clearly reference the original source.
- Preprints can be cited as long as a DOI or archive URL is available, and the citation clearly mentions that the contribution is a preprint. If a peer-reviewed journal publication for the same preprint exists, the official journal publication is the preferred source. See the [Preprints](#) section for more information.

4.1. Harvard Reference Style (Author-Date)

Many Frontiers journals use the Harvard referencing system, to find the correct reference style and resources for the journal you are submitting to please go to [this page](#). Reference examples are found below, for more examples of citing other documents and general questions regarding the Harvard reference style, please refer to the [Chicago Manual of Style](#).

4.1.1. In-text Citations

- For works by a single author, include the surname, followed by the year.
- For works by two authors, include both surnames, followed by the year.
- For works by more than two authors, include only the surname of the first author followed by et al., followed by the year.
- For Humanities and Social Sciences articles, include the page numbers.

4.1.2. Reference List

ARTICLE IN A PRINT JOURNAL

Sondheimer, N., and Lindquist, S. (2000). Rnq1: an epigenetic modifier of protein function in yeast. *Mol. Cell.* 5, 163-172.

ARTICLE IN AN ONLINE JOURNAL

Tahimic, C.G.T., Wang, Y., Bikle, D.D. (2013). Anabolic effects of IGF-1 signaling on the skeleton. *Front. Endocrinol.* 4:6. doi: 10.3389/fendo.2013.00006

ARTICLE OR CHAPTER IN A BOOK

Sorenson, P. W., and Caprio, J. C. (1998). "Chemoreception," in *The Physiology of Fishes*, ed. D. H. Evans (Boca Raton, FL: CRC Press), 375-405.

BOOK

Cowan, W. M., Jessell, T. M., and Zipursky, S. L. (1997). *Molecular and Cellular Approaches to Neural Development*. New York: Oxford University Press.

ABSTRACT

Hendricks, J., Applebaum, R., and Kunkel, S. (2010). A world apart? Bridging the gap between theory and applied social gerontology. *Gerontologist* 50, 284-293. Abstract retrieved from Abstracts in Social Gerontology database. (Accession No. 50360869)

WEBSITE

World Health Organization. (2018). E. coli. <https://www.who.int/news-room/fact-sheets/detail/e-coli> [Accessed March 15, 2018].

PATENT

Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. U.S. Patent No 6,090,051. Washington, DC: U.S. Patent and Trademark Office.

DATA

Perdiguero P, Venturas M, Cervera MT, Gil L, Collada C. Data from: Massive sequencing of Ulms minor's transcriptome provides new molecular tools for a genus under the constant threat of Dutch elm disease. Dryad Digital Repository. (2015) <http://dx.doi.org/10.5061/dryad.ps837>

THESES AND DISSERTATIONS

Smith, J. (2008) Post-structuralist discourse relative to phenomenological pursuits in the deconstructivist arena. [dissertation/master's thesis]. [Chicago (IL)]: University of Chicago

PREPRINT

Smith, J. (2008). Title of the document. Preprint repository name [Preprint]. Available at: <https://persistent-url> (Accessed March 15, 2018).

4.2. Vancouver Reference Style (Numbered)

Many Frontiers journals use the numbered referencing system, to find the correct reference style and resources for the journal you are submitting to please go to [this page](#).

Reference examples are found below, for more examples of citing other documents and general questions regarding the Vancouver reference style, please refer to [Citing Medicine](#).

4.2.1. In-text Citations

- Please apply the Vancouver system for in-text citations.
- In-text citations should be numbered consecutively in order of appearance in the text—identified by Arabic numerals in the parenthesis (use square brackets for Physics and Mathematics articles).

4.2.2. Reference List**ARTICLE IN A PRINT JOURNAL**

Sondheimer N, Lindquist S. Rnq1: an epigenetic modifier of protein function in yeast. *Mol Cell* (2000) 5:163-72.

ARTICLE IN AN ONLINE JOURNAL

Tahimic CGT, Wang Y, Bikle DD. Anabolic effects of IGF-1 signaling on the skeleton. *Front Endocrinol* (2013) 4:6. doi: 10.3389/fendo.2013.00006

ARTICLE OR CHAPTER IN A BOOK

Sorenson PW, Caprio JC. "Chemoreception,". In: Evans DH, editor. *The Physiology of Fishes*. Boca Raton, FL: CRC Press (1998). p. 375-405.

BOOK

Cowan WM, Jessell TM, Zipursky SL. *Molecular and Cellular Approaches to Neural Development*. New York: Oxford University Press (1997). 345 p.

ABSTRACT

Christensen S, Oppacher F. An analysis of Koza's computational effort statistic for genetic programming. In: Foster JA, editor. *Genetic Programming. EuroGP 2002: Proceedings of the 5th European Conference on Genetic Programming; 2002 Apr 3–5; Kinsdale, Ireland*. Berlin: Springer (2002). p. 182–91.

WEBSITE

World Health Organization. *E. coli* (2018). <https://www.who.int/news-room/fact-sheets/detail/e-coli> [Accessed March 15, 2018].

PATENT

Pagedas AC, inventor; Ancel Surgical R&D Inc., assignee. *Flexible Endoscopic Grasping and Cutting Device and Positioning Tool Assembly*. United States patent US 20020103498 (2002).

DATA

Perdiguero P, Venturas M, Cervera MT, Gil L, Collada C. Data from: Massive sequencing of Ulms minor's transcriptome provides new molecular tools for a genus under the constant threat of Dutch elm disease. Dryad Digital Repository. (2015) <http://dx.doi.org/10.5061/dryad.ps837>

THESES AND DISSERTATIONS

Smith, J. (2008) Post-structuralist discourse relative to phenomenological pursuits in the deconstructivist arena. [dissertation/master's thesis]. [Chicago (IL)]: University of Chicago

PREPRINT

Smith, J. Title of the document. Preprint repository name [Preprint] (2008). Available at: <https://persistent-url> (Accessed March 15, 2018).

7.2 Normas da revista *Genes*

Instructions for Authors

Shortcuts

- [Manuscript Submission Overview](#)
- [Manuscript Preparation](#)
- [Preparing Figures, Schemes and Tables](#)
- [Original Images for Blots and Gels Requirements](#)
- [Supplementary Materials, Data Deposit and Software Source Code](#)
- [Research and Publication Ethics](#)
- [Reviewer Suggestions](#)
- [English Corrections](#)
- [Preprints and Conference Papers](#)
- [Authorship](#)
- [Editorial Independence](#)
- [Conflict of Interests](#)
- [Editorial Procedures and Peer-Review](#)
- [Promoting Equity, Diversity and Inclusiveness Within MDPI Journals](#)
- [Resource Identification Initiative](#)

Submission Checklist

Please:

1. read the [Aims & Scope](#) to gain an overview and assess if your manuscript is suitable for this journal;
2. use the [Microsoft Word template](#) or [LaTeX template](#) to prepare your manuscript;
3. make sure that issues about [publication ethics](#), [research ethics](#), [copyright](#), [authorship](#), [figure formats](#), [data](#) and [references format](#) have been appropriately considered;
4. Ensure that all authors have approved the content of the submitted manuscript.
5. Authors are encouraged to add a [biography](#) (optional) to the submission and publish it.

Manuscript Submission Overview

Types of Publications

Genes has no restrictions on the length of manuscripts, provided that the text is concise and comprehensive. Full experimental details must be provided so that the results can be reproduced. *Genes* requires that authors publish all experimental controls and make full datasets available where possible (see the guidelines on [Supplementary Materials](#) and references to unpublished data).

Manuscripts submitted to *Genes* should neither be published previously nor be under consideration for publication in another journal. The main article types are as follows:

- *Articles*: Original research manuscripts. The journal considers all original research manuscripts provided that the work reports scientifically sound experiments and provides a substantial amount of new information. Authors should not unnecessarily divide their work into several related manuscripts, although short *Communications* of preliminary, but significant, results will be considered. The quality and impact of the study will be considered during peer review.
- *Reviews*: These provide concise and precise updates on the latest progress made in a given area of research. Systematic reviews should follow the PRISMA [guidelines](#).

Submission Process

Manuscripts for *Genes* should be submitted online at susy.mdpi.com. The submitting author, who is generally the corresponding author, is responsible for the manuscript during the submission and peer-review process. The submitting author must ensure that all eligible co-authors have been included in the author list (read the [criteria to qualify for authorship](#)) and that they have all read and approved the submitted version of the manuscript. To submit your manuscript, register and log in to the [submission website](#). Once you have registered, [click here to go to the submission form for Genes](#). All co-authors can see the manuscript details in the submission system, if they register and log in using the e-mail address provided during manuscript submission.

