

GUILHERME FERREIRA SIMIQUELI

**ENTROPY, MUTUAL INFORMATION, AND POPULATION STRUCTURE IN
GENOME-WIDE SELECTION**

Thesis submitted to the Universidade Federal de Viçosa, in partial fulfilment of the requirements of Genetics and Breeding Graduate Program for degree of *Doctor Scientiae*.

Advisor: Marcos Deon Vilela de Resende

**VIÇOSA – MINAS GERAIS
2020**

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa**

T

S589e
2020
Simiqueli, Guilherme Ferreira, 1988-
Entropy, mutual information, and population structure
in genome-wide selection / Guilherme Ferreira Simiqueli. –
Viçosa, MG, 2020.
122 f. : il. (algumas color.) ; 29 cm.

Inclui anexos.

Orientador: Marcos Deon Vilela de Resende.

Tese (doutorado) - Universidade Federal de Viçosa.

Inclui bibliografia.

1. Genômica. 2. Melhoramento genético. 3. Estrutura populacional. 4. Entropia. 5. Predição. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22 ed. 572.86

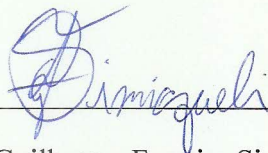
GUILHERME FERREIRA SIMIQUELI

**ENTROPY, MUTUAL INFORMATION, AND POPULATION STRUCTURE IN
GENOME-WIDE SELECTION**

Thesis submitted to the Universidade Federal de Viçosa, in partial fulfilment of the requirements of Genetics and Breeding Graduate Program for degree of *Doctor Scientiae*.

APPROVED: July 23, 2020.

Assent:



Guilherme Ferreira Simiqueli
Author



Marcos Deon Vilela de Resende
Advisor

*To my parents, sisters, and Kynynyn
for all love shared.*

ACKNOWLEDGEMENTS

I would like to acknowledge my parents, José and Célia, and my sisters, Ronara, Raquel, and Priscila for their love, confidence, patience, and unforgettable moments.

I would like to thank my advisor Marcos Deon for his outstanding perception of science, life, and for all powerful happy moments. I am also thankful for all professors who contributed to my professional skills and taught me to see the world through different ways. The diverse knowledge shows we always need to become the best of ourselves. I am thankful to the Postgraduate Program in Genetics and Breeding for all the support, mainly, the staff Marco Túlio and Odilon for their kindness, patience, and efficiency.

I would like to thank my relatives, mainly, my aunt Maria and my uncle Carlão for the support and friendship during my studies in Viçosa. I would like to thank all of my friends from Viçosa who were with me during my doctor degree, mainly, Caio Eleto, Eliane Freitas, Lori, Eliana, Ricardo, Helô, Rafael, Brígida, Álvaro, Dani, Luciana, Marcone, Natane, Rodrigo, Andrei Caique, Lívia, Gleidson, Carla, Caio, Tales, Cleiton, Danilo, João Victor, Michele, Rafael César, Vanderson, Douglas, Well, Cláudio, Júnior, Nathália Granato, Nero, Neuzinha, Alexandre, Chiquinho, Marquione, Dilson, Paulo, and all friends of biometry laboratory and so many others. In addition, I would like to thank my friends from Juiz de Fora, mainly, Thales, Cristina, Dina, Estela, Mariana, Lucas, Ricardo, Renata, Glauber, and Tássio.

I would like to thank the professor Leonardo Bhering and Camila Azevedo for my teaching internship journey. I would also like to thank my sister, Raquel, for English language editing and review of this study.

I am very grateful to my friend, Well Clarindo, who shared his routine, funny moments, and happiness of living in community. Further, I would like to thank all the pets Cindy, Aretha, Toddy and Kynynyn for patience, affection, and love.

I am very grateful to my friends Thiago Pires and Rafael Tassinari, who encourage me to study genetic and breeding. Last but not least, I would like to take this opportunity to thank all living being made me realize that there is always a chance to reborn, when all hope has gone. This is a gift and I am grateful to be here.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Arousing smiles, that's the real power.

ABSTRACT

SIMIQUÉLI, Guilherme Ferreira, D.Sc., Universidade Federal de Viçosa, July, 2020.
Entropy, mutual information, and population structure in genome-wide selection.
Advisor: Marcos Deon Vilela de Resende.

Different populations can compose the training set aiming for a better predictive ability of genomic prediction models. However, this practice has not always resulted in higher predictive ability and some studies have proposed to account population structure effect for a better prediction. Different strategies like principal components covariates, uni and multi-population models, alternative genomic relationships matrices, admixed proportions covariates, or a mix of them have been applied to genomic prediction. Thus, the first chapter aims to evaluate some combinations of these strategies to help the decision making about considering or not considering population structure on genomic prediction. Simulated polygenic traits with 0.1 and 0.5 heritability and real data were used to evaluate the strategies. Bias was lower, when multi-population model was used for low-heritability simulated trait. The accuracy of high-heritability trait was lower for strategies that used alternative genomic matrices that accounted for differences in allele frequency, only in admixed populations. Further, for real data, two commonly used genomic relationship matrices showed lower values of predictive ability for all traits, which are likely controlled by few quantitative trait loci. Therefore, accounting for population structure depends on trait heritability, trait architecture, and admixture level of population for obtaining lower bias without reduction of accuracy, and, consequently, success of genomic prediction. The second chapter address the fact that random k-fold cross-validation in genome wide selection can provide high estimates of predictive ability, due to the high degree of kinship between the training and validation sets. However, many breeding tree populations are less genetically related to the training sets and have different levels of phenotypic diversity. Therefore, this chapter proposed novel methods of splitting cross-validation sets, accounting genetic similarity and phenotypic diversity estimated via mutual information and entropy, respectively. These methods also verified how distribution of phenotypic and genotypic information affects genome wide selection of trees. The methods trustworthily fitted models, according to the entropy of tree breeding populations and their genetic relatedness to the training sets. Validation sets with more phenotypic diversity showed higher predictive ability and lower bias. Therefore, the phenotypic diversity should be added in tree breeding populations for

higher genetic gain and better estimation of genomic breeding values and a consistent long-term tree breeding success.

Keywords: Population structure. Accuracy. Bias. Mutual information. Entropy. K-fold cross-validation.

RESUMO

SIMIQUELI, Guilherme Ferreira, D.Sc., Universidade Federal de Viçosa, julho de 2020. **Entropia, informação mútua e estrutura de populações na seleção genômica ampla.** Orientador: Marcos Deon Vilela de Resende.

Na predição genômica, diferentes populações podem compor o conjunto de treinamento para melhorar a capacidade preditiva. Entretanto, esta prática não tem resultado em maiores capacidades preditivas e alguns estudos propuseram acomodar o efeito de estrutura populacional para melhor predição. Diferentes estratégias como componentes principais, modelos uni e multipopulacionais, matrizes alternativas de parentesco genômico, proporção de indivíduos misturados ou uma mistura destas estratégias tem sido empregada na predição genômica. Portanto, o objetivo deste primeiro capítulo foi avaliar algumas combinações destas estratégias para ajudar no processo de decisão sobre considerar ou não o efeito de estrutura populacional na predição genômica. Duas características poligênicas foram simuladas com herdabilidade de 0,1 e 0,5 e dados reais foram utilizados na avaliação. O viés de predição foi menor quando modelos multipopulacionais foram empregados para característica simulada de baixa herdabilidade. A acurácia da característica com alta herdabilidade (0,5) em populações misturadas foi baixa para estratégias que utilizaram matrizes de parentesco genômico que consideravam diferenças na frequência alélica. Além disso, nos dados reais, duas matrizes alternativas de parentesco genômico apresentaram baixa capacidade preditiva para as características avaliadas, as quais são provavelmente governadas por poucos loci. Portanto, a acomodação de estrutura populacional depende da arquitetura genética da característica, da herdabilidade e do nível de mistura da população para obtenção de menor viés sem reduzir a acurácia e, conseqüentemente, sucesso da predição genômica. O segundo capítulo aborda a validação cruzada na seleção genômica ampla. Esta validação quando feita aleatoriamente ocasiona em altos valores das estimativas de capacidade preditiva, provavelmente, devido ao alto grau de parentesco entre os conjuntos de treinamento e validação. No entanto, muitas populações de melhoramento florestal são fracamente relacionadas geneticamente com os conjuntos de treinamento e possuem diferentes níveis de diversidade fenotípica. Portanto, este capítulo propôs novos métodos de separação dos conjuntos de validação cruzada, considerando a similaridade genética e a diversidade fenotípica, obtidas por meio da informação mútua e entropia, respectivamente. Esses novos métodos também verificaram como a distribuição das

informações fenotípicas e genotípicas afeta a seleção genômica ampla de espécies florestais. Os novos métodos ajustaram modelos mais confiáveis e que estão de acordo com a entropia das populações de melhoramento e sua relação genética com os conjuntos de treinamento. Os conjuntos de validação com maior diversidade fenotípica apresentaram maior capacidade preditiva e menor viés. Portanto, a diversidade fenotípica deve ser adicionada nas populações de melhoramento para maior ganho genético e melhor estimativa dos valores genéticos genômicos.

Palavras-chave: Estrutura populacional. Acurácia. Viés. Informação mútua. Entropia. Validação cruzada.

SUMMARY

<i>1. General introduction</i>	12
<i>2. References</i>	15
<i>3. Chapter 1: Strategies for accounting population structure on genomic prediction: implications for breeding programs</i>	19
<i>Abstract</i>	19
3.1 Introduction	19
3.2 Material and methods	22
3.2.1 Simulation of data	22
3.2.2 Real data	24
3.2.3 Strategies for accounting population structure effect	24
3.2.4 Evaluation of strategies	27
3.3 Results	28
3.3.1 Analyses of population structure	28
3.3.2 Predictive ability (<i>ryy</i>) and accuracy (<i>rgg</i>)	30
3.3.3 Prediction bias and bias of the true breeding value (<i>BiasTBV</i>)	31
3.3.4 Genomic heritability	32
3.4 Discussion	33
3.4.1 Predictive ability (<i>ryy</i>) and accuracy (<i>rgg</i>)	33
3.4.2 Bias and bias of true breeding values	34
3.4.3 Genomic heritability (<i>hg2</i>)	35
3.4.4 Strategy recommendations	36
3.5 Conclusions	36
3.6 Acknowledgments	37
3.7 References	37
3.8 Supplementary material	43
<i>4. Chapter 2: Entropy and mutual information in genome-wide selection: the splitting of k-fold cross-validation sets and implications for tree breeding</i>	62
<i>Abstract</i>	62
4.1 Introduction	62

4.2 Material and methods	65
4.2.1 Simulated data	65
4.2.2 Real data	67
4.2.3 K-fold cross-validation methods	68
4.2.4 Evaluation of k-fold cross-validation methods	72
4.3 Results	73
4.3.1 Genotypic and phenotypic similarity	73
4.3.2 Predictive ability and bias	75
4.3.3 Correlations between model parameters with phenotypic and genotypic information	77
4.4 Discussion	79
4.4.1 The effect of diversity on <i>ryy</i> values	79
4.4.2 The effect of diversity on β_1 and MSEP	80
4.4.3 The choice of k-fold cross-validation method for 20-fold cross-validation scheme	81
4.5 Conclusions	82
4.6 Data archiving statement	82
4.7 Acknowledgments	82
4.8 References	82
4.9 Appendix A: Derivation of avoidance of negative entropy	88
4.10 Attachments	88
4.10.1 Online resource 1	88
4.10.2 Online resource 2	94
5. General conclusions	122

1. GENERAL INTRODUCTION

Genome wide selection (GWS) has been applied to many breeding species and has shown great potential of selection gain in a shorter time for some traits. However, some breeding programs have many small populations and a genomic model for each population can result in lower predictive ability due to the low number of phenotypes (DAETWYLER et al., 2013; RESENDE et al., 2012). One solution to this is to combine all populations in the training set, but this practice has showed subtle increased accuracy and also varied between traits and populations (MAKGAHLELA et al., 2013; ZHOU et al., 2014). In addition, marker effects can also vary between populations (DE LOS CAMPOS; SORENSEN, 2014; LEHERMEIER; SCHON; DE LOS CAMPOS, 2015; WIENTJES et al., 2017) and, thus, accounting for population structure effect could obtain higher accuracy and lower bias of estimated breeding values (GEBVs).

Many strategies have been proposed to account population structure effects. These strategies include correction of population structure effect by principal component analysis (PCA) (AZEVEDO et al., 2017; DAETWYLER et al., 2012; LYRA et al., 2018), the use of different genomic matrices, inclusion of individual admixed proportions as fixed effects, and multi-population models (KAROUI et al., 2012; LEHERMEIER; SCHON; DE LOS CAMPOS, 2015; RIO et al., 2019; TECHNOW et al., 2012). Some studies have applied these strategies individually or by mixing them (LYRA et al., 2018; RIO et al., 2019). However, the lack of comparison between these strategies hinders the decision making about considering or not considering population structure in genomic prediction. Therefore, the first chapter of this study evaluated some strategies by combining uni or multi-population model, admixed proportions, and different genomic relationship matrices, which account or not account for population structure with simulated and real data.

The second chapter addressed the use of k-fold cross-validation in genome wide selection of tree species. The k-fold cross-validation has shown higher predictive values when training and validation sets are randomly separated due to high degree of relationship shared (CLARK et al., 2012; PSZCZOLA et al., 2012), shared relatives (PÉREZ-CABAL et al., 2012), and lower fixation index values between the training and validation sets (SCUTARI; MACKAY; BALDING, 2016). Hence, the use of models based on random k-fold cross-validation can overestimate predictive ability, when it is applied to tree breeding populations that are less related to the training sets used.

Many strategies for composing training sets have been proposed to produce more accurate genomic prediction. These strategies split the training and validation sets according to family (HULSMAN HANNA et al., 2015; RESENDE et al., 2017), population structure (GUO et al., 2014), generation (SAATCHI; WARD; GARRICK, 2013; SILVA et al., 2016), maximum kinship coefficient (HABIER et al., 2010), Wright kinship coefficients (BODDHIREDDY et al., 2014; CLARK et al., 2012; SAATCHI; WARD; GARRICK, 2013), identity by state clustering methods (BODDHIREDDY et al., 2014) and unrelated individuals (PÉREZ-CABAL et al., 2012; SILVA et al., 2016). Another alternative is to add individuals from different populations in the training set and, consequently, reduce the degree of kinship between the training and validation sets and enhance the model's applicability across distinct populations (DE ROOS; HAYES; GODDARD, 2009; TECHNOV et al., 2012). However, this alternative depends on quantity and quality of the newly added information (HOFFSTETTER et al., 2016; RINCENT; CHARCOSSET; MOREAU, 2017). This information comes from genetic (PSZCZOLA et al., 2012; RINCENT et al., 2012) and possibly phenotypic differences between individuals (ISIDRO et al., 2015). Consequently, the distribution of genotypic and phenotypic information among the training and validation sets can affect the fit of the model, as well as its applicability and perpetuation.

Thus, this second chapter proposed novel strategies to acquire k-fold cross-validation sets based on information theory. This theory can be used to estimate genotypic and phenotypic information with parameters such as mutual information and entropy, respectively (SHANNON, 1948). Entropy (H) measures information by the amount of uncertainty regarding the value of a random variable; thus, if the random variable is a constant, there is no entropy or information. Mutual information (I) is a symmetric, non-negative, and nonlinear measurement of the amount of information shared between random variables (COVER; THOMAS, 2012; SHANNON, 1948).

Information theory has proven potential in many areas of science such as physics by showing how information is a part of the second law of thermodynamics through its influence on non-equilibrium free energy (PARRONDO; HOROWITZ; SAGAWA, 2015). In statistical inference, the Kullback-Leibler distance (an information theory parameter) of the Gaussian model can be minimized and provides least square estimators (BASU; SHIOYA; PARK, 2011; PARDO, 2006; RESENDE, 2015). In genetics, genome-wide association studies have found genomic regions associated with animal production by applying methods based on I and H (BOROWSKA et al., 2017;

GRACZYK et al., 2017). In GWS, some methods based on I and H have been applied to obtain a representative set of single nucleotide polymorphisms (SNPs) and have found better predictive abilities instead of using all markers (HAWS et al., 2015; HE et al., 2016; LONG et al., 2007).

Therefore, in this study, mutual information was used as a measurement of genotypic information or similarity between folds and the entropy was used as an estimate of phenotypic information or phenotypic diversity of folds. These novel strategies should produce trustworthy fitted models, which can be applied to tree breeding, and other species populations, according to their phenotypic diversity and genotypic information to the training set. The influence of genotypic and phenotypic information distribution on parameters commonly used in genome-wide selection, such as predictive ability, bias, and genomic heritability was also verified.

2. REFERENCES

AZEVEDO, C. F. *et al.* Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*, v. 17, n. 4, p. 350–358, 2017.

BASU, A.; SHIOYA, H.; PARK, C. *Statistical inference: the minimum distance approach*. [S.l.]: Chapman and Hall/CRC, 2011. v. 39.

BODDHIREDDY, P. *et al.* Genomic predictions in Angus cattle: comparisons of sample size, response variables, and clustering methods for cross-validation. *Journal of Animal Science*, v. 92, n. 2, p. 485–497, 2014.

BOROWSKA, A. *et al.* Detection of pig genome regions determining production traits using an information theory approach. *Livestock Science*, v. 205, n. September, p. 31–35, 2017.

CLARK, S. A. *et al.* The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics, selection, evolution : GSE*, v. 44, p. 4, 2012.

COVER, T. M; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012.

DAETWYLER, H. D. *et al.* Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of Animal Science*, v. 90, n. 10, p. 3375–3384, 2012.

DAETWYLER, H. D. *et al.* Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, v. 193, n. 2, p. 347–365, 2013.

DE LOS CAMPOS, G.; SORENSEN, D. On the genomic analysis of data from structured populations. *Journal of Animal Breeding and Genetics*, v. 131, n. 3, p. 163–164, 2014.

DE ROOS, A. P W; HAYES, B. J.; GODDARD, M. E. Reliability of genomic predictions across multiple populations. *Genetics*, v. 183, n. 4, p. 1545–1553, 2009.

GRACZYK, M. *et al.* Detection of the important chromosomal regions determining production traits in meat-type chicken using entropy analysis. *British Poultry Science*, v. 58, n. 4, p. 358–365, 2017.

GUO, Z. *et al.* The impact of population structure on genomic prediction in

stratified populations. *Theoretical and applied genetics*, v. 127, n. 3, p. 749–762, 2014.

HABIER, D. *et al.* The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics, selection, evolution : GSE*, v. 42, n. 1, p. 5, 2010.

HAWS, D. C. *et al.* Variable-selection emerges on top in empirical comparison of whole-genome complex-trait prediction methods. *PLoS ONE*, v. 10, n. 10, p. 1–22, 2015.

HE, D. *et al.* MINT: mutual information based transductive feature selection for genetic trait prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 13, n. 3, p. 578–583, 2016.

HOFFSTETTER, A. *et al.* Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. *G3: Genes|Genomes|Genetics*, v. 6, n. 9, p. 2919–2928, 2016.

HULSMAN HANNA, L. L. *et al.* Cross-validation of genetic and genomic predictions of temperament in Nellore-Angus crossbreds. *Livestock Science*, v. 182, p. 28–33, 2015.

ISIDRO, J. *et al.* Training set optimization under population structure in genomic selection. *Theoretical and applied genetics*, v. 128, n. 1, p. 145–158, 2015.

KAROUI, S. *et al.* Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genetics Selection Evolution*, v. 44, n. 1, p. 1–10, 2012.

LEHERMEIER, C.; SCHON, C.C.; DE LOS CAMPOS, G. Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics*, v. 201, n. 1, p. 323–337, 2015.

LONG, N. *et al.* Machine learning procedure for selecting single nucleotide polymorphisms in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics*, v. 124, p. 377–389, 2007.

LYRA, D. H. *et al.* Controlling population structure in the genomic prediction of tropical maize hybrids. *Molecular Breeding*, v. 38, n. 10, 2018.

MAKGAHLELA, M. L. *et al.* Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *Journal of Animal Breeding and Genetics*, v. 130, n. 1, p. 10–19, 2013.

PARDO, L. *Statistical inference based on divergence measures*. [S.l.]: Chapman

and Hall/CRC, 2006. v. 185.

PARRONDO, J. M. R.; HOROWITZ, J. M.; SAGAWA, T. Thermodynamics of information. *Nature Physics*, v. 11, n. 2, p. 131–139, 2015.

PÉREZ-CABAL, M. A. *et al.* Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Frontiers in Genetics*, v. 3, n. February, p. 1–7, 2012.

PSZCZOLA, M. *et al.* Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, v. 95, n. 1, p. 389–400, 2012.

RESENDE, R. T. *et al.* Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. *Heredity*, v. 119, n. 4, p. 245–255, 2017.

RESENDE, M. D. V. *Genética quantitativa e de populações*. Suprema, Visconde do Rio Branco, p. 452, 2015.

RESENDE, M. D. V. *et al.* *Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial*. [S.l: s.n.], 2012.

RINCENT, R.; CHARCOSSET, A.; MOREAU, L. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics*, v. 130, n. 11, p. 2231–2247, 2017.

RINCENT, R. *et al.* Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, v. 192, n. 2, p. 715–728, 2012.

RIO, S. *et al.* Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theoretical and Applied Genetics*, v. 132, n. 1, p. 81–96, 2019.

SAATCHI, M.; WARD, J.; GARRICK, D. J. Accuracies of direct genomic breeding values in Hereford beef cattle using national or international training populations. *Journal of Animal Science*, v. 91, n. 4, p. 1538–1551, 2013.

SCUTARI, M.; MACKAY, I.; BALDING, D. Using genetic distance to infer the

accuracy of genomic prediction. *PLoS genetics*, v. 12, n. 9, p. 1-19, 2016.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. July, October, p. 379–423, 1948.

SILVA, R. M. O. *et al.* Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. *Journal of Animal Science*, v. 94, n. 9, p. 3613–3623, 2016.

TECHNOW, F. *et al.* Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theoretical and Applied Genetics*, v. 125, n. 6, p. 1181–1194, 2012.

WIJNTJES, Y. C. J. *et al.* Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics*, v. 207, n. 2, p. 503–515, 2017.

ZHOU, L. *et al.* Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *Journal of Animal Breeding and Genetics*, v. 131, n. 4, p. 249–257, 2014.

3. CHAPTER 1: Strategies for accounting population structure on genomic prediction: implications for breeding programs

Abstract

Genomic prediction has shown potential for some breeding programs, when training sets are composed by different populations. However, these diverse training sets have not always resulted in higher predictive ability. Thus, studies have proposed strategies accounting for population structure aiming better accuracy, like uni and multi-population models, alternative genomic relationships matrices, principal components covariates, admixed proportions covariates, or a mix of them. Nevertheless, this mix of strategies difficult the decision making about considering or not population structure. Therefore, this study aimed to evaluated some combinations of these strategies on genomic prediction. Data were simulated for two traits (0.1 and 0.5 heritability) and two more real traits were analyzed to ten different strategies. These strategies considered uni and multi-population model, admixed proportions as fixed effects and different genomic matrices. The predictive ability, accuracy, prediction bias, bias of the true breeding value, and genomic heritability were estimated. The predictive ability showed no differences between strategies. The accuracy of high-heritability trait was lower in admixed populations considering the use of strategies with genomic matrices that accounted for population structure. Bias was lower when multi-population model was used for low-heritability simulated traits. For real data, two commonly used genomic relationship matrices showed lower values of predictive ability for both traits, which are likely controlled by few quantitative trait loci. Therefore, the effectiveness of strategies that account for population structure depends on trait heritability, trait architecture, and admixture level of population for obtaining lower bias without reduction of accuracy, and, consequently, success of genomic prediction.

Keywords: Accuracy, bias, population structure, multi-population model, genomic relationship matrix

3.1 Introduction

Genome wide selection (GWS) depends on linkage disequilibrium (LD) between makers and quantitative trait loci (QTLs). Hence, a great number of makers are needed to assure that there is LD between part of them with some QTLs, which is an assumption of GWS (Meuwissen *et al.*, 2001). The increasing number of markers and reduction of its cost has made possible the use of GWS in animal (Raymond *et al.*, 2018; Schultz and

Weigel, 2019), crop (Lyra *et al.*, 2018; Rio *et al.*, 2019), and forestry (Resende *et al.*, 2008, 2012, 2017; Grattapaglia and Resende, 2011) breeding with success for many traits, contributing for a shorter breeding cycle.

The use of higher marker density has shown to increase the predictive ability of GWS (de Roos *et al.*, 2009; Toosi *et al.*, 2010). However, an increase in marker density is not enough to obtain higher accuracy if the number of evaluated individuals is low (<1000) due to the laborious measurement of some traits (Resende, Silva, *et al.*, 2012; Daetwyler *et al.*, 2013). One possible solution is to include individuals from different populations, considering the quantity and quality of added information. Although, this practice has shown increase in predictive ability (de Roos *et al.*, 2009; Schulz-Streeck *et al.*, 2012; Technow *et al.*, 2012), this increase has been subtle and varied between populations and traits (Makgahlela *et al.*, 2013; Zhou *et al.*, 2014). However, the prediction across different populations has been effective, if all populations are included in the training set and genetic correlation between populations is high (de Roos *et al.*, 2009; Karoui *et al.*, 2012; Schulz-Streeck *et al.*, 2012). In addition, the predictive ability is higher for populations with small size and high genetic correlation with the larger population size in the training set (Karoui *et al.*, 2012).

The lower predictive ability across different populations is attributed to low genetic correlations between populations (Wientjes *et al.*, 2017), different LD phases (de Roos *et al.*, 2009; Chen *et al.*, 2013), and different patterns of LD between populations (Sawyer *et al.*, 2005), resulting in different marker effects and genomic breeding values (GEBVs) (de los Campos and Sorensen, 2014; Lehermeier *et al.*, 2015; Wientjes *et al.*, 2017). Thus, aiming a higher predictive ability and lower bias, different studies have proposed methods to account for population structure in GWS like principal component analysis (PCA) (Daetwyler *et al.*, 2012; Azevedo *et al.*, 2017; Lyra *et al.*, 2018), the use of different genomic matrices (Zheng *et al.*, 2013; Wientjes *et al.*, 2017), inclusion of individual admixed proportions as fixed effects (Makgahlela *et al.*, 2013; Thomassen *et al.*, 2013; Rio *et al.*, 2019), and multi-population models (Karoui *et al.*, 2012; Lehermeier *et al.*, 2015; Wientjes *et al.*, 2017).

Multi-population models estimate different marker effects per population and has shown no or small increase of predictive ability with varied values between traits and population (Karoui *et al.*, 2012; Schulz-Streeck *et al.*, 2012; Technow *et al.*, 2012; Lehermeier *et al.*, 2015; Rio *et al.*, 2019). However, these models seemed to be beneficial under low marker density and low LD (Technow *et al.*, 2012) and can be more effective

in larger sample sizes (Schulz-Streeck *et al.*, 2012; Lehermeier *et al.*, 2015). Further, multi-population models can estimate genetic correlation between populations and, thus, understand trait-specific differences in genetic architecture between populations (Brown *et al.*, 2016). However, not all traits are influenced by population structure and the use of principal components or admixed proportions to account population structure may be sufficient for a better prediction.

Principal component eigenvectors are extracted from genomic relationship matrix and used as fixed effects covariates to account for population structure effect and to avoid spurious associations between marker and phenotypes in genome wide association studies (GWAS) (Price *et al.*, 2006; Yu *et al.*, 2006; Conomos *et al.*, 2015). In GWS, the use of principal component eigenvectors as fixed effects covariates showed more accurate estimated breeding values (Azevedo *et al.*, 2017), reduction (Daetwyler *et al.*, 2012) or no effect in predictive ability (Lyra *et al.*, 2018). Moreover, this practice was proven to count population structure twice (Janss *et al.*, 2012) and using a decomposed genomic relationship without some principal component eigenvectors can deeply affect the estimates of variance components (Gianola *et al.*, 2016) and reduces predictive ability (Guo *et al.*, 2014).

Admixed proportions or ancestry proportions are estimated by considering that genome of individuals is composed by alleles from different populations and there are several methods currently available for estimating them (Pritchard *et al.*, 2000; Falush *et al.*, 2003; Alexander *et al.*, 2009; Raj *et al.*, 2014; Conomos *et al.*, 2015). These admixed proportions included as fixed effects has been effective for controlling spurious associations in GWAS models (Yu *et al.*, 2006; Conomos *et al.*, 2015). Thus, they could be used to account for population structure in GWS and, perhaps, included in multi-population models, which do not account for admixed individuals (Lehermeier *et al.*, 2015). However, the use of admixture proportions has shown no increase in predictive ability in GWS (Makgahlela *et al.*, 2013; Thomassen *et al.*, 2013; Lyra *et al.*, 2018; Rio *et al.*, 2019).

In GWAS, a matrix that considers admixed individual proportions in its construction has been suggested to estimate kinship coefficients (Thornton *et al.*, 2012; Conomos *et al.*, 2016). This matrix estimates kinship without influence of population structure and has been applied in GWS (Rio *et al.*, 2019). Moreover, many different genomic relationship matrices have been proposed to account population structure (Erbe *et al.*, 2012; Chen *et al.*, 2013; Wientjes *et al.*, 2017). The use of different genomic

matrices has shown less unbiased estimated breeding values (Erbe *et al.*, 2012), no evident increase or decrease of genomic reliabilities and bias (Rio *et al.*, 2019), and accurate estimates of genetic correlations between populations (Wientjes *et al.*, 2017). Some of these studies just suggested these genomic matrices, but did not compare them with others in terms of predictive ability and bias (Chen *et al.*, 2013; Wientjes *et al.*, 2017).

Furthermore, studies have applied different strategies to account for population structure by combining fixed effects, models, and different genomic relationship matrices (Lyra *et al.*, 2018; Rio *et al.*, 2019). The lack of comparison between these strategies difficult the decision making about considering or not considering population structure in GWS. Thus, this study aimed to evaluated some strategies by combining models, admixed proportions, and genomic relationship matrices that account or not account for population structure with simulated and real data.

3.2 Material and methods

3.2.1 Simulation of data

The simulated genome was based on a *Eucalyptus grandis* species ($2n = 22$ chromosomes) with a total size of 13 Morgans by using “HaploSim” package (Coster and Bastiaansen, 2009) in software R (R Core Team, 2019). In total, 10 000 loci were equally distributed along all the simulated genome and, thus, six hundred gametes were combined, resulting in three hundred non-inbred individuals (N). The recombination number followed a Poisson distribution and the recombination positions were randomly chosen. Reproduction between genotypes occurred over 6 000 generations with a mutation rate of 10^{-5} to reach an equilibrium state between genetic drift and mutation (Daetwyler *et al.*, 2013). This mutation rate was higher than those found for *E. grandis* (4.96×10^{-9} to 4.8×10^{-7} per generation and base pairs) (Silva-Junior and Grattapaglia, 2015).

After 6 000 generations, one hundred gametes were separated into three populations (*Homo 1*, *Homo 2*, and *Homo 3*) with N_e of 99 each and random crossings were simulated over the course of 75 generations for each one. Gametes from these three population were mixed to obtain two admixed populations (*Mix 4* and *Mix 5*). Population *Mix 4* was composed by 50, 25, and 25 gametes of populations *Homo 1*, *Homo 2*, and *Homo 3*, respectively and *Mix 5* was composed by 45, 35, and 20 gametes of populations *Homo 1*, *Homo 2*, and *Homo 3*, respectively. All populations had 100 genotypes and

each of them contributed two descendent ($\bar{k} = 2$) to next generation with equal number of descendent per parent ($\sigma_k^2 = 0$). Thus, the effective population size (N_e) was estimated by using the following equation: $(2N - 1)/(\bar{k} - 1 + \sigma_k^2/\bar{k})$, where \bar{k} is the mean number of descendants per parent, N is the total number of individuals, and σ_k^2 is the variance in the number of progeny per parent (Crow and Kimura, 1970), consequently, N_e was equal to 99 in this simulation for each population ($N = 100$, $\bar{k} = 2$, and $\sigma_k^2 = 0$). After mixing, random crossing occurred in all populations during generation 6 076 along with partial diallel cross per population based on a prior pedigree in generation 6 077. The pedigree consisted of 15 different genotypes (10×5) with 4 individuals per family and, consequently, 200 individuals per population and 1 000 individuals in total population (metapopulation). The mutation rate was null to avoid low frequency markers in last two generations. The total number of single nucleotide polymorphism (SNP) markers was 4 768 at the end of simulation.

The simulation of phenotypes consisted of five hundred QTLs randomly allocated to loci with minor allele frequencies over 0.01 for each of the two simulated traits. The QTL effect (α) was simulated according to average effect of allelic substitution and was given by $\alpha = a + d(1 - 2p)$, a is the additive effect, d is the genotypic effect of the heterozygous, and p is the QTL allele frequency. The a effect followed Gaussian distribution ($a \sim N(0, 0.25)$) for each trait. The d effect was estimated by multiplying a and degree of dominance ($DD = d/a$), which followed Gaussian distribution ($DD \sim N(0, 1/9)$). After obtaining a and d , α was estimated for each population according to allele frequency (p) of each QTL. Hence, populations had different QTL effects from each other. The true breeding values (TBVs) for each individual in each population were obtained by multiplying the incidence matrix of QTL and QTL effects. Random residuals (e) were added ($e \sim N(0, \sigma_e^2)$), where σ_e^2 was given by $\sigma_e^2 = \sigma_{TBV}^2(1 - h^2)/\sigma_{TBV}^2$, σ_{TBV}^2 is the variance of TBV in each population and h^2 (trait heritability) was 0.1 and 0.5 for traits 1 (DBH-01) and 2 (DBH-05), respectively. A fixed mean of 50 (cm) was added to simulate diameter at breast height. The supplemental figure 3.S1 resumes all simulation procedure employed in this study.

LD was estimated by r^2 , which is equivalent to the linear correlation between alleles from different loci. The LD was corrected by population structure (Mangin *et al.*, 2012). Nonlinear regressions between physical distance (Mb) and r^2 were fitted based on a drift-recombination model (Hill and Weir, 1988). LD estimates (r^2) were obtained using the “LDcorSV” package (Desrousseaux *et al.*, 2017) in R software (R Core Team,

2019). The genome-wide average recombination rate found for *E. grandis* is, approximately, 3.18 cM Mb^{-1} (Silva-Junior and Grattapaglia, 2015), thus, a recombination rate of 3 cM Mb^{-1} was considered in this study. The LD phase persistence (r_{PP}) was estimated by the Pearson correlation of LD values between populations for each 0.1 Mb bin.

3.2.2 Real data

Rice (*Oryza sativa* L.) data (Zhao *et al.*, 2011) from a worldwide panel with 413 accessions were used for this study. In this dataset, there are five rice populations (AUS, IND – Indica, TEJ – Temperate japonica, TRJ – Tropical japonica, and Aromatic) and an admixed group (ADMIX) that was not assigned to any population, but was considered a separated population in this study. These populations were established according to four principal components obtained using ENGEINSOFT software (Price *et al.*, 2006). All rice cultivars were genotyped with SNP chip 44K (Affymetrix). Flowering time (Arkansas) (FT) and panicle length (PL) were randomly chosen for this study among 34 traits.

The Aromatic population (14 lines) was removed from the data. Missing marker data were input from Zhao *et al.* (2011) and BEAGLE software (Browning and Browning, 2009). Individuals without values of FT and PL were removed and the final dataset was composed by 337 individuals and 36 816 markers. This dataset was chosen because it has a strong population structure (Zhao *et al.*, 2011; Guo *et al.*, 2014; Isidro *et al.*, 2015; Lehermeier *et al.*, 2015).

3.2.3 Strategies for accounting population structure effect

The adopted strategies involved different models (uni and multipopulation models), different genomic relationship matrices, and the genome population proportion of each individual (admixed proportions). Table 3.1 summarizes all strategies and one of them did not account population structure (*UG* strategy).

Table 3.1 Strategies for accounting population structure, their codification used in this study (Cod.), type of model applied (Models), fixed effects and the genomic relationship matrix.

Cod. ^a	Models	Fixed effects	Relationship matrix
<i>UG</i>	Uni-population	overall mean	<i>G</i>
<i>UFG</i>	Uni-population	admixed proportions	<i>G</i>
<i>UFG_A</i>	Uni-population	admixed proportions	<i>G_A</i>
<i>UFG_α</i>	Uni-population	admixed proportions	<i>G_α</i>

UFG_W	Uni-population	admixed proportions	G_W
MG	Multi-population	overall mean	G
MFG	Multi-population	admixed proportions	G
MFG_A	Multi-population	admixed proportions	G_A
MFG_α	Multi-population	admixed proportions	G_α
MFG_W	Multi-population	admixed proportions	G_W

^a U : uni-population model, M : multi-population model, F : admixed proportions used as fixed effects, and genomic relationship matrices: G (VanRaden, 2008), G_α (Chen *et al.*, 2013), G_A (Thornton *et al.*, 2012), and G_W (Wientjes *et al.*, 2017).

The admixed proportions were estimated by ADMIXTURE method (Alexander *et al.*, 2009) with package “radmixture” (Bian, 2017) in software R (R Core Team, 2019), resulting in a matrix ($N \times K$) with population genome proportions of K populations for N individuals. The genomic relationship matrices were G (VanRaden, 2008), G_α (Chen *et al.*, 2013), G_W (Wientjes *et al.*, 2017), and G_A (Thornton *et al.*, 2012). The G was estimated by WW'/m , where W is centralized maker matrix and $m = \sum_{i=1}^M 2p_i(1 - p_i)$ where p_i is average allelic frequency of all populations for the i^{th} marker, and M is the total number of markers. The G_A matrix was similar to that of Rio *et al.* (2019) and Conomos *et al.* (2016), and it was estimated as suggested by Thornton *et al.* (2012). The G_α (Chen *et al.*, 2013), G_A (Thornton *et al.*, 2012), and G_W (Wientjes *et al.*, 2017) were estimated by equation 1, 2, and 3, respectively.

$$\hat{G}_{\alpha_{ij}} = \frac{\sum_{s=1}^M (g_{isk} - 2p_{sk})(g_{j sk} - 2p_{sk})}{2 \sum_{s=1}^M [p_{sk}(1 - p_{sk})p_{sk}(1 - p_{sk})]^{0.5}} \quad \text{Eq. 1}$$

$\hat{G}_{\alpha_{ij}}$ (Chen *et al.*, 2013) is the genomic relationship estimator between individuals i and j , g_{isk} and $g_{j sk}$ are the number of alleles (2, 1, or 0) of individuals i and j , respectively for the s^{th} SNP in the k^{th} population, and p_{sk} is the allele frequency of k^{th} population for the s^{th} SNP.

$$\hat{G}_{A_{ij}} = \frac{\sum_{s=1}^M (g_{is} - 2\hat{\mu}_{is})(g_{js} - 2\hat{\mu}_{js})}{2 \sum_{s=1}^M [\hat{\mu}_{is}(1 - \hat{\mu}_{is})\hat{\mu}_{js}(1 - \hat{\mu}_{js})]^{0.5}} \quad \text{Eq. 2}$$

$\hat{G}_{A_{ij}}$ (Thornton *et al.*, 2012) is the genomic relationship estimator between individuals i and j based on admixture proportions, g_{is} and g_{js} are the number of alleles (2, 1, or 0) of individual i and j , respectively for the s^{th} SNP, $\hat{\mu}_{is}$ and $\hat{\mu}_{js}$ are the individual allele frequency of individuals i and j , respectively for s^{th} SNP. The $\hat{\mu}_{is}$ and $\hat{\mu}_{js}$ are estimated by multiplying admixed proportion matrix ($N \times K$) and allele frequency of populations ($K \times S$).

$$\hat{G}_{W_{ij}} = \frac{\sum_{s=1}^M (g_{isk} - 2p_{sk})(g_{j sk} - 2p_{sk})}{[\sum_{s=1}^M 2p_{sk}(1 - p_{sk})]^{0.5} [\sum_{s=1}^M 2p_{sk}(1 - p_{sk})]^{0.5}} \quad \text{Eq. 3}$$

\hat{G}_{Wij} (Wientjes *et al.*, 2017) is the genomic relationship estimator between individuals i and j , g_{isk} and g_{jsk} are the number of alleles (2, 1, or 0) of individuals i and j , respectively for the s^{th} SNP in the k^{th} population, and p_{sk} is the allele frequency of k^{th} population for the s^{th} SNP.