Accepted File Formats

Authors must use the [Microsoft Word template](#) or [LaTeX template](#) to prepare their manuscript. Using the template file will substantially shorten the time to complete copy-editing and publication of accepted manuscripts. The total amount of data for all files must not exceed 120 MB. If this is a problem, please contact the Editorial Office genes@mdpi.com. Accepted file formats are:

- *Microsoft Word*: Manuscripts prepared in Microsoft Word must be converted into a single file before submission. When preparing manuscripts in Microsoft Word, the [Genes Microsoft Word template file](#) must be used. Please insert your graphics (schemes, figures, *etc.*) in the main text after the paragraph of its first citation.
- *LaTeX*: Manuscripts prepared in LaTeX must be collated into one ZIP folder (including all source files and images, so that the Editorial Office can recompile the submitted PDF). When preparing manuscripts in LaTeX, please use the [Genes LaTeX template files](#). You can now also use the online application [writeLaTeX](#) to submit articles directly to *Genes*. The MDPI LaTeX template file should be selected from the [writeLaTeX template gallery](#).
- *Supplementary files*: May be any format, but it is recommended that you use common, non-proprietary formats where possible (see [below](#) for further details).

Disclaimer: Usage of these templates is exclusively intended for submission to the journal for peer-review, and strictly limited to this purpose and it cannot be used for posting online on preprint servers or other websites.

Free Format Submission

Genes now accepts free format submission:

- We do not have strict formatting requirements, but all manuscripts must contain the required sections: Author Information, Abstract, Keywords, Introduction, Materials & Methods, Results, Conclusions, Figures and Tables with Captions, Funding Information, Author Contributions, Conflict of Interest and other Ethics Statements. Check the Journal [Instructions for Authors](#) for more details.
- Your references may be in any style, provided that you use the consistent formatting throughout. It is essential to include author(s) name(s), journal or book title, article or chapter title (where required), year of publication, volume and issue (where appropriate) and pagination. DOI numbers (Digital Object Identifier) are not mandatory but highly encouraged. The bibliography software package *EndNote*, [Zotero](#), *Mendeley*, *Reference Manager* are recommended.
- When your manuscript reaches the revision stage, you will be requested to format the manuscript according to the journal guidelines.

Cover Letter

A cover letter must be included with each manuscript submission. It should be concise and explain why the content of the paper is significant, placing the findings in the context of existing work. It should explain why the manuscript fits the scope of the journal.

Any prior submissions of the manuscript to MDPI journals must be acknowledged. If this is the case, it is strongly recommended that the previous manuscript ID is provided in the submission system, which will ease your current submission process. The names of proposed and excluded reviewers should be provided in the submission system, not in the cover letter.

All cover letters are required to include the statements:

- We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal.
- All authors have approved the manuscript and agree with its submission to (journal name).

Author Biography

Authors are encouraged to add a biography (maximum 150 words) to the submission and publish it. This should be a single paragraph and should contain the following points:

1. Authors' full names followed by current positions;
2. Education background including institution information and year of graduation (type and level of degree received);
3. Work experience;
4. Current and previous research interests;
5. Memberships of professional societies and awards received.

Note for Authors Funded by the National Institutes of Health (NIH)

This journal automatically deposits papers to PubMed Central after publication of an issue. Authors do not need to separately submit their papers through the NIH Manuscript Submission System (NIHMS, <http://nihms.nih.gov/>).

[\[Return to top\]](#)

Manuscript Preparation

General Considerations

- **Research manuscripts** should comprise:
 - **Front matter:** Title, Author list, Affiliations, Abstract, Keywords

- **Research manuscript sections:** Introduction, Materials and Methods, Results, Discussion, Conclusions (optional).
- **Back matter:** Supplementary Materials, Acknowledgments, Author Contributions, Conflicts of Interest, [References](#).
- **Review manuscripts** should comprise the [front matter](#), literature review sections and the [back matter](#). The template file can also be used to prepare the front and back matter of your review manuscript. It is not necessary to follow the remaining structure. Structured reviews and meta-analyses should use the same structure as research articles and ensure they conform to the [PRISMA](#) guidelines.
- **Graphical Abstract:**

A graphical abstract (GA) is an image that appears alongside the text abstract in the Table of Contents. In addition to summarizing the content, it should represent the topic of the article in an attention-grabbing way. Moreover, it should not be exactly the same as the Figure in the paper or just a simple superposition of several subfigures. Note that the GA must be original and unpublished artwork. Any postage stamps, currency from any country, or trademarked items should not be included in it.

The GA should be a high-quality illustration or diagram in any of the following formats: PNG, JPEG, TIFF, or SVG. Written text in a GA should be clear and easy to read, using one of the following fonts: Times, Arial, Courier, Helvetica, Ubuntu or Calibri.

The minimum required size for the GA is 560 × 1100 pixels (height × width). The size should be of high quality in order to reproduce well.
- **Acronyms/Abbreviations/Initialisms** should be defined the first time they appear in each of three sections: the abstract; the main text; the first figure or table. When defined for the first time, the acronym/abbreviation/initialism should be added in parentheses after the written-out form.
- **SI Units** (International System of Units) should be used. Imperial, US customary and other units should be converted to SI units whenever possible.
- **Accession numbers** of RNA, DNA and protein sequences used in the manuscript should be provided in the Materials and Methods section. Also see the section on [Deposition of Sequences and of Expression Data](#).
- **Equations:** If you are using Word, please use either the Microsoft Equation Editor or the MathType add-on. Equations should be editable by the editorial office and not appear in a picture format.
- **Research Data and supplementary materials:** Note that publication of your manuscript implies that you must make all materials, data, and protocols associated with the publication available to readers. Disclose at the submission stage any restrictions on the availability of materials or information. Read the information about [Supplementary Materials](#) and Data Deposit for additional guidelines.
- **Preregistration:** Where authors have preregistered studies or analysis plans, links to the preregistration must be provided in the manuscript.
- **Guidelines and standards:** MDPI follows standards and guidelines for certain types of research. See https://www.mdpi.com/editorial_process for further information.

[\[Return to top\]](#)

Front Matter

These sections should appear in all manuscript types

- **Title:** The title of your manuscript should be concise, specific and relevant. It should identify if the study reports (human or animal) trial data, or is a systematic review,

meta-analysis or replication study. When gene or protein names are included, the abbreviated name rather than full name should be used.

- **Author List and Affiliations:** Authors' full first and last names must be provided. The initials of any middle names can be added. The PubMed/MEDLINE standard format is used for affiliations: complete address information including city, zip code, state/province, and country. At least one author should be designated as corresponding author, and his or her email address and other details should be included at the end of the affiliation section. Please read the [criteria to qualify for authorship](#).
- **Abstract:** The abstract should be a total of about 200 words maximum. The abstract should be a single paragraph and should follow the style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlight the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied. Include any relevant preregistration numbers, and species and strains of any animals used. 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article: it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.
- **Keywords:** Three to ten pertinent keywords need to be added after the abstract. We recommend that the keywords are specific to the article, yet reasonably common within the subject discipline.

Research Manuscript Sections

- **Introduction:** The introduction should briefly place the study in a broad context and highlight why it is important. It should define the purpose of the work and its significance, including specific hypotheses being tested. The current state of the research field should be reviewed carefully and key publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work and highlight the main conclusions. Keep the introduction comprehensible to scientists working outside the topic of the paper.
- **Materials and Methods:** They should be described with sufficient detail to allow others to replicate and build on published results. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited. Give the name and version of any software used and make clear whether computer code used is available. Include any pre-registration codes.
- **Results:** Provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.
- **Discussion:** Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible and limitations of the work highlighted. Future research directions may also be mentioned. This section may be combined with Results.
- **Conclusions:** This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.
- **Patents:** This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

[\[Return to top\]](#)

Back Matter

- **Supplementary Materials:** Describe any supplementary material published online alongside the manuscript (figure, tables, video, spreadsheets, etc.). Please indicate the name and title of each element as follows Figure S1: title, Table S1: title, etc.

- Funding:** All sources of funding of the study should be disclosed. Clearly indicate grants that you have received in support of your research work and if you received funds to cover publication costs. Note that some funders will not refund article processing charges (APC) if the funder and grant number are not clearly and correctly identified in the paper. Funding information can be entered separately into the submission system by the authors during submission of their manuscript. Such funding information, if available, will be deposited to FundRef if the manuscript is finally published.

Please add: “This research received no external funding” or “This research was funded by [name of funder] grant number [xxx]” and “The APC was funded by [XXX]” in this section. Check carefully that the details given are accurate and use the standard spelling of funding agency names at <https://search.crossref.org/funding>, any errors may affect your future funding.
- Acknowledgments:** In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).
- Author Contributions:** Each author is expected to have made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; or have drafted the work or substantively revised it; AND has approved the submitted version (and version substantially edited by journal staff that involves the author’s contribution to the study); AND agrees to be personally accountable for the author’s own contributions and for ensuring that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and documented in the literature.

For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; Methodology, X.X.; Software, X.X.; Validation, X.X., Y.Y. and Z.Z.; Formal Analysis, X.X.; Investigation, X.X.; Resources, X.X.; Data Curation, X.X.; Writing – Original Draft Preparation, X.X.; Writing – Review & Editing, X.X.; Visualization, X.X.; Supervision, X.X.; Project Administration, X.X.; Funding Acquisition, Y.Y.”, please turn to the [CRediT taxonomy](#) for the term explanation. For more background on CRediT, see [here](#). **Authorship must include and be limited to those who have contributed substantially to the work. Please read the section concerning the [criteria to qualify for authorship](#) carefully**.
- Institutional Review Board Statement:** In this section, please add the Institutional Review Board Statement and approval number for studies involving humans or animals. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” OR “Ethical review and approval were waived for this study, due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals. You might also choose to exclude this statement if the study did not involve humans or animals.
- Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients

who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

- **Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section “[MDPI Research Data Policies](#)”. You might choose to exclude this statement if the study did not report any data.
- **Conflicts of Interest:** Authors must identify and declare any personal circumstances or interest that may be perceived as influencing the representation or interpretation of reported research results. If there is no conflict of interest, please state “The authors declare no conflict of interest.” Any role of the funding sponsors in the choice of research project; design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. *Genes* does not publish studies funded partially or fully by the tobacco industry. Any projects funded by industry must pay special attention to the full declaration of funder involvement. If there is no role, please state “The sponsors had no role in the design, execution, interpretation, or writing of the study”. For more details please see [Conflict of Interest](#).
- **References:** References must be numbered in order of appearance in the text (including table captions and figure legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as [EndNote](#), [ReferenceManager](#) or [Zotero](#) to avoid typing mistakes and duplicated references. We encourage citations to data, computer code and other citable research material. If available online, you may use reference style 9. below.
- Citations and References in Supplementary files are permitted provided that they also appear in the main text and in the reference list.

In the text, reference numbers should be placed in square brackets [], and placed before the punctuation; for example [1], [1–3] or [1,3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 10). or [6] (pp. 101–105).

The reference list should include the full title, as recommended by the ACS style guide. Style files for [Endnote](#) and [Zotero](#) are available.

References should be described as follows, depending on the type of work:

Journal Articles:

1. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name* **Year**, *Volume*, page range.

Books and Book Chapters:

2. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, Year; pp. 154–196.

3. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, Year; Volume 3, pp. 154–196.

Unpublished materials intended for publication:

4. Author 1, A.B.; Author 2, C. Title of Unpublished Work (optional). Correspondence Affiliation, City, State, Country. year, *status (manuscript in preparation; to be submitted)*.

5. Author 1, A.B.; Author 2, C. Title of Unpublished Work. *Abbreviated Journal Name* year, *phrase indicating stage of publication (submitted; accepted; in press)*.

Unpublished materials not intended for publication:

6. Author 1, A.B. (Affiliation, City, State, Country); Author 2, C. (Affiliation, City, State, Country). Phase describing the material, year. (phase: Personal communication; Private communication; Unpublished work; etc.)

□ Conference Proceedings:

7. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In *Title of the Collected Work* (if available), Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference; Editor 1, Editor 2, Eds. (if available); Publisher: City, Country, Year (if available); Abstract Number (optional), Pagination (optional).

□ Thesis:

8. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.

□ Websites:

9. Title of Site. Available online: URL (accessed on Day Month Year).

Unlike published works, websites may change over time or disappear, so we encourage you create an archive of the cited website using a service such as [WebCite](#). Archived websites should be cited using the link provided as follows:

10. Title of Site. URL (archived on Day Month Year).

See the [Reference List and Citations Guide](#) for more detailed information.

[\[Return to top\]](#)

Preparing Figures, Schemes and Tables

- File for Figures and Schemes must be provided during submission in a single zip archive and at a sufficiently high resolution (minimum 1000 pixels width/height, or a resolution of 300 dpi or higher). Common formats are accepted, however, TIFF, JPEG, EPS and PDF are preferred.
- *Genes* can publish multimedia files in articles or as supplementary materials. Please contact the editorial office for further information.
- All Figures, Schemes and Tables should be inserted into the main text close to their first citation and must be numbered following their number of appearance (Figure 1, Scheme I, Figure 2, Scheme II, Table 1, *etc.*).
- All Figures, Schemes and Tables should have a short explanatory title and caption.
- All table columns should have an explanatory heading. To facilitate the copy-editing of larger tables, smaller fonts may be used, but no less than 8 pt. in size. Authors should use the Table option of Microsoft Word to create tables.
- Authors are encouraged to prepare figures and schemes in color (RGB at 8-bit per channel). There is no additional cost for publishing full color graphics.

[\[Return to top\]](#)

Original Images for Blots and Gels Requirements

For the main text, please ensure that:

- All experimental samples and controls used for one comparative analysis are run on the same blot/gel.
- Image processing methods, such as adjusting the brightness or contrast, do not alter or distort the information in the figure and are applied to every pixel. High-contrast blots/gels are discouraged.
- Cropped blots/gels present in the main text retain all important information and bands.
- You have checked figures for duplications and ensured the figure legends are clear and accurate. Please include all relevant information in the figure legends and clearly indicate any re-arrangement of lanes.

In order to ensure the integrity and scientific validity of blots (including, but not limited to, Western blots) and the reporting of gel data, original, uncropped and unadjusted images should be uploaded as Supporting Information files at the time of initial submission.

A single PDF file or a zip folder including all the original images reported in the main figure and supplemental figures should be prepared. Authors should annotate each original image, corresponding to the figure in the main article or supplementary materials, and label each lane or loading order. All experimental samples and controls used for one comparative analysis should be run on the same blot/gel image. For quantitative analyses, please provide the blots/gels for each independent biological replicate used in the analysis.

[\[Return to top\]](#)

Supplementary Materials, Data Deposit and Software Source Code

MDPI Research Data Policies

MDPI is committed to supporting open scientific exchange and enabling our authors to achieve best practices in sharing and archiving research data. We encourage all authors of articles published in MDPI journals to share their research data. Individual journal guidelines can be found at the journal 'Instructions for Authors' page. Data sharing policies concern the minimal dataset that supports the central findings of a published study. Generated data should be publicly available and cited in accordance with journal guidelines.

MDPI data policies are informed by [TOP Guidelines](#) and [FAIR Principles](#).

Where ethical, legal or privacy issues are present, data should not be shared. The authors should make any limitations clear in the Data Availability Statement upon submission. Authors should ensure that data shared are in accordance with consent provided by participants on the use of confidential data.

Data Availability Statements provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study.

Below are suggested Data Availability Statements:

- Data available in a publicly accessible repository
The data presented in this study are openly available in [repository name e.g., FigShare] at [[doi](#)], reference number [reference number].
- Data available in a publicly accessible repository that does not issue DOIs
Publicly available datasets were analyzed in this study. This data can be found here: [link/accession number]
- Data available on request due to restrictions eg privacy or ethical
The data presented in this study are available on request from the corresponding author. The data are not publicly available due to [insert reason here]
- 3rd Party Data
Restrictions apply to the availability of these data. Data was obtained from [third party] and are available [from the authors / at URL] with the permission of [third party].
- Data sharing not applicable
No new data were created or analyzed in this study. Data sharing is not applicable to this article.
- Data is contained within the article or supplementary material
The data presented in this study are available in [insert article or supplementary material here]

Data citation:

- [dataset] Authors. Year. Dataset title; Data repository or archive; Version (if any); Persistent identifier (e.g., DOI).

Computer Code and Software

For work where novel computer code was developed, authors should release the code either by depositing in a recognized, public repository such as [GitHub](#) or uploading as supplementary

information to the publication. The name, version, corporation and location information for all software used should be clearly indicated. Please include all the parameters used to run software/programs analyses.

Supplementary Material

Additional data and files can be uploaded as "Supplementary Files" during the manuscript submission process. The supplementary files will also be available to the referees as part of the peer-review process. Any file format is acceptable; however, we recommend that common, non-proprietary formats are used where possible.

References in Supplementary Files

Citations and References in Supplementary files are permitted provided that they also appear in the reference list of the main text.

Unpublished Data

Restrictions on data availability should be noted during submission and in the manuscript. "Data not shown" should be avoided: authors are encouraged to publish all observations related to the submitted manuscript as Supplementary Material. "Unpublished data" intended for publication in a manuscript that is either planned, "in preparation" or "submitted" but not yet accepted, should be cited in the text and a reference should be added in the References section. "Personal Communication" should also be cited in the text and reference added in the References section. (see also the MDPI reference list and citations style guide).