Uni-population model was given by $y = X\beta + Zg + e$, y is phenotypic vector of all individuals (N), β is the fixed vector, g is the genetic effects vector ($Zg | \sigma_g^2 \sim N(0, G\sigma_g^2)$), and e is the residual effects vector ($e \sim N(0, I\sigma_e^2)$). X and Z are incidence matrices of β and g , respectively. A scaled inverse chi-squared distribution was established as a prior distribution for genetic variance (σ_g^2) ($\sigma_g^2 \sim X^{-2}(df_g, S_g)$) and the same prior distribution was adopted for residual variance (σ_e^2) ($\sigma_e^2 \sim X^{-2}(df_e, S_e)$). The hyperparameters df_g , S_g , df_e , and S_e had values of 4, 3, 4, and 3, respectively (Lehermeier *et al.*, 2015). Multi-population model considered populations like different traits of a multi-trait model, using the following model:

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{ik} \end{bmatrix} = \begin{bmatrix} x_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & z_k \end{bmatrix} \begin{bmatrix} g_1 \\ \vdots \\ g_k \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_k \end{bmatrix}$$

y_{ik} is phenotypic value of the i^{th} individual in the k^{th} population, β_k is the fixed effects vector of the k^{th} population, g_k is the genomic value of the individual within the

k^{th} population, ($z_k g_k | \Sigma_g \sim MVN \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} G_{11}\sigma_1^2 & \dots & G_{1k}\sigma_{1k}^2 \\ \vdots & \ddots & \vdots \\ G_{k1}\sigma_{k1}^2 & \dots & G_{kk}\sigma_k^2 \end{bmatrix} \right)$), and e_k is the residual

effects vector of the k^{th} population ($e_k \sim MVN \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} I\sigma_{e1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I\sigma_{ek}^2 \end{bmatrix} \right)$). x_k and z_k are

the incidence matrices of β_k and g_k of the k^{th} population, respectively. Σ_g is the covariance matrix between the k^{th} populations and followed the inverse-Wishart distribution with ν degrees of freedom, and the ω scale parameter as a prior distribution ($\Sigma_g \sim W^{-1}(\omega, \nu)$). A scaled inverse chi-squared distribution was established for σ_{ek}^2 as a prior distribution ($\sigma_{ek}^2 \sim X^{-2}(df_e, S_e)$) where df_e is the degree of freedom and S_e is the scale parameter. The hyperparameters df_e , S_e , ν , and ω had values of 4, 3, 5, and 4, respectively (Lehermeier *et al.*, 2015). Thus, the *MG* strategy was the same as that of Lehermeier *et al.* (2015) (*MG-BLUP* model).

Strategies with uni-population model (*UG*, *UFG*, *UFG_A*, *UFG _{α}* , and *UFG_W*) were fitted through the Bayesian Genomic Best Linear Unbiased Predictor (*GBLUP*) with the

BGLR package (Pérez and de los Campos, 2014). Strategies with multi-population models (MG , MFG , MFG_A , MFG_α , and MFG_W) were fitted with the MTM package (de los Campos and Grüneberg, 2016). In total, 80 000 iterations were used, with a burn-in (iterations removed) of 30 000 and a thinning of 5. Markov chain convergence was evaluated using graphics and the Raftery and Lewis (1992) and Geweke (1992) methods with the “coda” package (Plummer *et al.*, 2010). All statistical analyses were performed in R software (R Core Team, 2019).

3.2.4 Evaluation of strategies

In total, 10% and 20% of individuals in each population were randomly sampled to compose the validation set in simulated and real data, respectively. This procedure samples individuals according to the population size and shows to perform better in structured populations (Isidro *et al.*, 2015). This procedure was repeated 40 times to evaluate all strategies. The strategies were compared by the following parameters: predictive ability ($r_{y\hat{y}}$), bias, accuracy ($r_{g\hat{g}}$), and bias of TBV ($Bias_{TBV}$) for each population. The last two parameters were evaluated only for simulated data. The $r_{y\hat{y}}$ was estimated by Pearson correlation between predicted phenotypes and phenotypes of validation sets, the $r_{g\hat{g}}$ was also estimated by Pearson correlation between TBV and GEBVs of the validation set phenotypes. Bias and $Bias_{TBV}$ were estimated by one minus the linear regression coefficient between predicted phenotypes and phenotypes of validation sets, and between TBV and GEBVs of validation sets, respectively. The predicted genotypes were given by summing the GEBVs and the estimated fixed effects (admixed proportions or the overall mean, depending on the strategy). Confidence intervals with 5% probability of error type I were constructed to compare these parameters for each strategy, population, and trait.

The genomic heritability (h_g^2) was estimated for all the strategies and populations, and was given by $h_g^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$, σ_g^2 is the genomic variance and σ_e^2 is the residual variance. Credible intervals with 5% probability of error type I were also constructed for h_g^2 . All the genomic relationship matrices used in the strategies (G , G_A , G_α , and G_W) had their population structure visualized by principal component analyses. The fixation index (F_{ST}) (Wright, 1965) between populations was also estimated by using the method of Weir and Cockerham (1984) with “hierfstat” package (Goudet and Jombart, 2015) in R software (R Core Team, 2019).

3.3 Results

3.3.1 Analyses of population structure

The matrices G_A , G_α , and G_W did not capture most of population structure like G matrix for real and simulated data (Fig 3.1). The explanation of variance by the first two principal components were lower for G_A , G_α , and G_W than G . The heatmap of these matrices also showed the population structured captured by them in supplemental figures 3.S2 and 3.S3 for simulated and real data, respectively. The LD of metapopulation and each population and the LD phase persistence (r_{PP}) between populations were shown in supplemental figure 3.S4. The fixation index (F_{ST}) was higher (> 0.20) between populations *Homo* 1, 2, and 3 and between *Homo* 3 and *Mix* 4 and 5 based on markers for simulated data, which is according to simulation (Table 3.2). The F_{ST} estimated with real data was higher than those found for simulated data. However, population ADMIX showed the lower F_{ST} value with population TRJ (Table 3.2).

Table 3.2 Fixation index estimated with markers between populations for simulated and real data.

Data	Populations	Populations			
		<i>Homo</i> 2	<i>Homo</i> 3	<i>Mix</i> 4	<i>Mix</i> 5
Simulated	<i>Homo</i> 1	0.3723	0.3845	0.1008	0.1333
	<i>Homo</i> 2		0.3858	0.1899	0.1697
	<i>Homo</i> 3			0.2442	0.2088
	<i>Mix</i> 4				0.0361
Real		AUS	IND	TEJ	TRJ
	ADMIX	0.5040	0.4941	0.2611	0.1544
	AUS		0.4780	0.7889	0.7152
	IND			0.7665	0.6933
	TEJ				0.5118

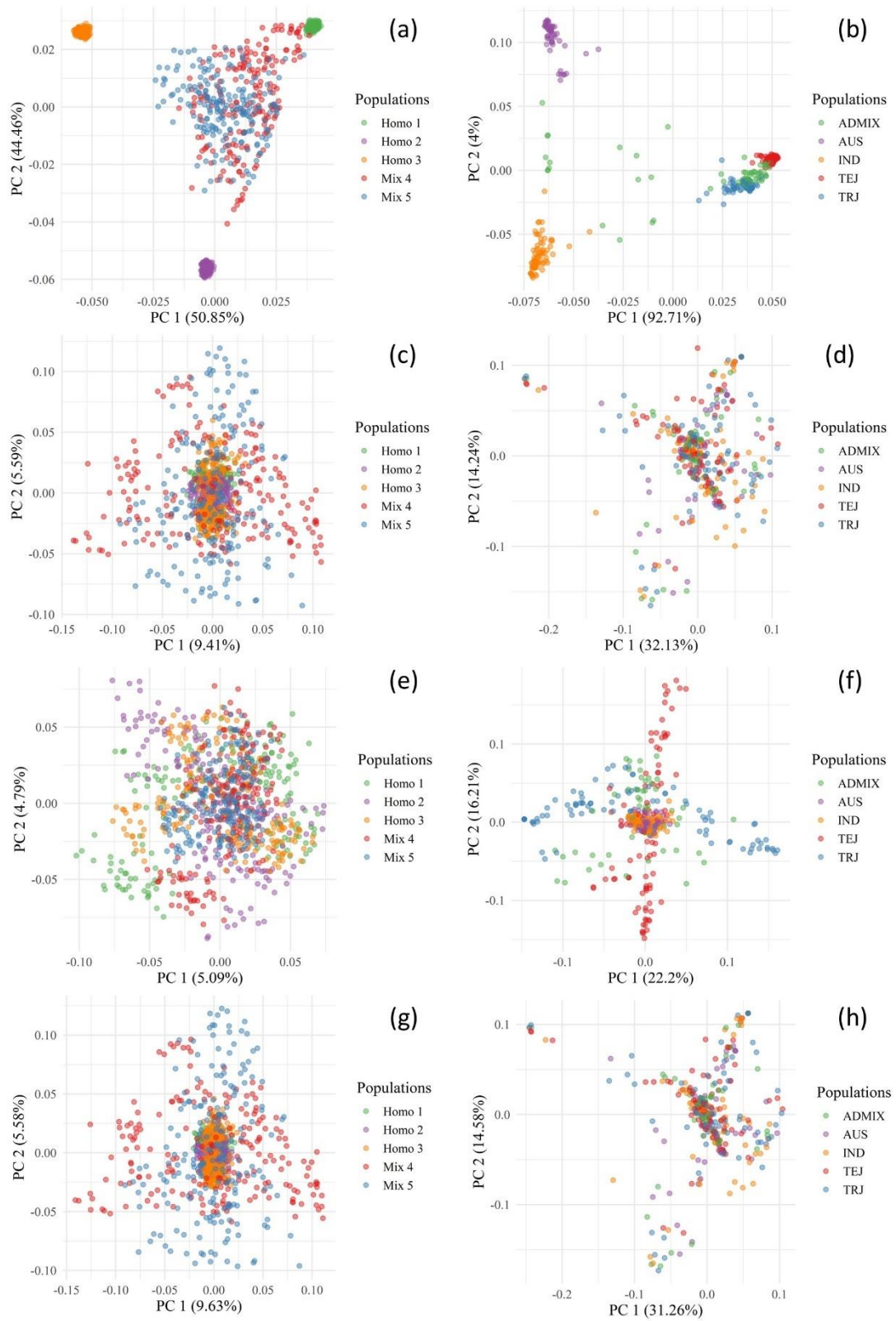


Fig. 3.1 Principal component analysis of matrices G (a and b), G_α (c and d), G_A (e and f), and G_W (g and h) for simulated (a, c, e, and g), and real data (b, d, f, and h).

3.3.2 Predictive ability ($r_{y\hat{y}}$) and accuracy ($r_{g\hat{g}}$)

All strategies had the same $r_{y\hat{y}}$ ($p \geq 0.05$) in all populations for both simulated traits (Supplemental Fig 3.S5 and 3.S6 and tables 3.S1 and 3.S2) and also showed the same $r_{g\hat{g}}$ ($p \geq 0.05$) in all populations for DBH-01 (Supplemental Fig 3.S7 and table 3.S1). However, strategies with matrices G_A , G_α , and G_W showed lower $r_{g\hat{g}}$ values ($p < 0.05$) than UG and MG strategies for high heritability trait (DBH-05) with uni and multi-population models in *Mix 4* and *Mix 5* populations, respectively (Fig 3.2 and table 3.S2). Hence, applying matrices that account for the population structure is not recommended for high heritability traits for genomic prediction of admixed individuals. Moreover, for real data, all strategies with G_W and G_α showed lower ($p < 0.05$) $r_{y\hat{y}}$ values than the most of strategies and populations for both traits (Supplemental Fig 3.S8 and 3.S9, and table 3.S3).

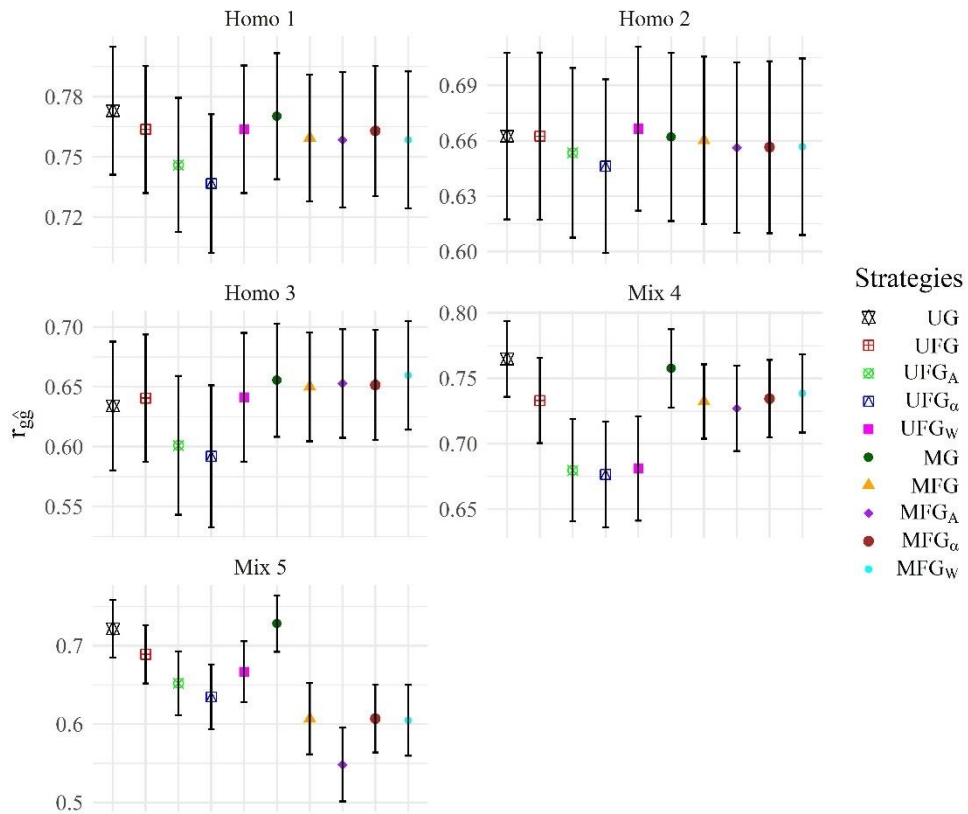


Fig. 3.2 Accuracy ($r_{g\hat{g}}$) of DBH-05 ($h^2 = 0.5$) in homogeneous (*Homo 1, 2, and 3*) and admixed populations (*Mix 4 and 5*) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are 95% confidence intervals.

3.3.3 Prediction bias and bias of the true breeding value ($Bias_{TBV}$)

Most of strategies showed bias equal to zero ($p \geq 0.05$) in all populations for high heritability trait (DBH-05) (Supplemental Fig 3.S10 and table 3.S2). However, bias were closest to zero for strategies with multi-population model and admixed proportions as fixed effects (MFG , MFG_A , MFG_α , and MFG_W) to most of populations in low heritability trait (DBH-01) (Supplemental Fig 3.S11 and table 3.S1). In addition, this result was more evident for $Bias_{TBV}$ of DBH-01, which highlighted the underestimation of GEBVs ($Bias_{TBV}$ lower than zero and $p < 0.05$) for uni-population model (Fig 3.3 and table 3.S2). However, differences between strategies for $Bias_{TBV}$ was reduced in high heritability trait (DBH-05) (Supplemental Fig 3.S12 and table 3.S1). Hence, accounting for population structure through multi-population model is crucial to achieve lower bias for low-heritability traits.

Moreover, the $Bias_{TBV}$ of DBH-01 showed values closest to zero for strategies (UFG_A , UFG_α , and UFG_W) than strategy UG in populations *Homo* 1 and 3 (Fig 3.3). Thus, the matrices G_A , G_α , and G_W slightly contributed for controlling bias in uni-population model. However, considering real data, matrices G_α and G_W showed larger confidence intervals for bias in UFG_α and UFG_W strategy and bias was not equal to zero ($p < 0.05$) in MFG_α and MFG_W strategy for most of populations and traits (Supplemental Fig 3.S13 and 3.S14 and table 3.S3).

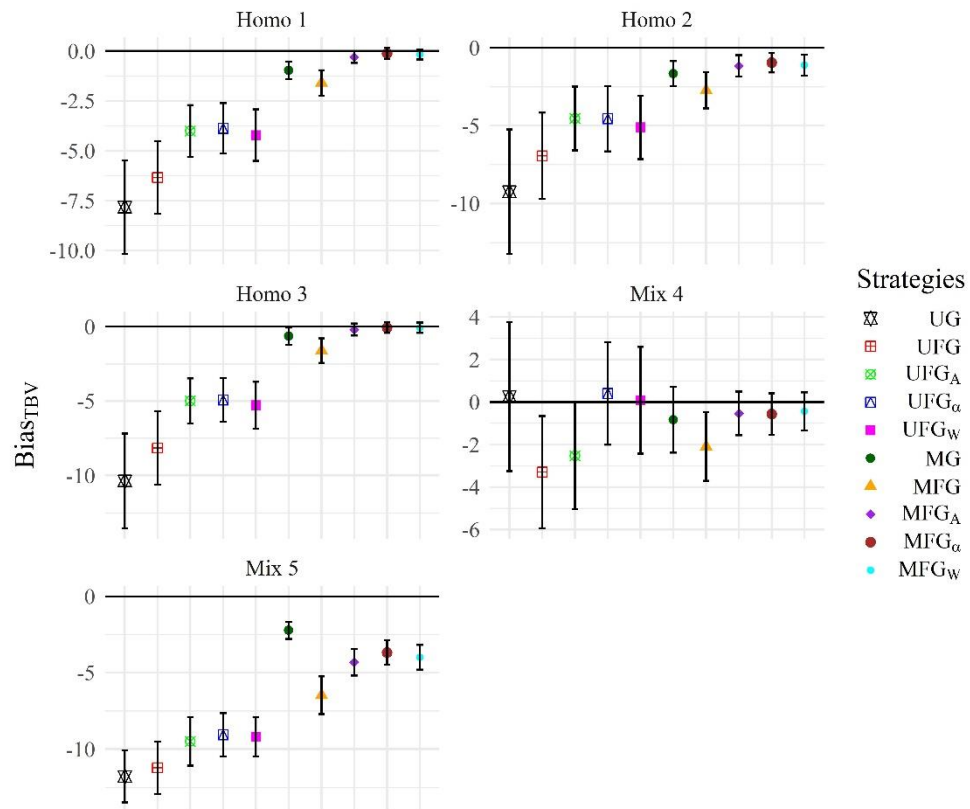


Fig. 3.3 Bias of true breeding value ($Bias_{TBV}$) of DBH-01 ($h^2 = 0.1$) in homogeneous (*Homo* 1, 2, and 3) and admixed populations (*Mix* 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are 95% confidence intervals.