Remote Hosting and Large Data Sets

Data may be deposited with specialized service providers or institutional/subject repositories, preferably those that use the DataCite mechanism. Large data sets and files greater than 60 MB must be deposited in this way. For a list of other repositories specialized in scientific and experimental data, please consult databib.org or re3data.org. The data repository name, link to the data set (URL) and accession number, doi or handle number of the data set must be provided in the paper. The journal [Data](#) also accepts submissions of data set papers.

Deposition of Sequences and of Expression Data

New sequence information must be deposited to the appropriate database prior to submission of the manuscript. Accession numbers provided by the database should be included in the submitted manuscript. Manuscripts will not be published until the accession number is provided.

- *New nucleic acid sequences* must be deposited in one of the following databases: [GenBank](#), [EMBL](#), or [DDBJ](#). Sequences should be submitted to only one database.
- *New high throughput sequencing (HTS) datasets* (RNA-seq, ChIP-Seq, degradome analysis, ...) must be deposited either in the [GEO database](#) or in the NCBI's [Sequence Read Archive \(SRA\)](#).
- *New microarray data* must be deposited either in the [GEO](#) or the [ArrayExpress](#) databases. The "Minimal Information About a Microarray Experiment" (MIAME) guidelines published by the Microarray Gene Expression Data Society must be followed.
- *New protein sequences* obtained by protein sequencing must be submitted to UniProt (submission tool [SPIN](#)). Annotated protein structure and its reference sequence must be submitted to [RCSB of Protein Data Bank](#).

All sequence names and the accession numbers provided by the databases must be provided in the Materials and Methods section of the article.

Deposition of Proteomics Data

Methods used to generate the proteomics data should be described in detail and we encourage authors to adhere to the "[Minimum Information About a Proteomics Experiment](#)". All

generated mass spectrometry raw data must be deposited in the appropriate public database such as [ProteomeXchange](#), [PRIDE](#) or [jPOST](#). At the time of submission, please include all relevant information in the materials and methods section, such as repository where the data was submitted and link, data set identifier, username and password needed to access the data.

[\[Return to top\]](#)

Research and Publication Ethics

Research Ethics

Research Involving Human Subjects

When reporting on research that involves human subjects, human material, human tissues, or human data, authors must declare that the investigations were carried out following the rules of the Declaration of Helsinki of 1975 (<https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/>), revised in 2013. According to point 23 of this declaration, an approval from the local institutional review board (IRB) or other appropriate ethics committee must be obtained before undertaking the research to confirm the study meets national and international guidelines. As a minimum, a statement including the project identification code, date of approval, and name of the ethics committee or institutional review board must be stated in Section ‘Institutional Review Board Statement’ of the article.

Example of an ethical statement: "All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of XXX (Project identification code)."

For non-interventional studies (e.g. surveys, questionnaires, social media research), all participants must be fully informed if the anonymity is assured, why the research is being conducted, how their data will be used and if there are any risks associated. As with all research involving humans, ethical approval from an appropriate ethics committee must be obtained prior to conducting the study. If ethical approval is not required, authors must either provide an exemption from the ethics committee or are encouraged to cite the local or national legislation that indicates ethics approval is not required for this type of study. Where a study has been granted exemption, the name of the ethics committee which provided this should be stated in Section ‘Institutional Review Board Statement’ with a full explanation regarding why ethical approval was not required.

A written informed consent for publication must be obtained from participating patients. Data relating to individual participants must be described in detail, but private information identifying participants need not be included unless the identifiable materials are of relevance to the research (for example, photographs of participants’ faces that show a particular symptom). Patients’ initials or other personal identifiers must not appear in any images. For manuscripts that include any case details, personal information, and/or images of patients, authors must obtain signed informed consent for publication from patients (or their relatives/guardians) before submitting to an MDPI journal. Patient details must be anonymized as far as possible, e.g., do not mention specific age, ethnicity, or occupation where they are not relevant to the conclusions. A [template permission form](#) is available to download. A blank version of the form used to obtain permission (without the patient names or signature) must be uploaded with your submission. Editors reserve the right to reject any submission that does not meet these requirements.

You may refer to our sample form and provide an appropriate form after consulting with your affiliated institution. For the purposes of publishing in MDPI journals, a consent, permission, or release form should include unlimited permission for publication in all formats (including print, electronic, and online), in sublicensed and reprinted versions (including translations and derived works), and in other works and products under open access license. To respect patients’ and any other individual’s privacy, please do not send signed forms. The journal reserves the right to ask authors to provide signed forms if necessary.

If the study reports research involving vulnerable groups, an additional check may be performed. The submitted manuscript will be scrutinized by the editorial office and upon request, documentary evidence (blank consent forms and any related discussion documents from the ethics board) must be supplied. Additionally, when studies describe groups by race, ethnicity, gender, disability, disease, etc., explanation regarding why such categorization was needed must be clearly stated in the article.

Ethical Guidelines for the Use of Animals in Research

The editors will require that the benefits potentially derived from any research causing harm to animals are significant in relation to any cost endured by animals, and that procedures followed are unlikely to cause offense to the majority of readers. Authors should particularly ensure that their research complies with the commonly-accepted '3Rs [1]':

- Replacement of animals by alternatives wherever possible,
- Reduction in number of animals used, and
- Refinement of experimental conditions and procedures to minimize the harm to animals.

Authors must include details on housing, husbandry and pain management in their manuscript.

For further guidance authors should refer to the Code of Practice for the Housing and Care of Animals Used in Scientific Procedures [2], American Association for Laboratory Animal Science [3] or European Animal Research Association [4].

If national legislation requires it, studies involving vertebrates or higher invertebrates must only be carried out after obtaining approval from the appropriate ethics committee. As a minimum, the project identification code, date of approval and name of the ethics committee or institutional review board should be stated in Section 'Institutional Review Board Statement'. Research procedures must be carried out in accordance with national and institutional regulations. Statements on animal welfare should confirm that the study complied with all relevant legislation. Clinical studies involving animals and interventions outside of routine care require ethics committee oversight as per the American Veterinary Medical Association. If the study involved client-owned animals, informed client consent must be obtained and certified in the manuscript report of the research. Owners must be fully informed if there are any risks associated with the procedures and that the research will be published. If available, a high standard of veterinary care must be provided. Authors are responsible for correctness of the statements provided in the manuscript.

If ethical approval is not required by national laws, authors must provide an exemption from the ethics committee, if one is available. Where a study has been granted exemption, the name of the ethics committee that provided this should be stated in Section 'Institutional Review Board Statement' with a full explanation on why the ethical approval was not required.

If no animal ethics committee is available to review applications, authors should be aware that the ethics of their research will be evaluated by reviewers and editors. Authors should provide a statement justifying the work from an ethical perspective, using the same utilitarian framework that is used by ethics committees. Authors may be asked to provide this even if they have received ethical approval.

MDPI endorses the ARRIVE guidelines (arriveguidelines.org/) for reporting experiments using live animals. Authors and reviewers must use the ARRIVE guidelines as a checklist, which can be found at <https://arriveguidelines.org/sites/arrive/files/documents/ARRIVE%20Compliance%20Questionnaire.pdf>. Editors reserve the right to ask for the checklist and to reject submissions that do not adhere to these guidelines, to reject submissions based on ethical or animal welfare concerns or if the procedure described does not appear to be justified by the value of the work presented.

1. NSW Department of Primary Industries and Animal Research Review Panel. Three Rs. Available online: <https://www.animaethics.org.au/three-rs>
2. Home Office. Animals (Scientific Procedures) Act 1986. Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for Scientific Purposes. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/388535/CoPanimalsWeb.pdf
3. American Association for Laboratory Animal Science. The Scientific Basis for Regulation of Animal Care and Use. Available online: <https://www.aalas.org/about-aalas/position-papers/scientific-basis-for-regulation-of-animal-care-and-use>
4. European Animal Research Association. EU regulations on animal research. Available online: <https://www.eara.eu/animal-research-law>

Research Involving Cell Lines

Methods sections for submissions reporting on research with cell lines should state the origin of any cell lines. For established cell lines the provenance should be stated and references must also be given to either a published paper or to a commercial source. If previously unpublished *de novo* cell lines were used, including those gifted from another laboratory, details of institutional review board or ethics committee approval must be given, and confirmation of written informed consent must be provided if the line is of human origin.

An example of Ethical Statements:

The HCT116 cell line was obtained from XXXX. The MLH1⁺ cell line was provided by XXXXX, Ltd. The DLD-1 cell line was obtained from Dr. XXXX. The DR-GFP and SA-GFP reporter plasmids were obtained from Dr. XXX and the Rad51K133A expression vector was obtained from Dr. XXXX.