3.3.4 Genomic heritability

The overestimation of genomic heritability values (h_g^2) occurred when the simulated values were under the inferior limit of the 95% credibility interval, and the underestimation occurred when the simulated values were above the superior limit of the 95% credibility interval. Thus, genomic heritability was underestimated for all strategies by uni-population models for DBH-01 trait (Supplemental figure 3.S15 and table 3.S4) and underestimated for strategy UFG_α in most of populations for DBH-05 (Supplemental figure 3.S16 and table 3.S4). Multi-population models showed larger credible intervals for both simulated traits (Fig 3.S15 and 3.S16). For real data, the strategies with G_α and G_W matrix showed lower h_g^2 values for both traits in most of populations, mainly, for uni-population models and for flowering time trait (Supplemental figures 3.S17 and 3.S18 and table 3.S5). Further, strategies UG , UFG , and UFG_A showed lower ($p < 0.05$) h_g^2 values than MG , MFG , and MFG_A in most of populations for flowering time.

3.4 Discussion

Many strategies have been applied to account for population structure, aiming a better prediction in combined population data (Azevedo *et al.*, 2017; Lyra *et al.*, 2018; Rio *et al.*, 2019). However, the advantages and disadvantages of these strategies have varied between traits and populations (Karoui *et al.*, 2012; Schulz-Streeck *et al.*, 2012; Zhou *et al.*, 2014; Lehermeier *et al.*, 2015), which hinder the choice of the most suitable strategy. In this study, some strategies were explored including uni and multi-population models, admixture proportions as fixed effects, different genomic relationship matrices, real and simulated traits, homogenous and admixed populations. Thus, this study should be used as a guide to choose suitable strategies, when combined populations are used for genomic prediction.

3.4.1 Predictive ability ($r_{y\hat{y}}$) and accuracy ($r_{g\hat{g}}$)

The strategies that accounted for population structure showed no or small increase of $r_{y\hat{y}}$ to most of traits and populations as observed in others studies (Karoui *et al.*, 2012; Technow *et al.*, 2012; Thomasen *et al.*, 2013; Rio *et al.*, 2019). However, other studies found advantages of accounting for population structure with multi-population models depending on population and trait, even if the $r_{y\hat{y}}$ is slightly better (Schulz-Streeck *et al.*, 2012; Makgahlela *et al.*, 2013; Zhou *et al.*, 2014; Lehermeier *et al.*, 2015; Lyra *et al.*, 2018). In this study, strategies with modified genomic relationship matrix to account for population structure showed lower $r_{g\hat{g}}$ for high heritability traits ($h^2 = 0.5$) in admixed populations with simulated data (Fig 3.2). Thus, accounting for population structure effect by modifying the genomic relationship matrix and including admixed proportions as fixed effects can reduce accuracy in admixed populations for high-heritability traits and should be avoided.

Furthermore, in this study, the type of genomic relationship matrix influenced $r_{y\hat{y}}$ in real data (Fig 3.S8 and 3.S9) independently of the model (uni or multi-population), likely due to the trait genetic architecture. The trait was polygenic (500 QTLs) in simulated data while, in real data, flowering time and panicle length have likely been controlled by few QTLs of larger effect (Zhao *et al.*, 2011; Spindel *et al.*, 2015; Sun *et al.*, 2017) and genetic architecture of panicle length also showed differences among subpopulations (Zhao *et al.*, 2011). Other studies did not find low $r_{y\hat{y}}$ when applying G_W (Schultz and Weigel, 2019) or G_α (Hidalgo *et al.*, 2015; Duhnen *et al.*, 2017) and modifications of G_α and G_W has shown higher $r_{y\hat{y}}$ for animal traits (Veroneze *et al.*, 2015;

Raymond *et al.*, 2018). However, in these studies, most of traits were productive (milk-production, soybean yield, soybean protein yield, litter birth weight), which are likely controlled by many QTLs (Falconer and Mackay, 1996). Thus, the genomic relationship matrices G_α and G_W should be used with caution for genomic prediction of traits controlled by few genes.

3.4.2 Bias and bias of true breeding values

Lower bias is important for the selection of individuals based on a combination of their genomic values, for the mating contributions proportional to their GEBVs (Daetwyler *et al.*, 2013), and comparison of individuals across generations (Henderson *et al.*, 1959). Hence, bias analysis has been suggested as an important parameter for evaluation of genomic prediction models (Vitezica *et al.*, 2011; Daetwyler *et al.*, 2013). In population structure context, some studies found slightly varied bias among traits and populations, considering uni and multi-population models, and no advantage of applying multi-population model was evident (Karoui *et al.*, 2012; Makgahlela *et al.*, 2013; Hidalgo *et al.*, 2015).

In this study, multi-population models showed lower or no bias or $Bias_{TBV}$ for low heritability traits ($h^2 = 0.1$) in simulated data. Moreover, $Bias_{TBV}$ of low-heritability trait ($h^2 = 0.1$) highlighted that strategies using uni-population model can result in underestimation of GEBVs. Thus, accounting for population structure can avoid higher bias and $Bias_{TBV}$ for low-heritability traits, mainly by using multi-population model.

However, the use of uni-population model and the matrices G_A , G_α , and G_W slightly improved $Bias_{TBV}$ for polygenic low-heritability traits ($h^2 = 0.1$) in some populations. Hence, this is advantageous when there are many populations and multi-population models become infeasible due to the slow convergence of parameters. Further, strategies that used G_α and G_W matrices showed higher bias in most of populations with real data for both traits. However, some studies did not find higher bias when applying G_W (Schultz and Weigel, 2019), G_α matrix (Hidalgo *et al.*, 2015) or modifications of G_α matrix (Veroneze *et al.*, 2015) for most of traits. Thus, likewise for $r_{y\hat{y}}$, G_α and G_W matrices or any type of relationship matrix should be cautiously used for genomic prediction, depending on the trait.

3.4.3 Genomic heritability (h_g^2)

The uni-population models with G_α and G_W matrices underestimated h_g^2 values of high-heritability simulated trait in *Mix 5* population (Fig 3.S16) and showed lower h_g^2 values in all populations with real data for both of traits (Fig 3.S17 and 3.S18), in this study. Although, the G_W matrix was expected to accurately estimate h_g^2 as proposed by Wientjes *et al.* (2017), these authors applied G_W matrix with multi-population model for a high heritability simulated trait ($h^2 = 0.9$). Thus, likewise in this study, the estimated h_g^2 of the high heritability simulated trait ($h^2 = 0.5$) for multi-population model was not under or overestimated (Fig 3.S16).

However, the multi-population model strategies that used G_α and G_W matrices showed lower h_g^2 values with real data for flowering time in IND and TEJ populations (Fig 3.S17). Thus, likewise for $r_{y\hat{y}}$ and bias, these lower values indicates that these matrices are likely affected by traits controlled by few QTLs of larger effect. Further, the G_W matrix has already shown lower h_g^2 values, when was constructed with a restricted set of markers selected by a meta-GWAS analysis (Raymond *et al.*, 2018). However, when these authors constructed G_W with all marker, there were no under or overestimation for simulated trait heritability. Other study did not find differences in h_g^2 with genomic matrix weighted by marker effects, but only two populations were used (Zhou *et al.*, 2014). A modification of G_α matrix considering marker effects in its construction showed lower h_g^2 values depending on the populations in the training set (Veroneze *et al.*, 2015). Thus, there is likely an interaction between populations in training set, trait genetic architecture, and type of genomic matrix that affects the estimates of h_g^2 .

Moreover, G_A , G_α , and G_W estimate the genomic variance within the current population, while genomic variance estimated through genomic relationship proposed by VanRaden, (2008) has no clear definition in structured populations (Wientjes *et al.*, 2017). However, the UFG_A strategy did not show lower h_g^2 values for real data traits like UFG_α and UFG_W strategies (Fig 3.S17 and 3.S18). This was likely because the G_A matrix considers individual allele frequencies, instead of population allele frequency by weighting population allele frequencies with individual admixture proportions. Hence, if all individuals exclusively belonged to only one population, admixture proportions would be a matrix of zeros and ones and G_A would be the same as G_α and with the same numerator of G_W . However, the rice populations were admixed in this study and G_A was

likely more similar to G than G_α . Consequently, h_g^2 values of UFG_A strategy was higher than h_g^2 values of UFG_α and UFG_W , and the same as UG and UFG strategies for real data.

In this study, multi-population models showed higher genomic heritability than uni-population models for flowering time in most of populations (Fig 3.S17). Although, multi-population model has shown higher estimates of genomic heritability than uni-population models in real data, it is not possible to prove that it is an overestimation (Lehermeier *et al.*, 2015). However, other study did not find different values of genomic heritability between uni and multi-population models and the use of fixed population structure covariates in uni-population models (Lyra *et al.*, 2018). Although, multi-populations model seemed not to overestimate genomic heritabilities, this should be deeply investigated, mainly for traits controlled by few QTLs.

3.4.4 Strategy recommendations

The recommendations will depend mostly on the trait architecture and heritability. For traits controlled by few QTLs, the recommendations are the same for polygenic traits, except G_α and G_W matrix should be avoided for genomic prediction. Thus, considering polygenic and small heritability traits, no strategy to account for population structure is needed for higher accuracy, but if there is necessity of lower bias, multi-population model should be preferred. Further, if multi-population model is forbidden for some reason, uni-population model with modified genomic relationship matrices (UFG_α , UFG_A , and UFG_W) should be preferred.

Considering polygenic and high heritability traits, uni and multi-population model with modified genomic relationship matrices should be avoided in genomic prediction of admixed populations, due to a probable reduction of accuracy. In this case, strategies UG and MG are preferred. However, if there are many populations, uni-population model can be advantageous, due to the slow convergence of parameters with multi-population model.

3.5 Conclusions

The use of structured populations requires strategies, like multi-population model and/or modifications of genomic matrix for a lower bias for low-heritability traits. However, uni and multi-population model with modified genomic relationship matrices can reduce accuracy of high-heritability traits in genomic prediction of admixed populations. In addition, modified genomic matrix to account populations structure should be cautiously applied in genomic prediction, depending on the trait architecture.

Therefore, the strategy of accounting for population structure, when combined populations are used in the training set, should be carefully chosen according to trait heritability, genetic trait architecture, and admixture level of population to obtaining lower bias without reduction of accuracy, and, consequently, success of genomic prediction.

3.6 Acknowledgments

We are thankful for the financial support from of National Council for Scientific and Technological Development (CNPq: #155113/2017-8) and the Coordination for the Improvement of Higher Education Personnel (Finance Code 001). We are grateful to professors Andrei C. P. Nunes, Felipe L. da Silva, Rafael T. Resende, and Moysés Nascimento for their important comments, suggestions, and criticisms.

3.7 References

- Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.
- Azevedo CF, Resende MDV De, Silva FF, Nascimento M, Viana JMS, Valente MSF (2017). Population structure correction for genomic selection through eigenvector covariates. *Crop Breed Appl Biotechnol* **17**: 350–358.
- Bian B (2017). radmixture: calculate population stratification. *R package*.
- Brown BC, Ye CJ, Price AL, Zaitlen N (2016). Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet* **99**: 76–88.
- Browning BL, Browning SR (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**: 210–223.
- Chen L, Schenkel F, Vinsky M, Crews Jr. DH, Li C (2013). Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *J Anim Sci* **91**: 4669–4678.
- Conomos MP, Miller MB, Thornton TA (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**: 276–293.
- Conomos MP, Reiner AP, Weir BS, Thornton TA (2016). Model-free estimation of recent genetic relatedness. *Am J Hum Genet* **98**: 127–148.
- Coster A, Bastiaansen JWM (2009). HaploSim. *R Package version*: 1.8.

- Crow JF, Kimura M (1970). *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row, Publishers.
- Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**: 347–365.
- Daetwyler HD, Kemper KE, Werf JHJ van der, Hayes BJ (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci* **90**: 3375–3384.
- Desrousseaux D, Sandron F, Siberchicot A, Cierco-Ayrolles C, Mangin B, Siberchicot MA (2017). ‘LDcorSV’. *R package*.
- Duhnen A, Gras A, Teyssèdre S, Romestant M, Claustres B, Daydé J, *et al.* (2017). Genomic selection for yield and seed protein content in soybean: a study of breeding program data and assessment of prediction accuracy. *Crop Sci* **57**: 1325–1337.
- Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, *et al.* (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci* **95**: 4114–4129.
- Falconer DS, Mackay TFC (1996). *Introduction to quantitative genetics*. 4th edn. Prentice Hall, Essex.
- Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- Geweke J (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Stat* **4**: 641–649.
- Gianola D, Fariello MI, Naya H, Schön CC (2016). Genome-wide association studies with a genomic relationship matrix: a case study with wheat and arabidopsis. *G3 Genes, Genomes, Genet* **6**: 3241–3256.
- Goudet J, Jombart T (2015). hierfstat: estimation and tests of hierarchical F-statistics. *R Packag version 004-22* **10**.
- Grattapaglia D, Resende MD V (2011). Genomic selection in forest tree breeding. *Tree Genet Genomes* **7**: 241–255.
- Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, *et al.* (2014). The impact of

- population structure on genomic prediction in stratified populations. *Theor Appl Genet* **127**: 749–762.
- Henderson CR, Kempthorne O, Searle SR, Von Krosigk CM (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**: 192–218.
- Hidalgo AM, Bastiaansen JWM, Lopes MS, Harlizius B, Groenen MAM, de Koning DJ (2015). Accuracy of predicted genomic breeding values in purebred and crossbred pigs. *G3 Genes, Genomes, Genet* **5**: 1575–1583.
- Hill WG, Weir BS (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* **33**: 54–78.
- Isidro J, Jannink JL, Akdemir D, Poland J, Heslot N, Sorrells ME (2015). Training set optimization under population structure in genomic selection. *Theor Appl Genet* **128**: 145–158.
- Janss L, de los Campos G, Sheehan N, Sorensen D (2012). Inferences from genomic models in stratified populations. *Genetics* **192**: 693–704.
- Karoui S, Carabaño MJ, Díaz C, Legarra A (2012). Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet Sel Evol* **44**: 1–10.
- Lehermeier C, Schon CC, de los Campos G (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* **201**: 323–337.
- de los Campos G, Grüneberg A (2016). MTM (Multiple-Trait Model). *R package*.
- de los Campos G, Sorensen D (2014). On the genomic analysis of data from structured populations. *J Anim Breed Genet* **131**: 163–164.
- Lyra DH, Granato ÍSC, Morais PPP, Alves FC, dos Santos ARM, Yu X, *et al.* (2018). Controlling population structure in the genomic prediction of tropical maize hybrids. *Mol Breed* **38**.
- Makgahlela ML, Mäntysaari EA, Strandén I, Koivula M, Nielsen US, Sillanpää MJ, *et al.* (2013). Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *J Anim Breed Genet* **130**: 10–19.
- Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)* **108**: 285–291.

- Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.
- Pérez P, de los Campos G (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**: 483–495.
- Plummer M, Best N, Cowles K, Vines K (2010). coda: output analysis and diagnostics for MCMC. *R package* version 0.14-2.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.
- Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- R Core Team (2019). R: a language and environment for statistical computing.
- Raftery AE, Lewis SM (1992). [Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Stat Sci* **7**: 493–497.
- Raj A, Stephens M, Pritchard JK (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.
- Raymond B, Bouwman AC, Wientjes YCJ, Schrooten C, Houwing-Duistermaat J, Veerkamp RF (2018). Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genet Sel Evol* **50**: 1–14.
- Resende MDV De, Lopes PS, da Silva RL, Pires IE (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesqui Florest Bras* **56**: 63–77.
- Resende MD V, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, *et al.* (2012). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol* **194**: 116–128.
- Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, *et al.* (2017). Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. *Heredity (Edinb)* **119**: 245–255.

- Resende MD V, Silva FF, Lopes PS, Azevedo CF (2012). *Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial*. 1th edn. Universidade Federal de Viçosa, Viçosa (MG).
- Rio S, Mary-Huard T, Moreau L, Charcosset A (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theor Appl Genet* **132**: 81–96.
- de Roos APW, Hayes BJ, Goddard ME (2009). Reliability of genomic predictions across multiple populations. *Genetics* **183**: 1545–1553.
- Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, *et al.* (2005). Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* **13**: 677–686.
- Schultz NE, Weigel KA (2019). Inclusion of herd-mate data improves genomic prediction for milk-production and feed-efficiency traits within North American dairy herds. *J Dairy Sci* **102**: 11081–11091.
- Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP (2012). Genomic selection using multiple populations. *Crop Sci* **52**: 2453–2461.
- Silva-Junior OB, Grattapaglia D (2015). Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. *New Phytol* **208**: 830–845.
- Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, *et al.* (2015). Genomic selection and association mapping in rice (*Oryza Sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding line. *PLoS Genet* **11**: e1005350.
- Sun Z, Yin X, Ding J, Yu D, Hu M, Sun X, *et al.* (2017). QTL analysis and dissection of panicle components in rice using advanced backcross populations derived from *Oryza sativa* cultivars HR1128 and ‘Nipponbare’. *PLoS One* **12**: 1–13.
- Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet* **125**: 1181–1194.

- Thomasen JR, Sørensen AC, Su G, Madsen P, Lund MS, Guldbrandtsen B (2013). The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. *J Anim Sci* **91**: 3105–3112.
- Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N (2012). Estimating kinship in admixed populations. *Am J Hum Genet* **91**: 122–138.
- Toosi A, Fernando RL, Dekkers JCMM (2010). Genomic selection in admixed and crossbred populations. *J Anim Sci* **88**: 32–46.
- VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–4423.
- Veroneze R, Lopes MS, Hidalgo AM, Guimarães SEF, Silva FF, Harlizius B, *et al.* (2015). Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. *J Anim Sci* **93**: 1–21.
- Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011). Bias in genomic predictions for populations under selection. *Genet Res (Camb)* **93**: 357–366.
- Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N Y)* **38**: 1358–1370.
- Wientjes YCJJ, Bijma P, Vandenplas J, Calus MPLL (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. *Genetics* **207**: 503–515.
- Wright S (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution (N Y)*: 395–420.
- Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.
- Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, *et al.* (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* **2**: 467.
- Zheng YJ, Song Q, Chen SY (2013). Multiobjective fireworks optimization for variable-rate fertilization in oil crop production. *Appl Soft Comput J* **13**: 4253–4263.
- Zhou L, Lund MS, Wang Y, Su G (2014). Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *J Anim Breed Genet* **131**: 249–257.

3.8 Supplementary material

Table 3.S1 Mean values, lower (LL), and upper limit (UL) of 95% confidence interval of predictive ability ($r_{y\hat{y}}$), accuracy ($r_{g\hat{g}}$), bias ($Bias$), and true breeding value bias ($Bias_{TBV}$) for simulated trait DBH-01 ($h^2 = 0.1$) in all populations (Pop.) and strategies (Str.) (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type). Value in parentheses is the standard error (SE).

Pop.	Str.	$r_{y\hat{y}}$			$r_{g\hat{g}}$			$Bias$			$Bias_{TBV}$		
		Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL
Homo 1	<i>UG</i>	0.1 (0.0272)	0.05	0.16	0.31 (0.0345)	0.24	0.38	-10.92 (3.126)	-17.24	-4.6	-7.83 (1.1594)	-10.17	-5.49
	<i>UFG</i>	0.09 (0.0282)	0.03	0.15	0.33 (0.034)	0.26	0.4	-7.34 (2.5063)	-12.41	-2.28	-6.35 (0.9004)	-8.17	-4.53
	<i>UFG_A</i>	0.1 (0.0277)	0.04	0.15	0.31 (0.0337)	0.24	0.38	-5.51 (1.833)	-9.22	-1.81	-4.01 (0.6434)	-5.31	-2.71
	<i>UFG_α</i>	0.1 (0.0282)	0.04	0.15	0.3 (0.0344)	0.23	0.37	-5.46 (1.8365)	-9.17	-1.75	-3.88 (0.6268)	-5.15	-2.61
	<i>UFG_W</i>	0.09 (0.0279)	0.04	0.15	0.32 (0.0339)	0.25	0.39	-5.16 (1.849)	-8.89	-1.42	-4.22 (0.6393)	-5.51	-2.93
	<i>MG</i>	0.1 (0.0252)	0.05	0.15	0.35 (0.0316)	0.28	0.41	-1.47 (0.5859)	-2.66	-0.29	-0.97 (0.2183)	-1.41	-0.53
	<i>MFG</i>	0 (0.0292)	-0.06	0.06	0.31 (0.0319)	0.25	0.38	0.4 (0.5714)	-0.75	1.56	-1.6 (0.3158)	-2.24	-0.96
	<i>MFG_A</i>	0.05 (0.0275)	0	0.11	0.32 (0.0314)	0.26	0.38	0.07 (0.3843)	-0.71	0.85	-0.3 (0.1473)	-0.6	-0.01
	<i>MFG_α</i>	0.05 (0.0252)	0	0.1	0.32 (0.0301)	0.26	0.38	0.33 (0.3118)	-0.3	0.96	-0.13 (0.1409)	-0.41	0.16
	<i>MFG_W</i>	0.04 (0.0258)	-0.01	0.1	0.32 (0.0308)	0.26	0.38	0.36 (0.3173)	-0.28	1	-0.17 (0.1229)	-0.42	0.07
Homo 2	<i>UG</i>	-0.04 (0.0345)	-0.11	0.03	0.19 (0.0349)	0.12	0.26	5.67 (5.5156)	-5.48	16.82	-9.24 (1.9751)	-13.23	-5.25
	<i>UFG</i>	-0.05 (0.0349)	-0.12	0.03	0.19 (0.0344)	0.12	0.26	4.92 (4.3982)	-3.97	13.8	-6.93 (1.3689)	-9.7	-4.17
	<i>UFG_A</i>	-0.06 (0.0363)	-0.14	0.01	0.19 (0.0342)	0.12	0.26	5.84 (3.1818)	-0.59	12.27	-4.54 (1.0124)	-6.58	-2.49
	<i>UFG_α</i>	-0.07 (0.0357)	-0.14	0	0.19 (0.0346)	0.12	0.26	6.36 (3.0872)	0.12	12.6	-4.55 (1.0383)	-6.65	-2.45
	<i>UFG_W</i>	-0.04 (0.0344)	-0.11	0.03	0.21 (0.0347)	0.14	0.28	3.83 (2.9434)	-2.12	9.77	-5.11 (1.0054)	-7.14	-3.08
	<i>MG</i>	0 (0.032)	-0.07	0.06	0.25 (0.0324)	0.18	0.31	0.1 (1.038)	-2	2.2	-1.66 (0.3985)	-2.46	-0.85
	<i>MFG</i>	0 (0.0326)	-0.07	0.06	0.25 (0.0331)	0.19	0.32	0.3 (1.4105)	-2.55	3.15	-2.73 (0.5733)	-3.89	-1.57
	<i>MFG_A</i>	0 (0.0322)	-0.06	0.07	0.25 (0.0328)	0.18	0.32	0.06 (0.8706)	-1.7	1.82	-1.16 (0.3387)	-1.84	-0.47
	<i>MFG_α</i>	0 (0.0323)	-0.07	0.06	0.25 (0.0336)	0.18	0.32	0.18 (0.7838)	-1.41	1.76	-0.95 (0.3101)	-1.58	-0.33
	<i>MFG_W</i>	0 (0.032)	-0.07	0.06	0.25 (0.033)	0.18	0.32	0.36 (0.8041)	-1.26	1.99	-1.11 (0.3354)	-1.79	-0.43
Homo 3	<i>UG</i>	0.06 (0.0424)	-0.03	0.14	0.25 (0.0273)	0.19	0.3	-9.98 (7.2422)	-24.62	4.66	-10.36 (1.5728)	-13.54	-7.18
	<i>UFG</i>	0.06 (0.0409)	-0.02	0.14	0.25 (0.0284)	0.2	0.31	-8.01 (5.3392)	-18.8	2.78	-8.15 (1.2217)	-10.62	-5.68
	<i>UFG_A</i>	0.07 (0.0424)	-0.02	0.16	0.25 (0.0278)	0.19	0.3	-5.77 (3.6699)	-13.19	1.65	-4.99 (0.7493)	-6.5	-3.47
	<i>UFG_α</i>	0.07 (0.0421)	-0.02	0.15	0.25 (0.0279)	0.19	0.31	-5.74 (3.6846)	-13.19	1.71	-4.93 (0.7254)	-6.39	-3.46
	<i>UFG_W</i>	0.05 (0.0411)	-0.03	0.14	0.25 (0.0279)	0.2	0.31	-5.19 (3.5838)	-12.44	2.05	-5.28 (0.779)	-6.86	-3.71
	<i>MG</i>	0.05 (0.0393)	-0.03	0.13	0.24 (0.0275)	0.18	0.29	-1.67 (1.2881)	-4.27	0.93	-0.64 (0.285)	-1.22	-0.07
	<i>MFG</i>	0.05 (0.0388)	-0.03	0.12	0.25 (0.0278)	0.19	0.31	-2.61 (1.8266)	-6.3	1.08	-1.61 (0.4095)	-2.44	-0.78
	<i>MFG_A</i>	0.06 (0.0402)	-0.03	0.14	0.24 (0.0271)	0.18	0.29	-0.93 (0.8256)	-2.59	0.74	-0.19 (0.1974)	-0.59	0.2
	<i>MFG_α</i>	0.06 (0.0412)	-0.02	0.15	0.23 (0.0277)	0.18	0.29	-0.87 (0.8494)	-2.58	0.85	-0.06 (0.1807)	-0.42	0.31
	<i>MFG_W</i>	0.06 (0.0395)	-0.02	0.14	0.23 (0.0273)	0.18	0.29	-0.69 (0.7311)	-2.17	0.79	-0.07 (0.1725)	-0.42	0.28
Mix 4	<i>UG</i>	-0.11 (0.0282)	-0.17	-0.06	0.02 (0.0395)	-0.06	0.09	15.79 (3.9848)	7.74	23.85	0.25 (1.7358)	-3.25	3.76

	<i>UFG</i>	-0.01 (0.0393)	-0.09	0.07	0.1 (0.0367)	0.03	0.18	0.85 (1.6887)	-2.57	4.26	-3.3 (1.3043)	-5.93	-0.66
	<i>UFG_A</i>	-0.03 (0.0393)	-0.11	0.05	0.1 (0.0378)	0.02	0.17	1.5 (1.6497)	-1.83	4.83	-2.52 (1.2464)	-5.04	0
	<i>UFG_α</i>	-0.03 (0.0398)	-0.11	0.05	0 (0.0378)	-0.07	0.08	1.76 (1.6262)	-1.53	5.04	0.4 (1.1932)	-2.01	2.82
	<i>UFG_W</i>	-0.03 (0.0399)	-0.11	0.05	0.02 (0.0384)	-0.06	0.1	1.5 (1.6684)	-1.87	4.87	0.08 (1.2447)	-2.43	2.6
	<i>MG</i>	-0.1 (0.0301)	-0.16	-0.04	0.1 (0.0457)	0	0.19	4.14 (1.5105)	1.09	7.19	-0.83 (0.7669)	-2.38	0.72
	<i>MFG</i>	-0.01 (0.028)	-0.07	0.05	0.14 (0.0344)	0.07	0.21	0.98 (0.22)	0.53	1.42	-2.09 (0.798)	-3.7	-0.48
	<i>MFG_A</i>	-0.01 (0.0293)	-0.07	0.05	0.14 (0.0346)	0.07	0.21	0.94 (0.2327)	0.47	1.41	-0.53 (0.5097)	-1.56	0.5
	<i>MFG_α</i>	0 (0.0285)	-0.06	0.06	0.13 (0.0362)	0.06	0.21	0.89 (0.2253)	0.44	1.35	-0.57 (0.483)	-1.54	0.41
	<i>MFG_W</i>	-0.01 (0.0289)	-0.06	0.05	0.12 (0.0322)	0.06	0.19	0.95 (0.2282)	0.49	1.41	-0.43 (0.4419)	-1.33	0.46
	<i>UG</i>	0.13 (0.0385)	0.05	0.21	0.46 (0.0251)	0.41	0.52	-13.2 (3.3973)	-20.07	-6.34	-11.79 (0.8414)	-13.49	-10.1
	<i>UFG</i>	0.14 (0.0446)	0.05	0.23	0.5 (0.0248)	0.45	0.55	-3.66 (1.3202)	-6.33	-0.99	-11.22 (0.8496)	-12.94	-9.5
	<i>UFG_A</i>	0.14 (0.0445)	0.05	0.23	0.49 (0.027)	0.43	0.54	-3.65 (1.273)	-6.22	-1.07	-9.5 (0.7887)	-11.09	-7.9
	<i>UFG_α</i>	0.14 (0.044)	0.05	0.23	0.47 (0.0247)	0.42	0.52	-3.61 (1.2112)	-6.06	-1.17	-9.06 (0.7034)	-10.48	-7.64
	<i>UFG_W</i>	0.14 (0.0437)	0.05	0.22	0.47 (0.0242)	0.42	0.52	-3.55 (1.1934)	-5.96	-1.13	-9.2 (0.6332)	-10.48	-7.92
Mix 5	<i>MG</i>	0.08 (0.038)	0.01	0.16	0.47 (0.0272)	0.41	0.52	-2.09 (1.0384)	-4.19	0.01	-2.23 (0.2822)	-2.8	-1.66
	<i>MFG</i>	0.11 (0.0406)	0.03	0.2	0.43 (0.0289)	0.37	0.48	0.3 (0.2169)	-0.14	0.74	-6.47 (0.6117)	-7.71	-5.24
	<i>MFG_A</i>	0.11 (0.0405)	0.03	0.19	0.44 (0.0276)	0.38	0.49	0.32 (0.2096)	-0.11	0.74	-4.32 (0.4307)	-5.19	-3.45
	<i>MFG_α</i>	0.11 (0.0411)	0.02	0.19	0.42 (0.0288)	0.36	0.48	0.33 (0.2122)	-0.09	0.76	-3.68 (0.3912)	-4.47	-2.89
	<i>MFG_W</i>	0.11 (0.0403)	0.03	0.19	0.43 (0.0282)	0.37	0.48	0.33 (0.2071)	-0.09	0.75	-3.99 (0.4016)	-4.81	-3.18

Table 3.S2 Mean values, lower (LL), and upper limit (UL) of 95% confidence interval of predictive ability ($r_{y\hat{y}}$), accuracy ($r_{g\hat{g}}$), bias (*Bias*), and true breeding value bias (*Bias*_{TBV}) for simulated trait DBH-05 ($h^2 = 0.5$) in all populations (Pop.) and strategies (Str.) (*U* = uni-population model, *M* = multi-population model, *F* = admixed proportions used as fixed effects, *G* = genomic matrix type). Value in parentheses is the standard error (SE).

Pop.	Str.	$r_{y\hat{y}}$			$r_{g\hat{g}}$			<i>Bias</i>			<i>Bias</i> _{TBV}		
		Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL
Homo 1	<i>UG</i>	0.46 (0.0339)	0.39	0.53	0.77 (0.0158)	0.74	0.8	-0.11 (0.1041)	-0.32	0.1	-0.29 (0.0435)	-0.38	-0.2
	<i>UFG</i>	0.47 (0.0329)	0.4	0.53	0.76 (0.0157)	0.73	0.8	-0.11 (0.1014)	-0.31	0.1	-0.27 (0.0441)	-0.36	-0.18
	<i>UFG_A</i>	0.46 (0.033)	0.39	0.52	0.75 (0.0165)	0.71	0.78	-0.11 (0.1056)	-0.32	0.1	-0.28 (0.0488)	-0.38	-0.18
	<i>UFG_α</i>	0.44 (0.0343)	0.37	0.51	0.74 (0.0171)	0.7	0.77	-0.1 (0.1098)	-0.32	0.13	-0.29 (0.0497)	-0.39	-0.19
	<i>UFG_W</i>	0.46 (0.0333)	0.4	0.53	0.76 (0.0157)	0.73	0.8	-0.07 (0.1005)	-0.27	0.13	-0.24 (0.0432)	-0.33	-0.15
	<i>MG</i>	0.47 (0.0327)	0.4	0.54	0.77 (0.0155)	0.74	0.8	0.02 (0.094)	-0.17	0.21	-0.1 (0.0417)	-0.19	-0.02
	<i>MFG</i>	0.46 (0.0326)	0.4	0.53	0.76 (0.0156)	0.73	0.79	0.05 (0.0919)	-0.14	0.23	-0.07 (0.0422)	-0.16	0.01
	<i>MFG_A</i>	0.46 (0.0336)	0.39	0.53	0.76 (0.0167)	0.72	0.79	0.05 (0.095)	-0.15	0.24	-0.08 (0.0427)	-0.17	0.01
	<i>MFG_α</i>	0.46 (0.0325)	0.4	0.53	0.76 (0.0161)	0.73	0.8	0.06 (0.0882)	-0.12	0.23	-0.07 (0.0383)	-0.14	0.01
	<i>MFG_W</i>	0.46 (0.0328)	0.4	0.53	0.76 (0.0169)	0.72	0.79	0.07 (0.0883)	-0.11	0.25	-0.04 (0.0411)	-0.13	0.04
Homo 2	<i>UG</i>	0.43 (0.0272)	0.37	0.48	0.66 (0.0223)	0.62	0.71	0 (0.0659)	-0.13	0.14	-0.05 (0.0444)	-0.14	0.04
	<i>UFG</i>	0.43 (0.0265)	0.38	0.49	0.66 (0.0224)	0.62	0.71	0.01 (0.0633)	-0.12	0.14	-0.04 (0.044)	-0.13	0.05
	<i>UFG_A</i>	0.4 (0.0278)	0.34	0.46	0.65 (0.0227)	0.61	0.7	0.09 (0.0642)	-0.04	0.22	-0.02 (0.0445)	-0.11	0.07
	<i>UFG_α</i>	0.39 (0.0282)	0.34	0.45	0.65 (0.0232)	0.6	0.69	0.08 (0.0678)	-0.06	0.21	-0.03 (0.0463)	-0.13	0.06
	<i>UFG_W</i>	0.43 (0.0267)	0.38	0.49	0.67 (0.022)	0.62	0.71	0.02 (0.0628)	-0.11	0.14	-0.02 (0.0428)	-0.11	0.06
	<i>MG</i>	0.45 (0.0241)	0.4	0.5	0.66 (0.0225)	0.62	0.71	0.07 (0.0565)	-0.04	0.18	0.09 (0.0398)	0.01	0.17
	<i>MFG</i>	0.45 (0.0241)	0.4	0.5	0.66 (0.0224)	0.61	0.71	0.06 (0.0577)	-0.06	0.17	0.07 (0.0403)	-0.01	0.15
	<i>MFG_A</i>	0.44 (0.023)	0.4	0.49	0.66 (0.0229)	0.61	0.7	0.07 (0.0553)	-0.04	0.18	0.07 (0.0401)	-0.01	0.15
	<i>MFG_α</i>	0.44 (0.0239)	0.39	0.49	0.66 (0.023)	0.61	0.7	0.09 (0.0569)	-0.02	0.21	0.09 (0.0396)	0.01	0.17
	<i>MFG_W</i>	0.45 (0.0238)	0.41	0.5	0.66 (0.0237)	0.61	0.7	0.13 (0.0539)	0.02	0.23	0.15 (0.0392)	0.07	0.22
Homo 3	<i>UG</i>	0.49 (0.0308)	0.42	0.55	0.63 (0.0266)	0.58	0.69	-0.16 (0.0838)	-0.33	0.01	-0.04 (0.0562)	-0.15	0.07
	<i>UFG</i>	0.49 (0.0306)	0.43	0.56	0.64 (0.0263)	0.59	0.69	-0.17 (0.0823)	-0.33	0	-0.04 (0.0557)	-0.15	0.07
	<i>UFG_A</i>	0.45 (0.0318)	0.38	0.51	0.6 (0.0287)	0.54	0.66	-0.06 (0.0827)	-0.22	0.11	0.02 (0.0592)	-0.1	0.14
	<i>UFG_α</i>	0.44 (0.0323)	0.38	0.51	0.59 (0.0293)	0.53	0.65	-0.06 (0.086)	-0.24	0.11	0.01 (0.0617)	-0.11	0.14
	<i>UFG_W</i>	0.49 (0.0308)	0.43	0.56	0.64 (0.0266)	0.59	0.69	-0.13 (0.0798)	-0.29	0.03	-0.01 (0.0541)	-0.12	0.1
	<i>MG</i>	0.52 (0.0293)	0.46	0.58	0.66 (0.0234)	0.61	0.7	-0.11 (0.0797)	-0.27	0.05	0.05 (0.0481)	-0.05	0.15
	<i>MFG</i>	0.51 (0.0294)	0.45	0.57	0.65 (0.0226)	0.6	0.7	-0.12 (0.084)	-0.29	0.05	0.05 (0.0488)	-0.05	0.15
	<i>MFG_A</i>	0.52 (0.03)	0.46	0.58	0.65 (0.0224)	0.61	0.7	-0.12 (0.0849)	-0.3	0.05	0.05 (0.047)	-0.05	0.14
	<i>MFG_α</i>	0.52 (0.0308)	0.45	0.58	0.65 (0.0227)	0.61	0.7	-0.12 (0.0886)	-0.3	0.06	0.06 (0.0486)	-0.04	0.16
	<i>MFG_W</i>	0.52 (0.0292)	0.46	0.58	0.66 (0.0225)	0.61	0.7	-0.06 (0.0781)	-0.22	0.1	0.09 (0.0434)	0.01	0.18
Mix 4	<i>UG</i>	0.47 (0.026)	0.41	0.52	0.76 (0.0143)	0.74	0.79	0.13 (0.0716)	-0.01	0.28	0.02 (0.0329)	-0.05	0.09
	<i>UFG</i>	0.48 (0.0257)	0.42	0.53	0.73 (0.0161)	0.7	0.77	0.12 (0.0713)	-0.03	0.26	0.06 (0.0327)	-0.01	0.13
	<i>UFG_A</i>	0.46 (0.0255)	0.41	0.52	0.68 (0.0193)	0.64	0.72	0.08 (0.0732)	-0.07	0.23	0.02 (0.0353)	-0.05	0.09
	<i>UFG_α</i>	0.46 (0.0254)	0.4	0.51	0.68 (0.0199)	0.64	0.72	0.09 (0.0728)	-0.06	0.24	0.01 (0.0358)	-0.07	0.08

	<i>UFG_W</i>	0.47 (0.0255)	0.42	0.52	0.68 (0.0197)	0.64	0.72	0.09 (0.0753)	-0.06	0.24	0.05 (0.035)	-0.02	0.13
	<i>MG</i>	0.45 (0.0266)	0.4	0.51	0.76 (0.0148)	0.73	0.79	0.02 (0.0917)	-0.17	0.2	-0.11 (0.0448)	-0.2	-0.02
	<i>MFG</i>	0.45 (0.0258)	0.39	0.5	0.73 (0.014)	0.7	0.76	0.11 (0.0735)	-0.04	0.26	0.03 (0.0382)	-0.05	0.11
	<i>MFG_A</i>	0.44 (0.0268)	0.39	0.49	0.73 (0.0161)	0.69	0.76	0.12 (0.0742)	-0.03	0.27	-0.02 (0.0422)	-0.11	0.06
	<i>MFG_α</i>	0.44 (0.0262)	0.39	0.49	0.73 (0.0146)	0.7	0.76	0.14 (0.072)	-0.01	0.28	0.04 (0.0401)	-0.04	0.12
	<i>MFG_W</i>	0.45 (0.027)	0.39	0.5	0.74 (0.0148)	0.71	0.77	0.13 (0.0741)	-0.02	0.28	0.04 (0.0404)	-0.04	0.12
	<i>UG</i>	0.45 (0.0335)	0.39	0.52	0.72 (0.0182)	0.68	0.76	0.09 (0.0808)	-0.07	0.25	-0.07 (0.038)	-0.15	0.01
	<i>UFG</i>	0.46 (0.0337)	0.39	0.52	0.69 (0.0184)	0.65	0.73	0.09 (0.0798)	-0.07	0.26	-0.08 (0.0381)	-0.16	0
	<i>UFG_A</i>	0.45 (0.0338)	0.38	0.52	0.65 (0.0202)	0.61	0.69	0.01 (0.0893)	-0.17	0.19	-0.27 (0.0493)	-0.37	-0.17
	<i>UFG_α</i>	0.44 (0.0331)	0.37	0.51	0.63 (0.0205)	0.59	0.68	-0.02 (0.0924)	-0.21	0.16	-0.28 (0.0525)	-0.38	-0.17
Mix 5	<i>UFG_W</i>	0.45 (0.0335)	0.39	0.52	0.67 (0.0193)	0.63	0.71	0.04 (0.0856)	-0.14	0.21	-0.23 (0.0467)	-0.32	-0.13
	<i>MG</i>	0.45 (0.0332)	0.38	0.52	0.73 (0.0178)	0.69	0.76	0.09 (0.0834)	-0.08	0.25	-0.08 (0.0429)	-0.16	0.01
	<i>MFG</i>	0.43 (0.0326)	0.37	0.5	0.61 (0.0225)	0.56	0.65	0.17 (0.0751)	0.02	0.32	-0.25 (0.0622)	-0.38	-0.13
	<i>MFG_A</i>	0.43 (0.0321)	0.36	0.49	0.55 (0.0232)	0.5	0.6	0.19 (0.073)	0.04	0.33	-0.15 (0.0649)	-0.29	-0.02
	<i>MFG_α</i>	0.43 (0.0326)	0.36	0.49	0.61 (0.0215)	0.56	0.65	0.2 (0.074)	0.05	0.34	-0.22 (0.0622)	-0.34	-0.09
	<i>MFG_W</i>	0.43 (0.0324)	0.37	0.5	0.6 (0.0223)	0.56	0.65	0.18 (0.0752)	0.03	0.34	-0.2 (0.0644)	-0.33	-0.07

Table 3.S3 Mean values, lower (LL), and upper limit (UL) of 95% confidence interval of predictive ability (r_{yy}) and bias (*Bias*) for real data traits (flowering time (FT) and panicle length (PL)) in all populations (Pop.) and strategies (Str.) (*U* = uni-population model, *M* = multi-population model, *F* = admixed proportions used as fixed effects, *G* = genomic matrix type). Value in parentheses is the standard error (SE).