Research Involving Plants

Experimental research on plants (either cultivated or wild) including collection of plant material, must comply with institutional, national, or international guidelines. We recommend that authors comply with the [Convention on Biological Diversity](#) and the [Convention on the Trade in Endangered Species of Wild Fauna and Flora](#).

For each submitted manuscript supporting genetic information and origin must be provided. For research manuscripts involving rare and non-model plants (other than, e.g., *Arabidopsis thaliana*, *Nicotiana benthamiana*, *Oryza sativa*, or many other typical model plants), voucher specimens must be deposited in an accessible herbarium or museum. Vouchers may be requested for review by future investigators to verify the identity of the material used in the study (especially if taxonomic rearrangements occur in the future). They should include details of the populations sampled on the site of collection (GPS coordinates), date of collection, and document the part(s) used in the study where appropriate. For rare, threatened or endangered species this can be waived but it is necessary for the author to describe this in the cover letter.

Editors reserve the rights to reject any submission that does not meet these requirements.

An example of Ethical Statements:

Torenia fournieri plants were used in this study. White-flowered Crown White (CrW) and violet-flowered Crown Violet (CrV) cultivars selected from 'Crown Mix' (XXX Company, City, Country) were kindly provided by Dr. XXX (XXX Institute, City, Country).

Arabidopsis mutant lines (SALKxxxx, SAILxxxx,...) were kindly provided by Dr. XXX, institute, city, country).

Clinical Trials Registration

Registration

MDPI follows the International Committee of Medical Journal Editors (ICMJE) [guidelines](#) which require and recommend registration of clinical trials in a public trials

registry at or before the time of first patient enrollment as a condition of consideration for publication.

Purely observational studies do not require registration. A clinical trial not only refers to studies that take place in a hospital or involve pharmaceuticals, but also refer to all studies which involve participant randomization and group classification in the context of the intervention under assessment.

Authors are strongly encouraged to pre-register clinical trials with an international clinical trials register and cite a reference to the registration in the Methods section. Suitable databases include clinicaltrials.gov, [the EU Clinical Trials Register](http://www.eu-clinical-trials-register.eu) and those listed by the World Health Organisation [International Clinical Trials Registry Platform](http://www.international-clinical-trials-registry-platform.org).

Approval to conduct a study from an independent local, regional, or national review body is not equivalent to prospective clinical trial registration. MDPI reserves the right to decline any paper without trial registration for further peer-review. However, if the study protocol has been published before the enrolment, the registration can be waived with correct citation of the published protocol.

CONSORT Statement

MDPI requires a completed CONSORT 2010 [checklist](#) and [flow diagram](#) as a condition of submission when reporting the results of a randomized trial. Templates for these can be found here or on the CONSORT website (<http://www.consort-statement.org>) which also describes several CONSORT checklist extensions for different designs and types of data beyond two group parallel trials. At minimum, your article should report the content addressed by each item of the checklist.

[\[Return to top\]](#)

Sex and Gender in Research

We encourage our authors to follow the [‘Sex and Gender Equity in Research – SAGER – guidelines’](#) and to include sex and gender considerations where relevant. Authors should use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Article titles and/or abstracts should indicate clearly what sex(es) the study applies to. Authors should also describe in the background, whether sex and/or gender differences may be expected; report how sex and/or gender were accounted for in the design of the study; provide disaggregated data by sex and/or gender, where appropriate; and discuss respective results. If a sex and/or gender analysis was not conducted, the rationale should be given in the Discussion. We suggest that our authors consult the full [guidelines](#) before submission.

[\[Return to top\]](#)

Borders and Territories

Potential disputes over borders and territories may have particular relevance for authors in describing their research or in an author or editor correspondence address, and should be respected. Content decisions are an editorial matter and where there is a potential or perceived dispute or complaint, the editorial team will attempt to find a resolution that satisfies parties involved.

MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publication Ethics Statement

Genes is a member of the Committee on Publication Ethics ([COPE](#)). We fully adhere to its [Code of Conduct](#) and to its [Best Practice Guidelines](#).

The editors of this journal enforce a rigorous peer-review process together with strict ethical policies and standards to ensure to add high quality scientific works to the field of scholarly publication. Unfortunately, cases of plagiarism, data falsification, image manipulation,

inappropriate authorship credit, and the like, do arise. The editors of *Genes* take such publishing ethics issues very seriously and are trained to proceed in such cases with a zero tolerance policy.

Authors wishing to publish their papers in *Genes* must abide to the following:

- Any facts that might be perceived as a possible conflict of interest of the author(s) must be disclosed in the paper prior to submission.
- Authors should accurately present their research findings and include an objective discussion of the significance of their findings.
- Data and methods used in the research need to be presented in sufficient detail in the paper, so that other researchers can replicate the work.
- Raw data should preferably be publicly deposited by the authors before submission of their manuscript. Authors need to at least have the raw data readily available for presentation to the referees and the editors of the journal, if requested. Authors need to ensure appropriate measures are taken so that raw data is retained in full for a reasonable time after publication.
- Simultaneous submission of manuscripts to more than one journal is not tolerated.
- The journal accepts exact translations of previously published work. All submissions of translations must conform with our [policies on translations](#).
- If errors and inaccuracies are found by the authors after publication of their paper, they need to be promptly communicated to the editors of this journal so that appropriate actions can be taken. Please refer to our [policy regarding Updating Published Papers](#).
- Your manuscript should not contain any information that has already been published. If you include already published figures or images, please obtain the necessary permission from the copyright holder to publish under the CC-BY license. For further information, see the [Rights and Permissions](#) page.
- Plagiarism, data fabrication and image manipulation are not tolerated.
 - **Plagiarism is not acceptable** in *Genes* submissions.

Plagiarism includes copying text, ideas, images, or data from another source, even from your own publications, without giving any credit to the original source.

Reuse of text that is copied from another source must be between quotes and the original source must be cited. If a study's design or the manuscript's structure or language has been inspired by previous works, these works must be explicitly cited.

All MDPI submissions are checked for plagiarism using the industry standard software iThenticate. If plagiarism is detected during the peer review process, the manuscript may be rejected. If plagiarism is detected after publication, an investigation will take place and action taken in accordance with our policies.

- **Image files must not be manipulated or adjusted in any way** that could lead to misinterpretation of the information provided by the original image.

Irregular manipulation includes: 1) introduction, enhancement, moving, or removing features from the original image; 2) grouping of images that should obviously be presented separately (e.g., from different parts of the same gel, or from different gels); or 3) modifying the contrast, brightness or color balance to obscure, eliminate or enhance some information.

If irregular image manipulation is identified and confirmed during the peer review process, we may reject the manuscript. If irregular image manipulation

is identified and confirmed after publication, we may correct or retract the paper.

Our in-house editors will investigate any allegations of publication misconduct and may contact the authors' institutions or funders if necessary. If evidence of misconduct is found, appropriate action will be taken to correct or retract the publication. Authors are expected to comply with the best ethical publication practices when publishing with MDPI.

Citation Policy

Authors should ensure that where material is taken from other sources (including their own published writing) the source is clearly cited and that where appropriate permission is obtained.

Authors should not engage in excessive self-citation of their own work.

Authors should not copy references from other publications if they have not read the cited work.

Authors should not preferentially cite their own or their friends', peers', or institution's publications.

Authors should not cite advertisements or advertorial material.

In accordance with COPE guidelines, we expect that "original wording taken directly from publications by other researchers should appear in quotation marks with the appropriate citations." This condition also applies to an author's own work. COPE have produced a discussion document on [citation manipulation](#) with recommendations for best practice.

[\[Return to top\]](#)

Reviewer Suggestions

During the submission process, please suggest three potential reviewers with the appropriate expertise to review the manuscript. The editors will not necessarily approach these referees. Please provide detailed contact information (address, homepage, phone, e-mail address). The proposed referees should neither be current collaborators of the co-authors nor have published with any of the co-authors of the manuscript within the last five years. Proposed reviewers should be from different institutions to the authors. You may identify appropriate Editorial Board members of the journal as potential reviewers. You may suggest reviewers from among the authors that you frequently cite in your paper.

[\[Return to top\]](#)

English Corrections

To facilitate proper peer-reviewing of your manuscript, it is essential that it is submitted in grammatically correct English. Advice on some specific language points can be found [here](#).

If you are not a native English speaker, we recommend that you have your manuscript professionally edited before submission or read by a native English-speaking colleague. This can be carried out by MDPI's [English editing service](#). Professional editing will enable reviewers and future readers to more easily read and assess the content of submitted manuscripts. All accepted manuscripts undergo language editing, however **an additional fee will be charged** to authors if very extensive English corrections must be made by the Editorial Office: pricing is according to the service [here](#).