Pop.	Str.	FT						PL					
		r_{yy}			<i>Bias</i>			r_{yy}			<i>Bias</i>		
		Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL	Mean (SE)	LL	UL
ADMIX	<i>UG</i>	0.65 (0.033)	0.58	0.72	-0.01 (0.0541)	-0.12	0.1	0.56 (0.0325)	0.49	0.62	-0.09 (0.0693)	-0.22	0.05
	<i>UFG</i>	0.64 (0.0351)	0.57	0.71	0.02 (0.0584)	-0.1	0.13	0.53 (0.0358)	0.46	0.61	-0.04 (0.0762)	-0.19	0.12
	<i>UFG_A</i>	0.63 (0.0353)	0.55	0.7	-0.02 (0.0586)	-0.14	0.1	0.51 (0.0368)	0.44	0.59	-0.08 (0.0892)	-0.26	0.1
	<i>UFG_α</i>	0.48 (0.0403)	0.4	0.56	-0.64 (0.1441)	-0.93	-0.34	0.28 (0.0399)	0.2	0.36	0.18 (0.123)	-0.07	0.43
	<i>UFG_W</i>	0.48 (0.0405)	0.4	0.56	-0.64 (0.1441)	-0.93	-0.34	0.29 (0.0405)	0.21	0.37	0.17 (0.1232)	-0.08	0.42
	<i>MG</i>	0.67 (0.0322)	0.61	0.74	0.07 (0.0573)	-0.04	0.19	0.52 (0.0353)	0.45	0.59	0.19 (0.0623)	0.07	0.32
	<i>MFG</i>	0.65 (0.0351)	0.58	0.73	0.16 (0.0647)	0.03	0.29	0.48 (0.0374)	0.4	0.56	0.31 (0.0632)	0.18	0.44
	<i>MFG_A</i>	0.65 (0.0349)	0.58	0.72	0.18 (0.052)	0.08	0.29	0.45 (0.0398)	0.37	0.53	0.35 (0.0659)	0.22	0.49
	<i>MFG_α</i>	0.3 (0.0409)	0.21	0.38	0.59 (0.0626)	0.46	0.71	0.41 (0.0345)	0.34	0.48	0.29 (0.0783)	0.13	0.45
	<i>MFG_W</i>	0.32 (0.0419)	0.23	0.4	0.56 (0.0622)	0.43	0.69	0.41 (0.0342)	0.34	0.48	0.28 (0.086)	0.11	0.45
AUS	<i>UG</i>	0.46 (0.0332)	0.39	0.52	-0.16 (0.0965)	-0.36	0.03	0.53 (0.0285)	0.47	0.58	-0.26 (0.114)	-0.49	-0.03
	<i>UFG</i>	0.47 (0.0333)	0.4	0.54	-0.19 (0.0998)	-0.39	0.01	0.52 (0.0289)	0.46	0.57	-0.25 (0.1179)	-0.49	-0.01
	<i>UFG_A</i>	0.44 (0.0412)	0.35	0.52	-0.04 (0.1114)	-0.26	0.19	0.49 (0.0309)	0.43	0.55	-0.16 (0.0975)	-0.35	0.04
	<i>UFG_α</i>	0.09 (0.0477)	-0.01	0.18	-4.79 (2.3257)	-9.49	-0.09	-0.02 (0.0463)	-0.11	0.07	0.76 (0.5564)	-0.37	1.88
	<i>UFG_W</i>	0.1 (0.0486)	0	0.19	-5.18 (2.4972)	-10.23	-0.14	-0.03 (0.0435)	-0.12	0.06	0.85 (0.5408)	-0.25	1.94
	<i>MG</i>	0.42 (0.0303)	0.36	0.48	-0.17 (0.1257)	-0.42	0.09	0.45 (0.0303)	0.39	0.52	0.06 (0.0977)	-0.14	0.26
	<i>MFG</i>	0.47 (0.0376)	0.39	0.54	0.09 (0.0997)	-0.11	0.29	0.42 (0.0369)	0.35	0.5	0.2 (0.1237)	-0.05	0.45
	<i>MFG_A</i>	0.42 (0.0374)	0.35	0.5	0.07 (0.1025)	-0.13	0.28	0.39 (0.0354)	0.32	0.46	0.25 (0.0949)	0.06	0.44
	<i>MFG_α</i>	0.13 (0.0528)	0.02	0.24	0.63 (0.1264)	0.37	0.88	0.14 (0.0459)	0.05	0.23	0.6 (0.1413)	0.32	0.89
	<i>MFG_W</i>	0.19 (0.0525)	0.08	0.29	0.38 (0.2079)	-0.04	0.8	0.11 (0.0475)	0.01	0.2	0.74 (0.1633)	0.41	1.07
IND	<i>UG</i>	0.55 (0.0216)	0.51	0.6	-0.01 (0.0744)	-0.16	0.14	0.21 (0.0375)	0.13	0.28	0.4 (0.1105)	0.17	0.62
	<i>UFG</i>	0.55 (0.0214)	0.51	0.59	0.01 (0.0706)	-0.13	0.16	0.18 (0.0379)	0.1	0.26	0.45 (0.1142)	0.22	0.68
	<i>UFG_A</i>	0.55 (0.0224)	0.5	0.59	0.03 (0.0648)	-0.1	0.16	0.2 (0.037)	0.12	0.27	0.38 (0.1173)	0.15	0.62
	<i>UFG_α</i>	-0.33 (0.0341)	-0.4	-0.26	15.62 (2.02)	11.53	19.7	-0.08 (0.0278)	-0.13	-0.02	1.51 (0.2185)	1.07	1.95
	<i>UFG_W</i>	-0.34 (0.0342)	-0.41	-0.27	16.7 (2.0906)	12.48	20.93	-0.09 (0.0276)	-0.14	-0.03	1.68 (0.225)	1.23	2.14
	<i>MG</i>	0.55 (0.0242)	0.5	0.6	0.14 (0.063)	0.01	0.26	0.19 (0.0417)	0.1	0.27	0.51 (0.1145)	0.28	0.74
	<i>MFG</i>	0.52 (0.0263)	0.47	0.57	0.25 (0.0655)	0.12	0.39	0.31 (0.0413)	0.23	0.39	0.41 (0.0828)	0.24	0.58
	<i>MFG_A</i>	0.51 (0.0283)	0.46	0.57	0.26 (0.0583)	0.15	0.38	0.3 (0.0401)	0.22	0.38	0.43 (0.0777)	0.27	0.59
	<i>MFG_α</i>	0.23 (0.04)	0.15	0.31	0.39 (0.1166)	0.16	0.63	0.22 (0.0464)	0.13	0.31	0.62 (0.0906)	0.44	0.8
	<i>MFG_W</i>	0.23 (0.0367)	0.16	0.3	0.46 (0.1052)	0.25	0.68	0.22 (0.0499)	0.12	0.32	0.59 (0.1013)	0.38	0.79
TEJ	<i>UG</i>	0.7 (0.0216)	0.65	0.74	-0.25 (0.0661)	-0.38	-0.11	0.45 (0.0263)	0.39	0.5	-0.38 (0.1121)	-0.61	-0.16
	<i>UFG</i>	0.7 (0.0216)	0.65	0.74	-0.22 (0.0633)	-0.34	-0.09	0.45 (0.0257)	0.4	0.51	-0.35 (0.101)	-0.55	-0.15
	<i>UFG_A</i>	0.69 (0.0207)	0.65	0.73	-0.1 (0.0575)	-0.21	0.02	0.43 (0.0271)	0.38	0.49	-0.01 (0.0771)	-0.17	0.15

	<i>UFG_α</i>	0.08 (0.0385)	0	0.15	0.22 (0.3723)	-0.54	0.97	0.2 (0.0409)	0.12	0.29	0.22 (0.1851)	-0.15	0.59
	<i>UFG_W</i>	0.08 (0.0386)	0	0.16	0.2 (0.3725)	-0.56	0.95	0.2 (0.0419)	0.11	0.28	0.24 (0.1917)	-0.15	0.63
	<i>MG</i>	0.69 (0.0208)	0.65	0.73	0.08 (0.0483)	-0.01	0.18	0.41 (0.026)	0.36	0.47	0.07 (0.0928)	-0.12	0.26
	<i>MFG</i>	0.69 (0.0226)	0.65	0.74	0.08 (0.0523)	-0.03	0.18	0.45 (0.0271)	0.4	0.51	0.12 (0.0783)	-0.04	0.28
	<i>MFG_A</i>	0.7 (0.0211)	0.66	0.74	0.09 (0.0438)	0	0.17	0.46 (0.0297)	0.4	0.52	0.22 (0.0657)	0.09	0.36
	<i>MFG_α</i>	-0.06 (0.0314)	-0.12	0	1.12 (0.1752)	0.77	1.47	0.06 (0.0336)	-0.01	0.13	0.84 (0.1013)	0.63	1.04
	<i>MFG_W</i>	-0.01 (0.0352)	-0.08	0.06	0.86 (0.1882)	0.48	1.24	0.03 (0.0333)	-0.04	0.09	0.91 (0.0984)	0.71	1.11
	<i>UG</i>	0.38 (0.0333)	0.31	0.45	0.17 (0.0828)	0.01	0.34	0.43 (0.028)	0.37	0.48	0.08 (0.0843)	-0.09	0.25
	<i>UFG</i>	0.37 (0.0353)	0.3	0.44	0.2 (0.0867)	0.03	0.38	0.43 (0.0281)	0.37	0.49	0.13 (0.083)	-0.03	0.3
	<i>UFG_A</i>	0.32 (0.0328)	0.25	0.39	0.31 (0.0779)	0.15	0.46	0.42 (0.0279)	0.36	0.47	0.13 (0.0821)	-0.04	0.3
	<i>UFG_α</i>	0.01 (0.0356)	-0.06	0.08	1.13 (0.7648)	-0.41	2.68	0.26 (0.032)	0.2	0.33	-0.18 (0.1826)	-0.55	0.19
TRJ	<i>UFG_W</i>	0.01 (0.0353)	-0.06	0.08	1.13 (0.732)	-0.35	2.61	0.28 (0.0324)	0.21	0.34	-0.26 (0.1961)	-0.66	0.13
	<i>MG</i>	0.34 (0.0416)	0.25	0.42	-0.56 (0.2453)	-1.06	-0.07	0.42 (0.0253)	0.37	0.47	0.18 (0.073)	0.03	0.32
	<i>MFG</i>	0.3 (0.036)	0.23	0.37	0.23 (0.1019)	0.02	0.44	0.37 (0.0262)	0.31	0.42	0.36 (0.063)	0.24	0.49
	<i>MFG_A</i>	0.26 (0.0363)	0.18	0.33	0.31 (0.103)	0.1	0.51	0.36 (0.0258)	0.31	0.42	0.35 (0.063)	0.22	0.47
	<i>MFG_α</i>	0.11 (0.0366)	0.04	0.18	0.61 (0.1685)	0.27	0.95	0.13 (0.0275)	0.07	0.18	0.71 (0.0617)	0.59	0.84
	<i>MFG_W</i>	0.11 (0.0352)	0.04	0.18	0.62 (0.1507)	0.31	0.92	0.14 (0.0285)	0.08	0.19	0.7 (0.0613)	0.57	0.82

Table 3.S4 Mean values, lower (LL), and upper limit (UL) of 95% credible interval of genomic heritability for simulated traits (DBH-01 ($h^2 = 0.1$) and DBH-05 ($h^2 = 0.5$)) in all populations (Pop.) and strategies (Str.) (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type). Value in parentheses is the standard error (SE).

Pop.	Str.	DBH-01				DBH-05			
		Simulated h^2	Mean (SE)	LL	UL	Simulated h^2	Mean (SE)	LL	UL
Homo 1	UG	0.08	0.01 (0.0052)	0	0.02	0.51	0.52 (0.0521)	0.42	0.63
	UFG	0.08	0.01 (0.008)	-0.01	0.02	0.51	0.53 (0.0523)	0.43	0.64
	UFG_A	0.08	0.01 (0.0074)	-0.01	0.02	0.51	0.42 (0.0495)	0.32	0.52
	UFG_α	0.08	0.01 (0.0073)	-0.01	0.02	0.51	0.4 (0.0495)	0.3	0.49
	UFG_W	0.08	0.01 (0.0078)	-0.01	0.02	0.51	0.47 (0.0526)	0.36	0.57
	MG	0.08	0.05 (0.0532)	-0.06	0.15	0.51	0.66 (0.1027)	0.46	0.87
	MFG	0.08	0.03 (0.039)	-0.05	0.11	0.51	0.67 (0.103)	0.47	0.87
	MFG_A	0.08	0.05 (0.0546)	-0.06	0.16	0.51	0.6 (0.1111)	0.38	0.82
	MFG_α	0.08	0.06 (0.0638)	-0.06	0.19	0.51	0.61 (0.1114)	0.39	0.83
	MFG_W	0.08	0.05 (0.0597)	-0.06	0.17	0.51	0.64 (0.1122)	0.42	0.86
Homo 2	UG	0.10	0.01 (0.0052)	0	0.02	0.45	0.52 (0.0521)	0.42	0.63
	UFG	0.10	0.01 (0.008)	-0.01	0.02	0.45	0.53 (0.0523)	0.43	0.64
	UFG_A	0.10	0.01 (0.0074)	-0.01	0.02	0.45	0.42 (0.0495)	0.32	0.52
	UFG_α	0.10	0.01 (0.0073)	-0.01	0.02	0.45	0.4 (0.0495)	0.3	0.49
	UFG_W	0.10	0.01 (0.0078)	-0.01	0.02	0.45	0.47 (0.0526)	0.36	0.57
	MG	0.10	0.04 (0.0385)	-0.04	0.11	0.45	0.68 (0.1205)	0.44	0.92
	MFG	0.10	0.02 (0.0285)	-0.03	0.08	0.45	0.67 (0.1231)	0.42	0.91
	MFG_A	0.10	0.03 (0.035)	-0.03	0.1	0.45	0.59 (0.132)	0.34	0.85
	MFG_α	0.10	0.04 (0.0389)	-0.04	0.12	0.45	0.61 (0.1305)	0.36	0.87
	MFG_W	0.10	0.04 (0.0361)	-0.04	0.11	0.45	0.67 (0.1257)	0.42	0.91
Homo 3	UG	0.10	0.01 (0.0052)	0	0.02	0.46	0.52 (0.0521)	0.42	0.63
	UFG	0.10	0.01 (0.008)	-0.01	0.02	0.46	0.53 (0.0523)	0.43	0.64
	UFG_A	0.10	0.01 (0.0074)	-0.01	0.02	0.46	0.42 (0.0495)	0.32	0.52
	UFG_α	0.10	0.01 (0.0073)	-0.01	0.02	0.46	0.4 (0.0495)	0.3	0.49
	UFG_W	0.10	0.01 (0.0078)	-0.01	0.02	0.46	0.47 (0.0526)	0.36	0.57
	MG	0.10	0.06 (0.0684)	-0.08	0.19	0.46	0.65 (0.1061)	0.44	0.86
	MFG	0.10	0.03 (0.0472)	-0.06	0.13	0.46	0.64 (0.1072)	0.43	0.85
	MFG_A	0.10	0.06 (0.0652)	-0.07	0.19	0.46	0.57 (0.1178)	0.33	0.8
	MFG_α	0.10	0.07 (0.0701)	-0.07	0.2	0.46	0.58 (0.1165)	0.35	0.81
	MFG_W	0.10	0.07 (0.0741)	-0.08	0.21	0.46	0.61 (0.1133)	0.39	0.83
Mix 4	UG	0.10	0.01 (0.0052)	0	0.02	0.50	0.52 (0.0521)	0.42	0.63
	UFG	0.10	0.01 (0.008)	-0.01	0.02	0.50	0.53 (0.0523)	0.43	0.64
	UFG_A	0.10	0.01 (0.0074)	-0.01	0.02	0.50	0.42 (0.0495)	0.32	0.52
	UFG_α	0.10	0.01 (0.0073)	-0.01	0.02	0.50	0.4 (0.0495)	0.3	0.49
	UFG_W	0.10	0.01 (0.0078)	-0.01	0.02	0.50	0.47 (0.0526)	0.36	0.57
	MG	0.10	0.02 (0.0232)	-0.02	0.07	0.50	0.46 (0.1136)	0.23	0.68
	MFG	0.10	0.02 (0.021)	-0.02	0.06	0.50	0.49 (0.1123)	0.27	0.71
	MFG_A	0.10	0.03 (0.0357)	-0.04	0.1	0.50	0.48 (0.1119)	0.26	0.7
	MFG_α	0.10	0.03 (0.0414)	-0.05	0.12	0.50	0.5 (0.1153)	0.28	0.73
	MFG_W	0.10	0.03 (0.0375)	-0.04	0.1	0.50	0.52 (0.1174)	0.29	0.75
Mix 5	UG	0.11	0.01 (0.0052)	0	0.02	0.57	0.52 (0.0521)	0.42	0.63
	UFG	0.11	0.01 (0.008)	-0.01	0.02	0.57	0.53 (0.0523)	0.43	0.64
	UFG_A	0.11	0.01 (0.0074)	-0.01	0.02	0.57	0.42 (0.0495)	0.32	0.52
	UFG_α	0.11	0.01 (0.0073)	-0.01	0.02	0.57	0.4 (0.0495)	0.3	0.49
	UFG_W	0.11	0.01 (0.0078)	-0.01	0.02	0.57	0.47 (0.0526)	0.36	0.57
	MG	0.11	0.04 (0.0366)	-0.04	0.11	0.57	0.55 (0.1068)	0.34	0.76
	MFG	0.11	0.02 (0.0211)	-0.02	0.06	0.57	0.44 (0.1292)	0.19	0.7
	MFG_A	0.11	0.03 (0.0287)	-0.03	0.08	0.57	0.45 (0.1287)	0.2	0.7
	MFG_α	0.11	0.03 (0.0313)	-0.03	0.09	0.57	0.46 (0.1261)	0.22	0.71
	MFG_W	0.11	0.03 (0.0295)	-0.03	0.08	0.57	0.48 (0.1312)	0.22	0.74

Table 3.S5 Mean values, lower (LL), and upper limit (UL) of 95% credible interval of genomic heritability for real data traits (flowering time (FT) and panicle length (PL)) in all populations (Pop.) and strategies (Str.) (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type). Value in parentheses is the standard error (SE).

Pop.	Str.	FT			PL		
		Mean (SE)	LL	UL	Mean (SE)	LL	UL
ADMIX	UG	0.7 (0.0675)	0.56	0.83	0.67 (0.0829)	0.51	0.84
	UFG	0.7 (0.0677)	0.57	0.83	0.67 (0.0897)	0.49	0.84
	UFG_A	0.51 (0.0812)	0.35	0.67	0.48 (0.1056)	0.27	0.68
	UFG_α	0.01 (0.0057)	0	0.02	0.05 (0.0228)	0.01	0.1
	UFG_W	0.01 (0.006)	0	0.02	0.06 (0.0248)	0.01	0.11
	MG	0.97 (0.0248)	0.92	1.02	0.85 (0.0916)	0.67	1.03
	MFG	0.98 (0.0202)	0.94	1.02	0.87 (0.0866)	0.7	1.04
	MFG_A	0.95 (0.0402)	0.87	1.03	0.82 (0.1055)	0.61	1.02
	MFG_α	0.71 (0.255)	0.21	1.21	0.62 (0.2311)	0.17	1.08
	MFG_W	0.76 (0.2431)	0.28	1.23	0.64 (0.2224)	0.2	1.07
AUS	UG	0.7 (0.0675)	0.56	0.83	0.67 (0.0829)	0.51	0.84
	UFG	0.7 (0.0677)	0.57	0.83	0.67 (0.0897)	0.49	0.84
	UFG_A	0.51 (0.0812)	0.35	0.67	0.48 (0.1056)	0.27	0.68
	UFG_α	0.01 (0.0057)	0	0.02	0.05 (0.0228)	0.01	0.1
	UFG_W	0.01 (0.006)	0	0.02	0.06 (0.0248)	0.01	0.11
	MG	0.76 (0.1338)	0.49	1.02	0.83 (0.1307)	0.58	1.09
	MFG	0.76 (0.1275)	0.51	1.01	0.84 (0.1326)	0.58	1.1
	MFG_A	0.65 (0.152)	0.36	0.95	0.77 (0.143)	0.49	1.05
	MFG_α	0.48 (0.2158)	0.06	0.9	0.67 (0.2139)	0.25	1.09
	MFG_W	0.53 (0.2012)	0.13	0.92	0.67 (0.2116)	0.26	1.09
IND	UG	0.7 (0.0675)	0.56	0.83	0.67 (0.0829)	0.51	0.84
	UFG	0.7 (0.0677)	0.57	0.83	0.67 (0.0897)	0.49	0.84
	UFG_A	0.51 (0.0812)	0.35	0.67	0.48 (0.1056)	0.27	0.68
	UFG_α	0.01 (0.0057)	0	0.02	0.05 (0.0228)	0.01	0.1
	UFG_W	0.01 (0.006)	0	0.02	0.06 (0.0248)	0.01	0.11
	MG	0.97 (0.019)	0.93	1.01	0.82 (0.1298)	0.56	1.07
	MFG	0.97 (0.0181)	0.94	1.01	0.81 (0.1262)	0.56	1.06
	MFG_A	0.95 (0.0335)	0.88	1.01	0.76 (0.131)	0.5	1.02
	MFG_α	0.13 (0.1194)	-0.11	0.36	0.75 (0.1616)	0.43	1.06
	MFG_W	0.18 (0.1669)	-0.14	0.51	0.76 (0.1518)	0.46	1.06
TEJ	UG	0.7 (0.0675)	0.56	0.83	0.67 (0.0829)	0.51	0.84
	UFG	0.7 (0.0677)	0.57	0.83	0.67 (0.0897)	0.49	0.84
	UFG_A	0.51 (0.0812)	0.35	0.67	0.48 (0.1056)	0.27	0.68
	UFG_α	0.01 (0.0057)	0	0.02	0.05 (0.0228)	0.01	0.1
	UFG_W	0.01 (0.006)	0	0.02	0.06 (0.0248)	0.01	0.11
	MG	0.95 (0.0351)	0.88	1.02	0.87 (0.12)	0.63	1.1
	MFG	0.94 (0.04)	0.86	1.02	0.85 (0.1317)	0.59	1.11
	MFG_A	0.82 (0.0806)	0.67	0.98	0.8 (0.133)	0.54	1.06
	MFG_α	0.09 (0.0974)	-0.1	0.28	0.27 (0.15)	-0.03	0.56
	MFG_W	0.1 (0.1091)	-0.11	0.31	0.29 (0.1618)	-0.03	0.6
TRJ	UG	0.7 (0.0675)	0.56	0.83	0.67 (0.0829)	0.51	0.84
	UFG	0.7 (0.0677)	0.57	0.83	0.67 (0.0897)	0.49	0.84
	UFG_A	0.51 (0.0812)	0.35	0.67	0.48 (0.1056)	0.27	0.68
	UFG_α	0.01 (0.0057)	0	0.02	0.05 (0.0228)	0.01	0.1
	UFG_W	0.01 (0.006)	0	0.02	0.06 (0.0248)	0.01	0.11
	MG	0.42 (0.1543)	0.12	0.72	0.73 (0.1059)	0.52	0.94
	MFG	0.44 (0.1545)	0.14	0.75	0.67 (0.138)	0.4	0.94
	MFG_A	0.26 (0.1295)	0.01	0.51	0.51 (0.1441)	0.22	0.79
	MFG_α	0.1 (0.1009)	-0.09	0.3	0.32 (0.1642)	0	0.64
	MFG_W	0.13 (0.1241)	-0.11	0.38	0.44 (0.204)	0.04	0.84

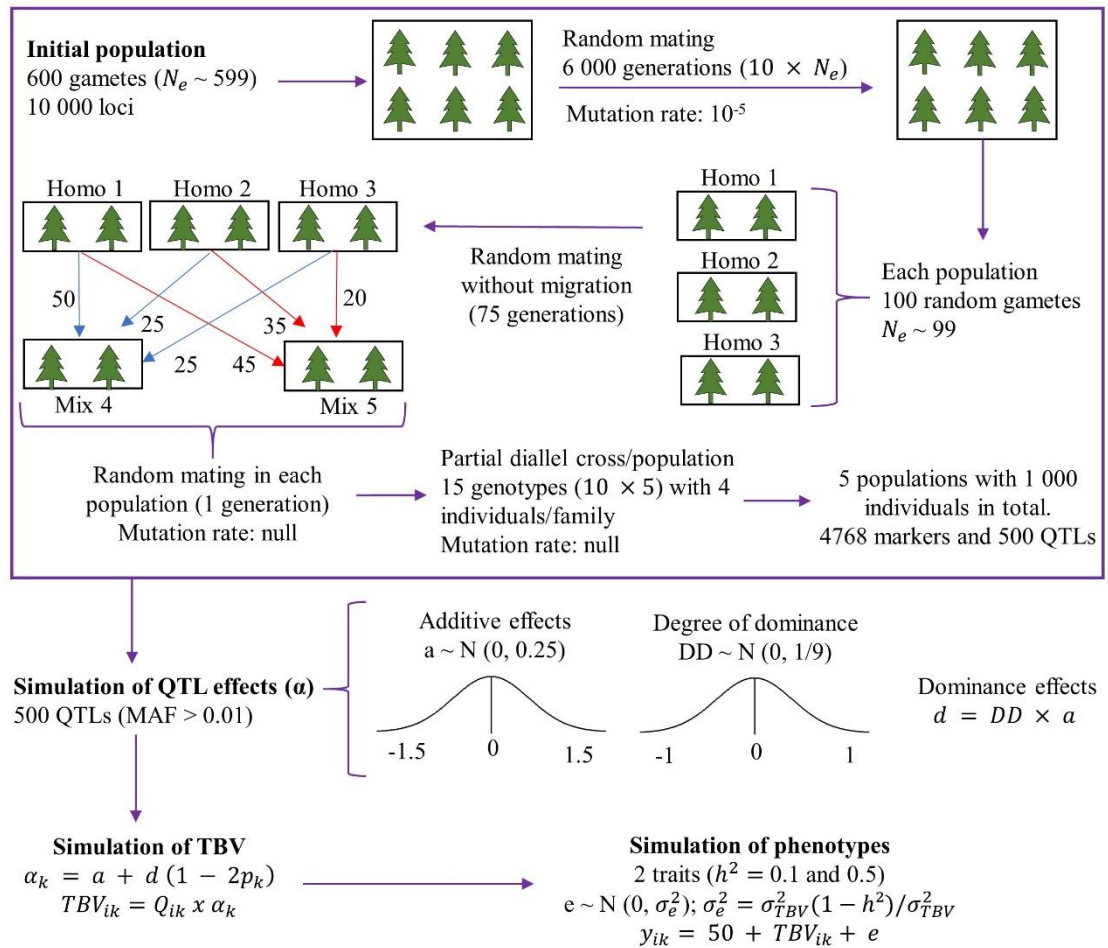


Fig. 3.S1 Simulation procedure scheme. α_k is the average effect of allelic substitution vector for quantitative trait loci (QTLs) of the k^{th} population, p_k is the QTL allele frequency vector of the k^{th} population, TBV_{ik} is the true breeding value (TBV) of the i^{th} individual in the k^{th} population, Q_{ik} is the QTL number per locus vector of the i^{th} individual in the k^{th} population, e is the residual effect, σ_{TBV}^2 is the variance of TBV in each population, and y_{ik} is the phenotype of the i^{th} individual in the k^{th} population.

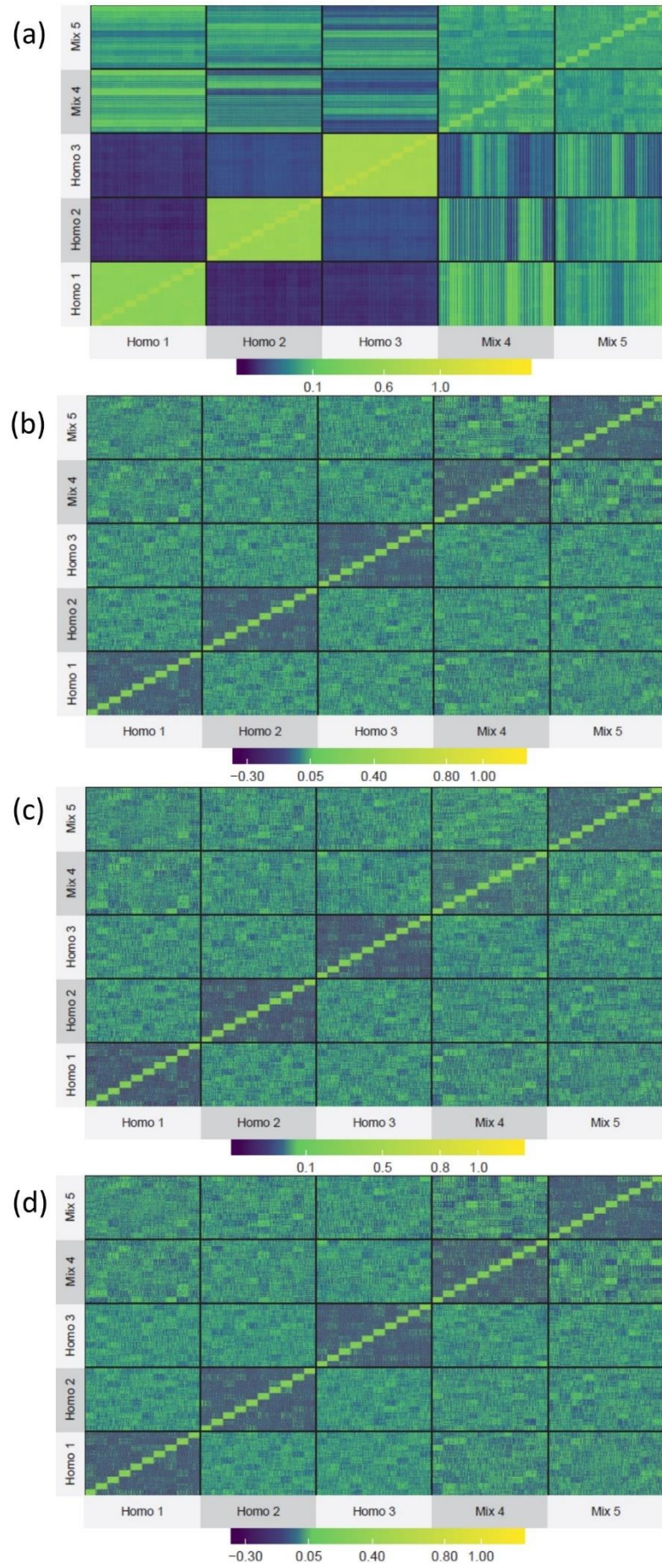


Fig. 3.S2 Heatmap of G (a), G_α (b), G_A (c), and G_W matrix (d) for simulated data.

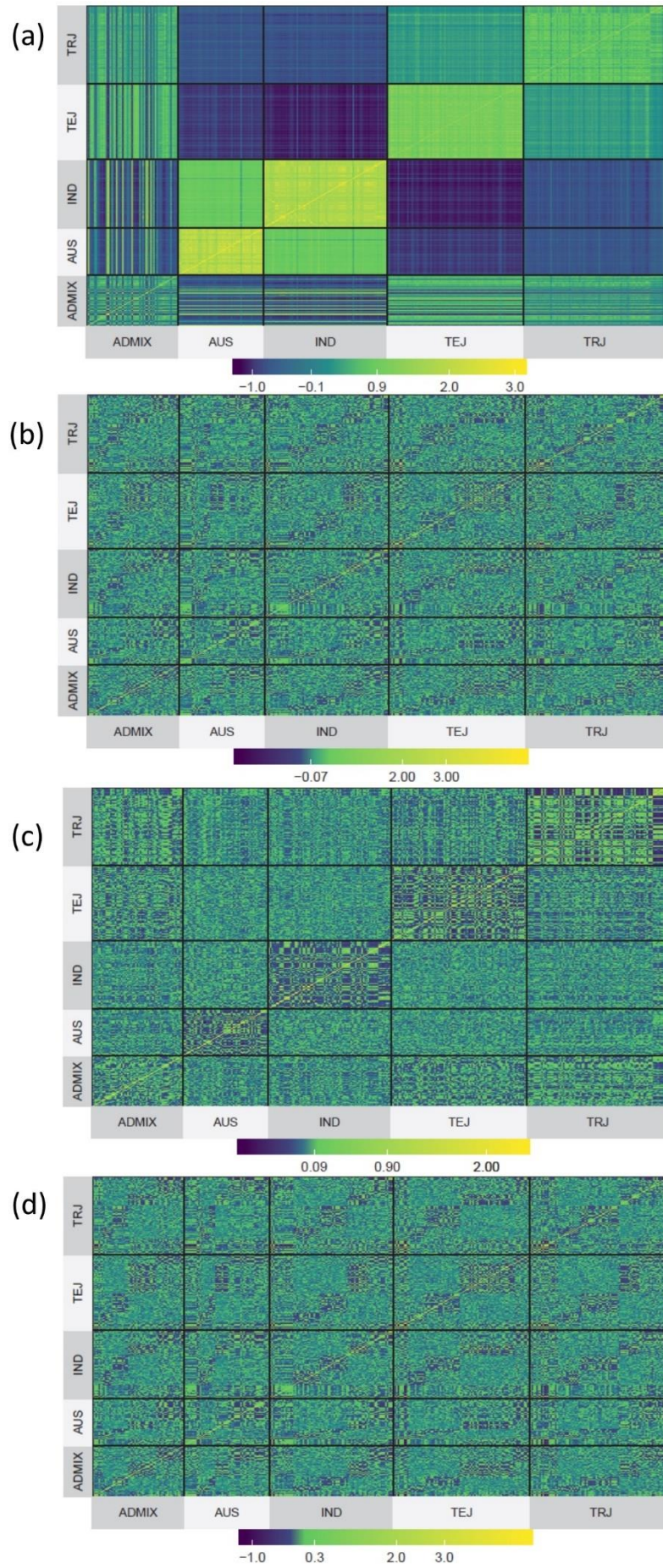


Fig. 3.S3 Heatmap of G (a), G_α (b), G_A (c), and G_W matrix (d) for real data.

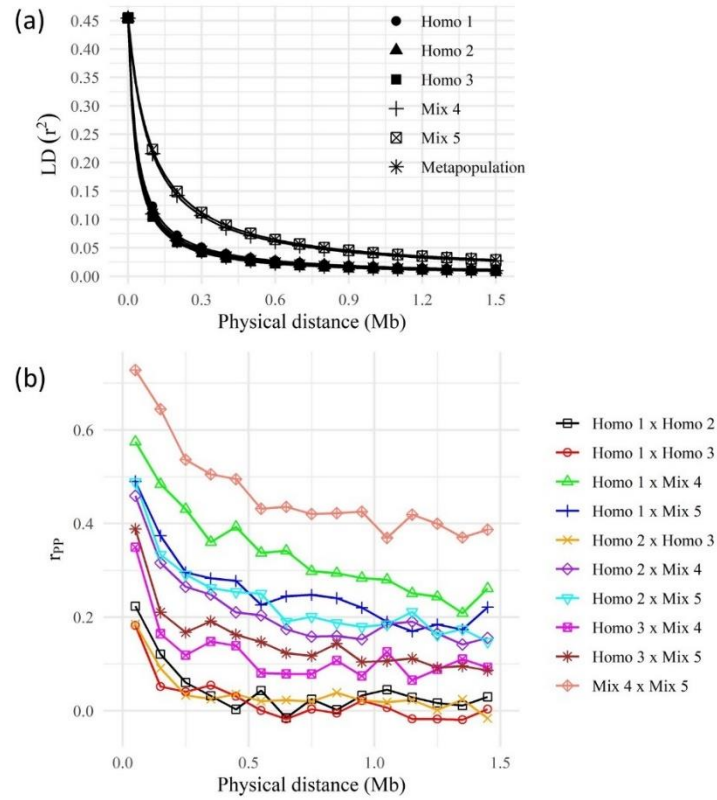


Fig. 3.S4 Linkage disequilibrium (r^2) (a) and linkage disequilibrium phase persistence (r_{PP}) between populations (b) as a function of physical distance (Mb) for simulated data.