[\[Return to top\]](#)

Preprints and Conference Papers

Genes accepts submissions that have previously been made available as preprints provided that they have not undergone peer review. A preprint is a draft version of a paper made available online before submission to a journal.

MDPI operates *Preprints*, a preprint server to which submitted papers can be uploaded directly after completing journal submission. Note that *Preprints* operates independently of the journal

and posting a preprint does not affect the peer review process. Check the *Preprints [instructions for authors](#)* for further information.

Expanded and high-quality conference papers can be considered as articles if they fulfill the following requirements: (1) the paper should be expanded to the size of a research article; (2) the conference paper should be cited and noted on the first page of the paper; (3) if the authors do not hold the copyright of the published conference paper, authors should seek the appropriate permission from the copyright holder; (4) authors are asked to disclose that it is conference paper in their cover letter and include a statement on what has been changed compared to the original conference paper. *Genes* does not publish pilot studies or studies with inadequate statistical power.

Unpublished conference papers that do not meet the above conditions are recommended to be submitted to the [Proceedings Series journals](#).

[\[Return to top\]](#)

Authorship

MDPI follows the International Committee of Medical Journal Editors ([ICMJE](#)) guidelines which state that, in order to qualify for authorship of a manuscript, the following criteria should be observed:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Those who contributed to the work but do not qualify for authorship should be listed in the acknowledgments. More detailed guidance on authorship is given by the [International Council of Medical Journal Editors \(ICMJE\)](#).

Any change to the author list should be approved by all authors including any who have been removed from the list. The corresponding author should act as a point of contact between the editor and the other authors and should keep co-authors informed and involve them in major decisions about the publication. We reserve the right to request confirmation that all authors meet the authorship conditions.

For more details about authorship please check [MDPI ethics website](#).

Reviewers Recommendation

Authors can recommend potential reviewers. Journal editors will check to make sure there are no conflicts of interest before contacting those reviewers, and will not consider those with competing interests. Reviewers are asked to declare any conflicts of interest. Authors can also enter the names of potential peer reviewers they wish to exclude from consideration in the peer review of their manuscript, during the initial submission progress. The editorial team will respect these requests so long as this does not interfere with the objective and thorough assessment of the submission.

Editorial Independence

Lack of Interference With Editorial Decisions

Editorial independence is of utmost importance and MDPI does not interfere with editorial decisions. All articles published by MDPI are peer reviewed and assessed by our independent editorial boards, and MDPI staff are not involved in decisions to accept manuscripts. When making an editorial decision, we expect the academic editor to make their decision based only upon:

- The suitability of selected reviewers;
- Adequacy of reviewer comments and author response;
- Overall scientific quality of the paper.

In all of our journals, in every aspect of operation, MDPI policies are informed by the mission to make science and research findings open and accessible as widely and rapidly as possible.

Editors and Editorial Staff as Authors

Editorial staff or editors shall not be involved in processing their own academic work. Submissions authored by editorial staff/editors will be assigned to at least two independent outside reviewers. Decisions will be made by other Editorial Board Members who do not have a conflict of interest with the author. Journal staff are not involved in the processing of their own work submitted to any MDPI journals.

Conflict of Interests

According to The International Committee of Medical Journal Editors, “Authors should avoid entering into agreements with study sponsors, both for-profit and non-profit, that interfere with authors’ access to all of the study’s data or that interfere with their ability to analyze and interpret the data and to prepare and publish manuscripts independently when and where they choose.”

All authors must disclose all relationships or interests that could inappropriately influence or bias their work. Examples of potential conflicts of interest include but are not limited to financial interests (such as membership, employment, consultancies, stocks/shares ownership, honoraria, grants or other funding, paid expert testimonies and patent-licensing arrangements) and non-financial interests (such as personal or professional relationships, affiliations, personal beliefs).

Authors can disclose potential conflicts of interest via the online submission system during the submission process. Declarations regarding conflicts of interest can also be collected via the [MDPI disclosure form](#). The corresponding author must include a summary statement in the manuscript in a separate section “Conflicts of Interest” placed just before the reference list. The statement should reflect all the collected potential conflict of interest disclosures in the form.

See below for examples of disclosures:

Conflicts of Interest: Author A has received research grants from Company A. Author B has received a speaker honorarium from Company X and owns stocks in Company Y. Author C has been involved as a consultant and expert witness in Company Z. Author D is the inventor of patent X.

If no conflicts exist, the authors should state:

Conflicts of Interest: The authors declare no conflicts of interest.

[\[Return to top\]](#)

Editorial Procedures and Peer-Review

Initial Checks

All submitted manuscripts received by the Editorial Office will be checked by a professional in-house *Managing Editor* to determine whether they are properly prepared and whether they follow the ethical policies of the journal, including those for human and animal experimentation. Manuscripts that do not fit the journal’s ethics policy or do not meet the standards of the journal will be rejected before peer-review. Manuscripts that are not properly prepared will be returned to the authors for revision and resubmission. After these checks, the *Managing Editor* will consult the journals’ *Editor-in-Chief* or *Associate Editors* to determine whether the manuscript fits the scope of the journal and whether it is scientifically sound. No judgment on the potential impact of the work will be made at this stage. Reject decisions at this stage will be verified by the *Editor-in-Chief*.

Peer-Review

Once a manuscript passes the initial checks, it will be assigned to at least two independent experts for peer-review. A single-blind review is applied, where authors' identities are known to reviewers. Peer review comments are confidential and will only be disclosed with the express agreement of the reviewer.

In the case of regular submissions, in-house assistant editors will invite experts, including recommendations by an academic editor. These experts may also include *Editorial Board Members* and Guest Editors of the journal. Potential reviewers suggested by the authors may also be considered. Reviewers should not have published with any of the co-authors during the past five years and should not currently work or collaborate with any of the institutions of the co-authors of the submitted manuscript.

Optional Open Peer-Review

The journal operates optional open peer-review: *Authors are given the option for all review reports and editorial decisions to be published alongside their manuscript. In addition, reviewers can sign their review, i.e., identify themselves in the published review reports.* Authors can alter their choice for open review at any time before publication, but once the paper has been published changes will only be made at the discretion of the *Publisher* and *Editor-in-Chief*. We encourage authors to take advantage of this opportunity as proof of the rigorous process employed in publishing their research. To guarantee impartial refereeing, the names of referees will be revealed only if the referees agree to do so, and after a paper has been accepted for publication.

Editorial Decision and Revision

All the articles, reviews and communications published in MDPI journals go through the peer-review process and receive at least two reviews. The in-house editor will communicate the decision of the academic editor, which will be one of the following:

- *Accept after Minor Revisions:*
The paper is in principle accepted after revision based on the reviewer's comments. Authors are given five days for minor revisions.
- *Reconsider after Major Revisions:*
The acceptance of the manuscript would depend on the revisions. The author needs to provide a point by point response or provide a rebuttal if some of the reviewer's comments cannot be revised. Usually, only one round of major revisions is allowed. Authors will be asked to resubmit the revised paper within a suitable time frame, and the revised version will be returned to the reviewer for further comments.
- *Reject and Encourage Resubmission:*
If additional experiments are needed to support the conclusions, the manuscript will be rejected and the authors will be encouraged to re-submit the paper once further experiments have been conducted.
- *Reject:*
The article has serious flaws, and/or makes no original significant contribution. No offer of resubmission to the journal is provided.

All reviewer comments should be responded to in a point-by-point fashion. Where the authors disagree with a reviewer, they must provide a clear response.

Author Appeals

Authors may appeal a rejection by sending an e-mail to the Editorial Office of the journal. The appeal must provide a detailed justification, including point-by-point responses to the reviewers' and/or Editor's comments. The *Managing Editor* of the journal will forward the manuscript and related information (including the identities of the referees) to the Editor-in-Chief, Associate Editor, or Editorial Board member. The academic Editor being consulted will be asked to give an advisory recommendation on the manuscript and may recommend acceptance, further peer-

review, or uphold the original rejection decision. A reject decision at this stage is final and cannot be reversed.

In the case of a special issue, the *Managing Editor* of the journal will forward the manuscript and related information (including the identities of the referees) to the *Editor-in-Chief* who will be asked to give an advisory recommendation on the manuscript and may recommend acceptance, further peer-review, or uphold the original rejection decision. A reject decision at this stage will be final and cannot be reversed.

Production and Publication

Once accepted, the manuscript will undergo professional copy-editing, English editing, proofreading by the authors, final corrections, pagination, and, publication on the www.mdpi.com website.

[\[Return to top\]](#)

Promoting Equity, Diversity and Inclusiveness Within MDPI Journals

Our Managing Editors encourage the Editors-in-Chief and Associate Editors to appoint diverse expert Editorial Boards. This is also reflective in our multi-national and inclusive workplace. We are proud to create equal opportunities without regard to gender, ethnicity, sexual orientation, age, religion, or socio-economic status. There is no place for discrimination in our workplace and editors of MDPI journals are to uphold these principles in high regard.

[\[Return to top\]](#)

Resource Identification Initiative

To improve the reproducibility of scientific research, the [Resource Identification Initiative](#) aims to provide unique persistent identifiers for key biological resources, including antibodies, cell lines, model organisms and tools.

We encourage authors to include unique identifiers - RRIDs- provided by the [Resource Identification Portal](#) in the dedicated section of the manuscript.

To help authors quickly find the correct identifiers for their materials, there is a single [website](#) where all resource types can be found and a ‘cite this’ button next to each resource, that contains a proper citation text that should be included in the methods section of the manuscript.

[\[Return to top\]](#)

7.3 Normas da revista *Journal of Animal Breeding and Genetics*

Sections

[1. Submission](#)

[2. Aims and Scope](#)

[3. Manuscript Categories and Requirements](#)

[4. Preparing Your Submission](#)

[5. Editorial Policies and Ethical Considerations](#)

[6. Author Licensing](#)

[7. Publication Process After Acceptance](#)

[8. Post Publication](#)

[9. Editorial Office Contact Details](#)

1. SUBMISSION AND PEER REVIEW PROCESS

New submissions should be made via the Research Exchange submission portal submission.wiley.com/journal/JBG. Should your manuscript proceed to the revision stage, you will be directed to make your revisions via the same submission portal. You may check the status of your submission at anytime by logging on to submission.wiley.com and

clicking the “My Submissions” button. For technical help with the submission system, please review our [FAQs](#) or contact submissionhelp@wiley.com.

Data Protection and Privacy

By submitting a manuscript to, or reviewing for, this publication, your name, email address, institutional affiliation, and other contact details the publication might require, will be used for the regular operations of the publication, including, when necessary, sharing with the publisher (Wiley) and partners for production and publication. The publication and the publisher recognize the importance of protecting the personal information collected from users in the operation of these services, and have practices in place to ensure that steps are taken to maintain the security, integrity, and privacy of the personal data collected and processed. You can learn more at <https://authorservices.wiley.com/statements/data-protection-policy.html>.

Preprint Policy

The *Journal of Animal Breeding and Genetics* will consider for review articles previously available as preprints. Authors may also post the [submitted version](#) of a manuscript to a preprint server at any time. Authors are requested to update any pre-publication versions with a link to the final published article.

2. AIMS AND SCOPE

The journal publishes original articles by international scientists on genomic selection, and any other topic related to breeding programmes, selection, quantitative genetics, genomics, diversity, evolution of domestic animals and analysis of efficiency and consequences of commercial breeding programs. Researchers, teachers, and the animal breeding industry will find the reports of interest.

3. MANUSCRIPT CATEGORIES AND REQUIREMENTS

The *Journal of Animal Breeding and Genetics* publishes:

- **Original Articles** – articles should contain reports of new research findings or conceptual analyses that make a significant contribution to knowledge. Ideally, manuscripts should be around 20 typewritten pages or less – although, longer papers may be considered at the Editor’s discretion.
- **Book Reviews** – books submitted for review are assigned to specialists in the same field. The reviewer does not receive financial remuneration for a review, but keeps the copy of the book sent to him or her for review. The review should include the complete bibliographical data on the book being reviewed: author’s surname and initials of prename(s). Title of the book, edition (if not the first edition), publisher, place of publication, year of publication, length in pages, number of figures and tables, type of binding (paperback, hardback), and retail price.

4. PREPARING YOUR SUBMISSION

The *Journal of Animal Breeding and Genetics* now offers Free Format submission for a simplified and streamlined submission process.

Before you submit, you will need:

- Your manuscript: this should be an editable file including text, figures, and tables, or separate files – whichever you prefer. All required sections should be contained in your manuscript, including abstract, introduction, methods, results, and conclusions. Figures and tables should have legends. Figures should be uploaded in the highest resolution possible. References may be submitted in any style or format, as long as it is consistent throughout the manuscript. Supporting information should be submitted in separate files. If the manuscript, figures or tables are difficult for you to read, they will also be difficult for the editors and reviewers, and the editorial office will send it back to you for revision. Your manuscript may also be sent back to you for revision if the quality of English language is poor.

- An ORCID ID, freely available at <https://orcid.org>. (*Why is this important? Your article, if accepted and published, will be attached to your ORCID profile. Institutions and funders are increasingly requiring authors to have ORCID IDs.*)
- The title page of the manuscript, including:
 - Your co-author details, including affiliation and email address. (*Why is this important? We need to keep all co-authors informed of the outcome of the peer review process.*)
 - Statements relating to our ethics and integrity policies, which may include any of the following:
 - data availability statement
 - funding statement
 - conflict of interest disclosure
 - ethics approval statement
 - permission to reproduce material from other sources
 - clinical trial registration

(Why are these important? We need to uphold rigorous ethical standards for the research we consider for publication)

To submit, login at <https://submission.wiley.com/journal/JBG> and create a new submission. Follow the submission steps as required and submit the manuscript.

Manuscripts can be uploaded either as a single document (containing the main text, tables and figures), or with figures and tables provided as separate files. Should your manuscript reach revision stage, figures and tables must be provided as separate files. The main manuscript file can be submitted in Microsoft Word (.doc or .docx).

5. EDITORIAL POLICIES AND ETHICAL CONSIDERATIONS

Editorial Review and Acceptance

The acceptance criteria for all papers is the quality and originality of the research and its significance to our readership. Except where otherwise stated, manuscripts are single-blind peer reviewed. Papers will only be sent to review if the Editor-in-Chief determines that the paper meets the appropriate quality and relevance requirements. Wiley's policy on confidentiality of the review process is [available here](#).

Data Sharing and Data Accessibility

Journal of Animal Breeding and Genetics recognizes the many benefits of archiving research data. The journal expects you to archive all the data from which your published results are derived in a public repository. The repository that you choose should offer you guaranteed preservation (see the registry of research data repositories at <https://www.re3data.org/>) and should help you make it findable, accessible, interoperable, and re-useable, according to FAIR Data Principles (<https://www.force11.org/group/fairgroup/fairprinciples>). All accepted manuscripts are required to publish a data availability statement to confirm the presence or absence of shared data. If you have shared data, this statement will describe how the data can be accessed, and include a persistent identifier (e.g., a DOI for the data, or an accession number) from the repository where you shared the data. Authors will be required to confirm adherence to the policy. If you cannot share the data described in your manuscript, for example for legal or ethical reasons, or do not intend to share the data then you must provide the appropriate data availability statement. *Journal of Animal Breeding and Genetics* notes that FAIR data sharing allows for access to shared data under restrictions (e.g., to protect confidential or proprietary information) but notes that the FAIR principles encourage you to share data in ways that are as open as possible (but that can be as closed as necessary).

Sample statements are available [here](#). Please note that the samples provided are examples of how the statements can be formatted – these can be modified accordingly depending on your requirements. If published, all statements will be placed in the heading of your manuscript.

As the *Journal of Animal Breeding and Genetics* publishes research linked to commercial breeding programmes, in these cases, authors may not be able to share their underlying data publicly due to license restrictions. Therefore, please find an example data availability statement for such cases here: ‘The data that support the findings of this study are available from (name of third party company). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the corresponding author with the permission of (name of third party company).’

If you are unsure of the suitability of your proposed data availability statement, please reach out to the journal’s Editorial Office for assistance: JABAG.office@wiley.com

Data Citation

Please also cite the data you have shared, like you would cite other sources that your article refers to, in your references section. You should follow the format for your data citations laid out in the Joint Declaration of Data Citation

Principles, <https://www.force11.org/datacitationprinciples>:

[dataset] Authors; Year; Dataset title; Data repository or archive; Version (if any); Persistent identifier (e.g. DOI)

Human Studies and Subjects

For manuscripts reporting medical studies involving human participants, we require a statement identifying the ethics committee that approved the study, and that the study conforms to recognized standards, for example: [Declaration of Helsinki](#); [US Federal Policy for the Protection of Human Subjects](#); or [European Medicines Agency Guidelines for Good Clinical Practice](#).

Images and information from individual participants will only be published where the authors have obtained the individual’s free prior informed consent. Authors do not need to provide a copy of the consent form to the publisher, however in signing the author license to publish authors are required to confirm that consent has been obtained. Wiley has a [standard patient consent form available](#).

Animal Studies

A statement indicating that the protocol and procedures employed were ethically reviewed and approved, and the name of the body giving approval, must be included in the Methods section of the manuscript. We encourage authors to adhere to animal research reporting standards, for example the [ARRIVE reporting guidelines](#) for reporting study design and statistical analysis; experimental procedures; experimental animals and housing and husbandry. Authors should also state whether experiments were performed in accordance with relevant institutional and national guidelines and regulations for the care and use of laboratory animals:

- US authors should cite compliance with the US National Research Council’s [Guide for the Care and Use of Laboratory Animals](#), the US Public Health Service’s [Policy on Humane Care and Use of Laboratory Animals](#), and [Guide for the Care and Use of Laboratory Animals](#).
- UK authors should conform to UK legislation under the [Animals \(Scientific Procedures\) Act 1986 Amendment Regulations \(SI 2012/3039\)](#).
- European authors outside the UK should conform to [Directive 2010/63/EU](#).

Clinical Trial Registration

We require that clinical trials are prospectively registered in a publicly accessible database and clinical trial registration numbers should be included in all papers that report their results. Please include the name of the trial register and your clinical trial registration number at the end of

your abstract. If your trial is not registered, or was registered retrospectively, please explain the reasons for this.

Research Reporting Guidelines

Accurate and complete reporting enables readers to fully appraise research, replicate it, and use it. We encourage authors to adhere to the following research reporting standards.

- [CONSORT](#)
- [SPIRIT](#)
- [PRISMA](#)
- [PRISMA-P](#)
- [STROBE](#)
- [CARE](#)
- [COREQ](#)
- [STARD](#) and [TRIPOD](#)
- [CHEERS](#)
- [the EQUATOR Network](#)
- [Future of Research Communications and e-Scholarship \(FORCE11\)](#)
- [ARRIVE guidelines](#)
- [National Research Council's Institute for Laboratory Animal Research guidelines: the Gold Standard Publication Checklist from Hooijmans and colleagues](#)
- [Minimum Information Guidelines from Diverse Bioscience Communities \(MIBBI\) website; Biosharing website](#)
- [REFLECT statement](#)

Species Names

Upon its first use in the title, abstract and text, the common name of a species should be followed by the scientific name (genus, species and authority) in parentheses. For well-known species, however, scientific names may be omitted from article titles. If no common name exists in English, the scientific name should be used only.

Genetic Nomenclature

Sequence variants should be described in the text and tables using both DNA and protein designations whenever appropriate. Sequence variant nomenclature must follow the current HGVS guidelines; see <http://varnomen.hgvs.org/>, where examples of acceptable nomenclature are provided.

Nucleotide Sequence Data

Nucleotide sequence data can be submitted in electronic form to any of the three major collaborative databases: DDBJ, EMBL or GenBank. It is only necessary to submit to one database as data are exchanged between DDBJ, EMBL and GenBank on a daily basis. The suggested wording for referring to accession-number information is: 'These sequence data have been submitted to the DDBJ/EMBL/GenBank databases under accession number U12345'. Addresses are as follows:

DNA Data Bank of Japan (DDBJ) <http://www.ddbj.nig.ac.jp>

EMBL Nucleotide Sequence Submissions <http://www.ebi.ac.uk>

GenBank <http://www.ncbi.nlm.nih.gov>

Conflict of Interest

The *Journal of Animal Breeding and Genetics* requires that all authors disclose any potential sources of conflict of interest. Any interest or relationship, financial or otherwise that might be perceived as influencing an author's objectivity is considered a potential source of conflict of interest. These must be disclosed when directly relevant or directly related to the work that the authors describe in their manuscript. Potential sources of conflict of interest include, but are not limited to, patent or stock ownership, membership of a company board of directors, membership of an advisory board or committee for a company, and consultancy for or receipt of speaker's

fees from a company. The existence of a conflict of interest does not preclude publication. If the authors have no conflict of interest to declare, they must also state this at submission. It is the responsibility of the corresponding author to review this policy with all authors and collectively to disclose with the submission ALL pertinent commercial and other relationships. The Conflict of Interest statement should be included within the main text file of your submission.

Funding

Authors should list all funding sources in the Acknowledgments section. Authors are responsible for the accuracy of their funder designation. If in doubt, please check the Open Funder Registry for the correct nomenclature: <http://www.crossref.org/fundingdata/registry.html>

Authorship

The list of authors should accurately illustrate who contributed to the work and how. All those listed as authors should qualify for authorship according to the following criteria:

- 1) Have made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data;
- 2) Been involved in drafting the manuscript or revising it critically for important intellectual content;
- 3) Given final approval of the version to be published. Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content; and
- 4) Agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in an Acknowledgments section (for example, to recognize contributions from people who provided technical help, collation of data, writing assistance, acquisition of funding, or a department chairperson who provided general support). Prior to submitting the article all authors should agree on the order in which their names will be listed in the manuscript.

Additional authorship options

Joint first or senior authorship: In the case of joint first authorship a footnote should be added to the author listing, e.g. ‘X and Y should be considered joint first author’ or ‘X and Y should be considered joint senior author.’

ORCID

As part of our commitment to supporting authors at every step of the publishing process, the *Journal of Animal Breeding and Genetics* requires the submitting author (only) to provide an ORCID iD when submitting a manuscript. This takes around 2 minutes to complete. Find more [information](#).

Publication Ethics

Journal of Animal Breeding and Genetics is a member of the [Committee on Publication Ethics \(COPE\)](#). Note this journal uses iThenticate’s CrossCheck software to detect instances of overlapping and similar text in submitted manuscripts. Read our Top 10 Publishing Ethics Tips for Authors [here](#). Wiley’s Publication Ethics Guidelines can be found at <https://authorservices.wiley.com/ethics-guidelines/index.html>

6. AUTHOR LICENSING

If your paper is accepted, the author identified as the formal corresponding author will receive an email prompting them to log in to Author Services, where via the Wiley Author Licensing Service (WALS) they will be required to complete a copyright license agreement on behalf of all authors of the paper.

Authors may choose to publish under the terms of the journal's standard copyright agreement, or [Open Access](#) under the terms of a Creative Commons License.

General information regarding licensing and copyright is available [here](#). To review the Creative Commons License options offered under Open Access, please [click here](#). (Note that certain funders mandate that a particular type of CC license has to be used; to check this please click [here](#).)

Self-Archiving definitions and policies. Note that the journal's standard copyright agreement allows for self-archiving of different versions of the article under specific conditions. Please click here for more detailed information about self-archiving definitions and policies.

Open Access fees: If you choose to publish using Open Access you will be charged a fee. A list of Article Publication Charges for Wiley journals is available [here](#).

Funder Open Access: Please click [here](#) for more information on Wiley's compliance with specific Funder Open Access Policies.

7. PUBLICATION PROCESS AFTER ACCEPTANCE

Accepted article received in production

When your accepted article is received by Wiley's production production team, you (corresponding authors) will receive an email asking you to login or register with [Author Services](#). You will be asked to sign a publication licence at this point.

Proofs

Authors will receive an e-mail notification with a link and instructions for accessing HTML page proofs online. Page proofs should be carefully proofread for any copyediting or typesetting errors. Online guidelines are provided within the system. No special software is required, all common browsers are supported. Authors should also make sure that any renumbered tables, figures, or references match text citations and that figure legends correspond with text citations and actual figures. Proofs must be returned within 48 hours of receipt of the email. Return of proofs via e-mail is possible in the event that the online system cannot be used or accessed.

Publication Charges

Color figures may be published online free of charge; however, the journal charges for publishing figures in colour in print. If the author supplies colour figures at Early View publication, they will be invited to complete a colour charge agreement in RightsLink for Author Services. The author will have the option of paying immediately with a credit or debit card, or they can request an invoice. If the author chooses not to purchase color printing, the figures will be converted to black and white for the print issue of the journal.

Early View

The journal offers rapid publication via Wiley's Early View service. [Early View](#) (Online Version of Record) articles are published on Wiley Online Library before inclusion in an issue. Once your article is published on Early View no further changes to your article are possible. Your Early View article is fully citable and carries an online publication date and DOI for citations.

8. POST PUBLICATION

Access and sharing

When your article is published online:

- You receive an email alert (if requested).
- You can share your published article through social media.
- As the author, you retain free access (after accepting the Terms & Conditions of use, you can view your article).
- The corresponding author and co-authors can nominate up to ten colleagues to receive a publication alert and free online access to your article.

Article Promotion Support

[Wiley Editing Services](#) offers professional video, design, and writing services to create shareable video abstracts, infographics, conference posters, lay summaries, and research news stories for your research – so you can help your research get the attention it deserves.

9. EDITORIAL OFFICE CONTACT DETAILS

JABAG.office@wiley.com

Author Guidelines updated 15th February 2021