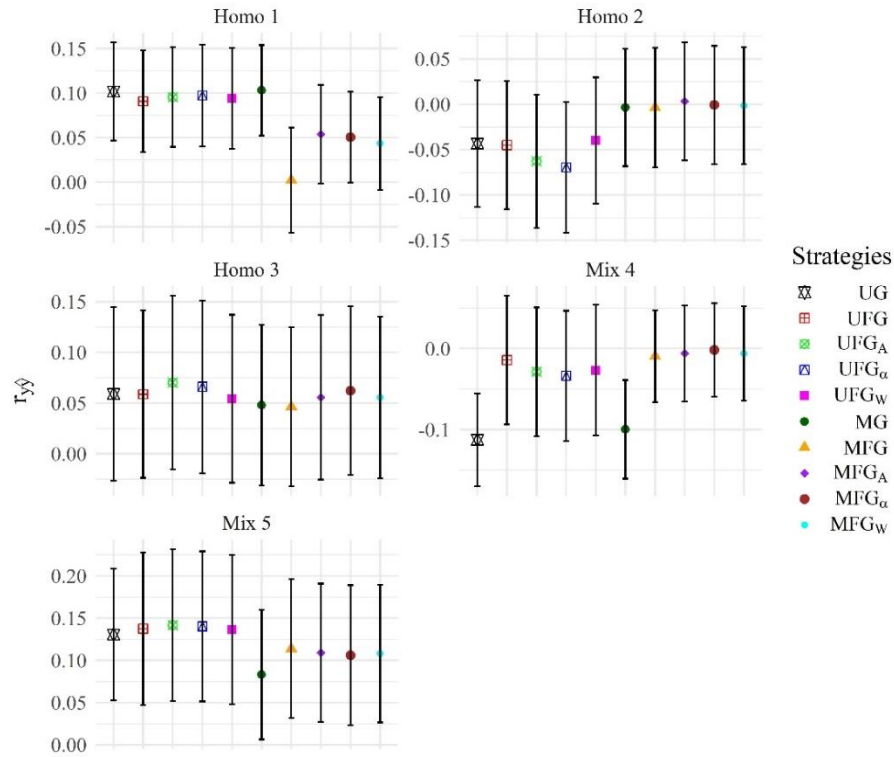


Fig. 3.S5 Predictive ability ($r_{\hat{y}y}$) of DBH-01 ($h^2 = 0.1$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

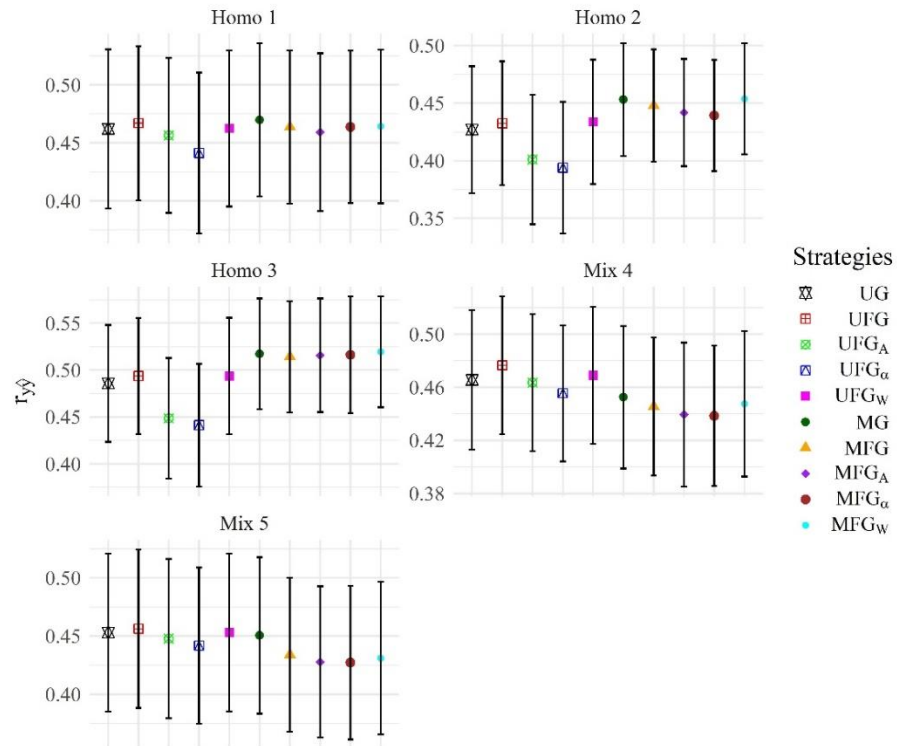


Fig. 3.S6 Predictive ability ($r_{y\hat{y}}$) of DBH-05 ($h^2 = 0.5$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

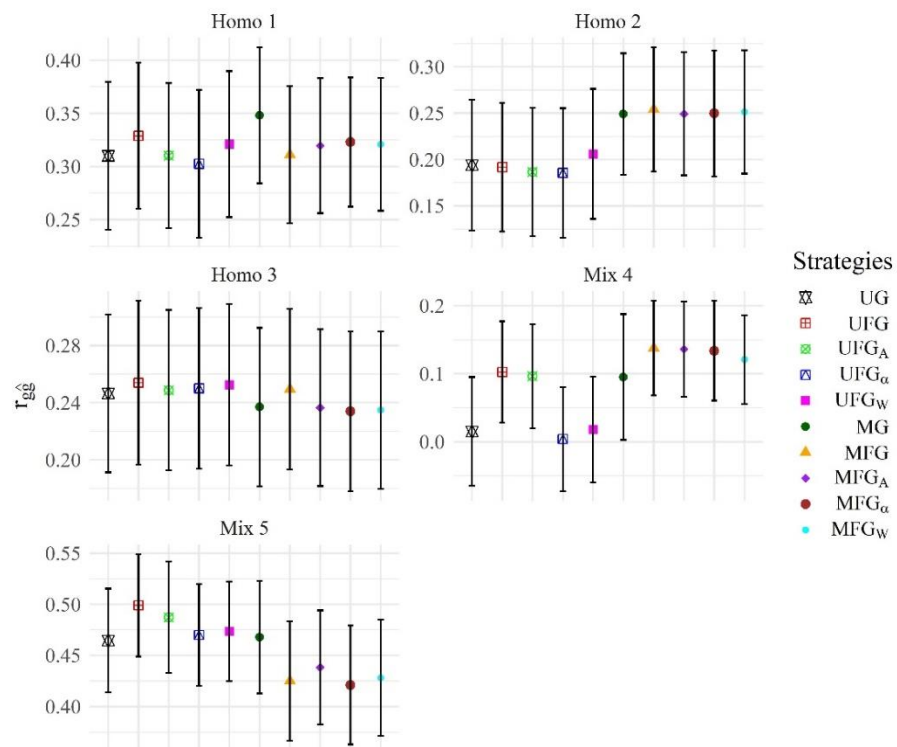


Fig. 3.S7 Accuracy ($r_{g\hat{g}}$) of DBH-01 ($h^2 = 0.1$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

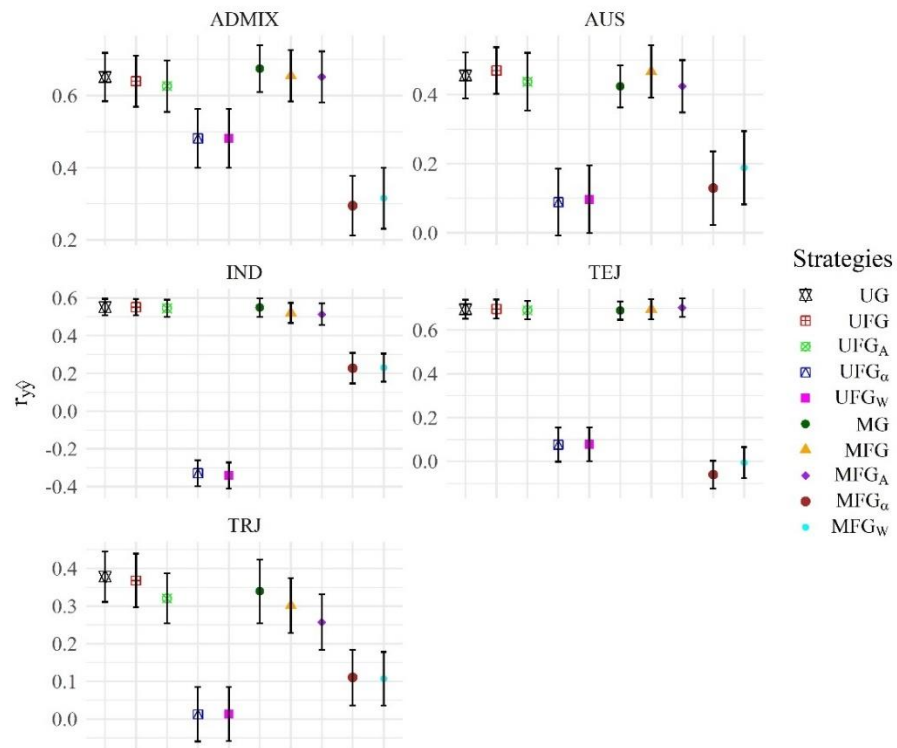


Fig. 3.S8 Predictive ability (r_{yy}) of flowering time in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% confidence intervals.

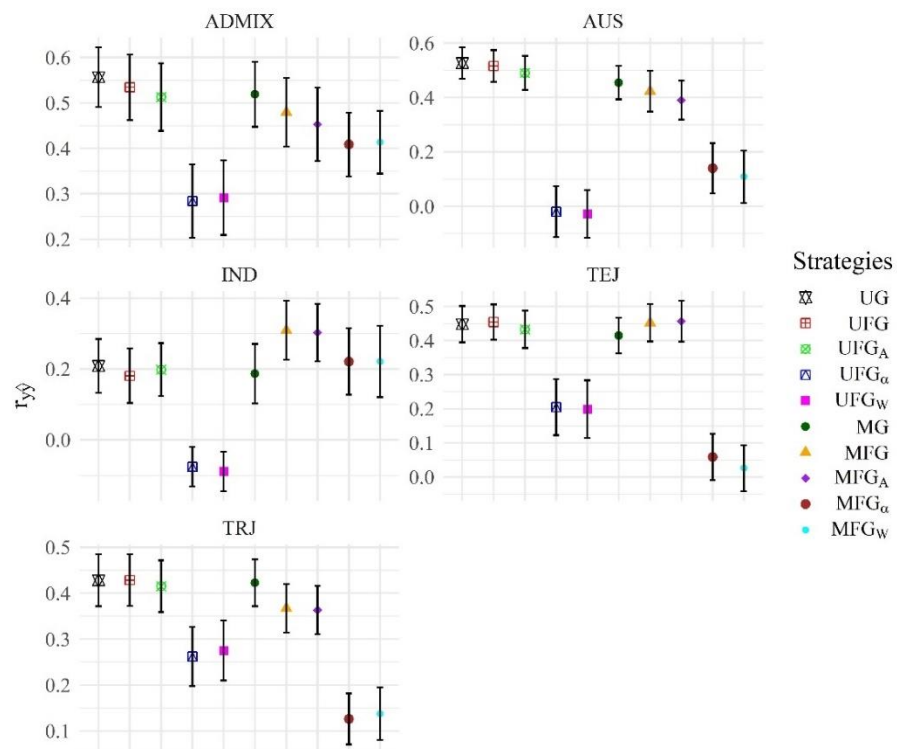


Fig. 3.S9 Predictive ability (r_{yy}) of panicle length in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% confidence intervals.

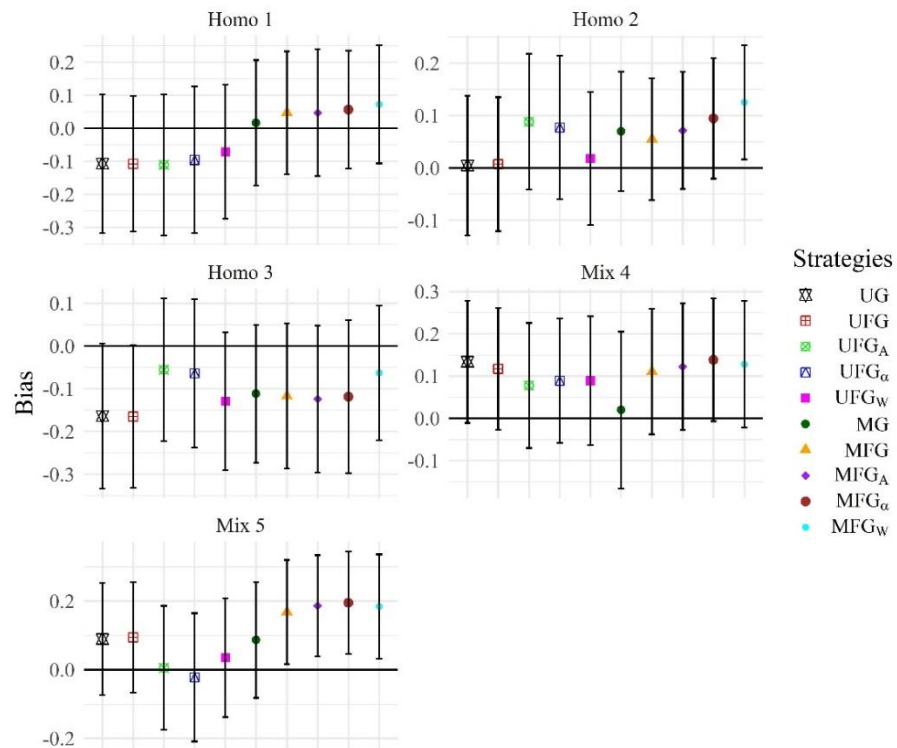


Fig. 3.S10 Bias of DBH-05 ($h^2 = 0.5$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

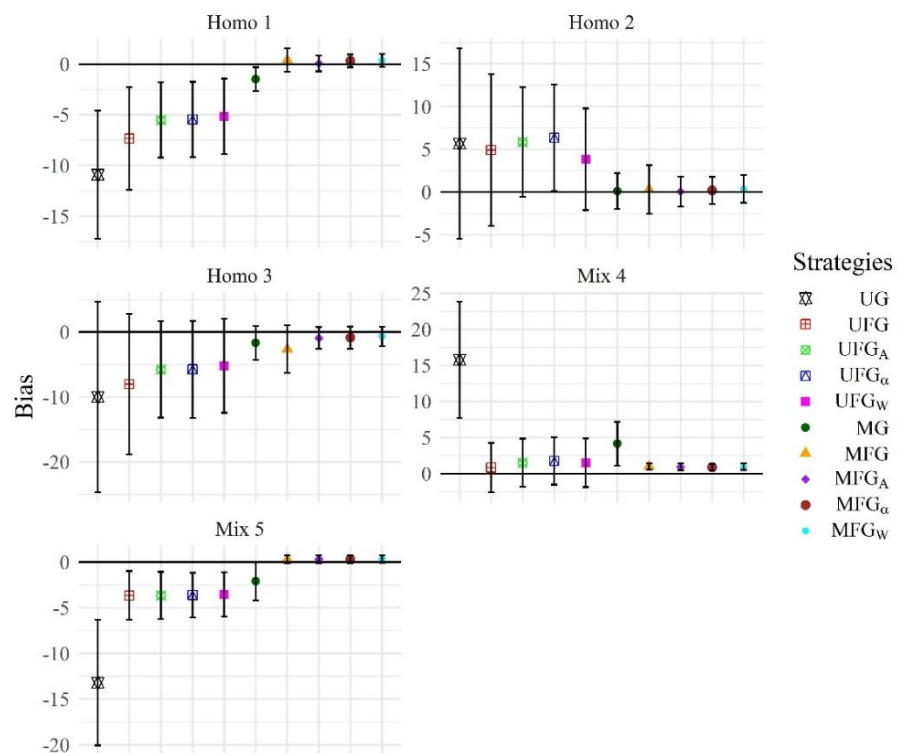


Fig. 3.S11 Bias of DBH-01 ($h^2 = 0.1$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

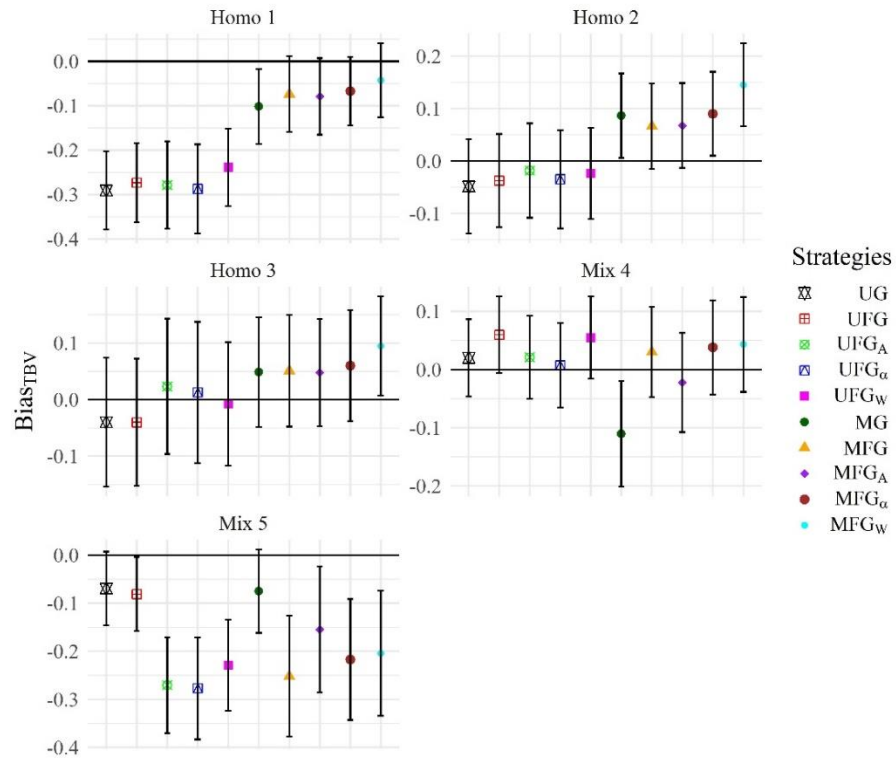


Fig. 3.S12 True breeding value bias ($Bias_{TBV}$) of DBH-05 ($h^2 = 0.5$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = ancestry proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% confidence intervals.

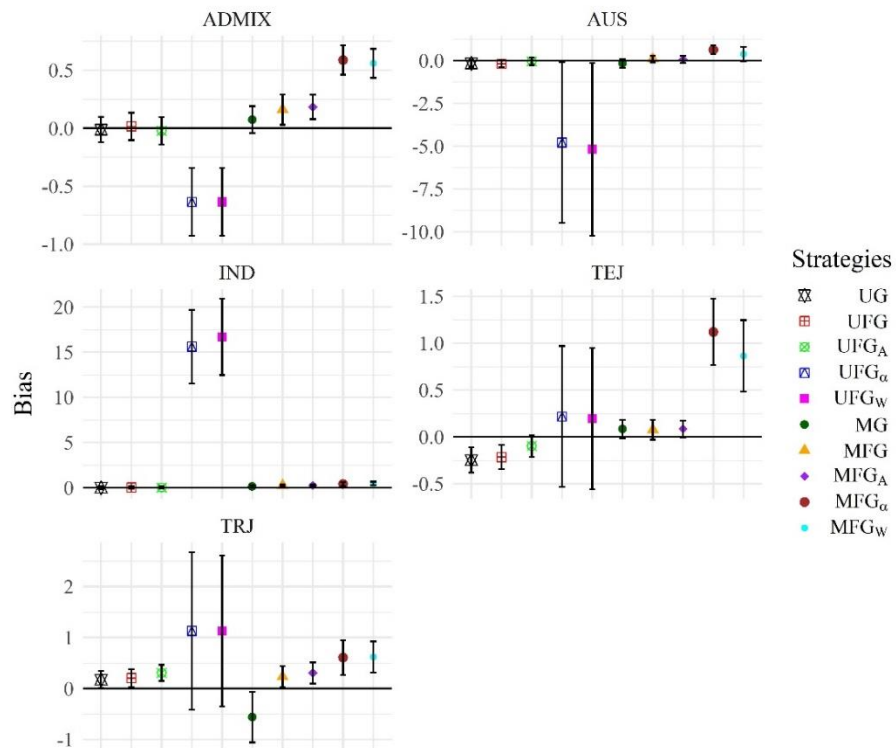


Fig. 3.S13 Bias of flowering time in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% confidence intervals.

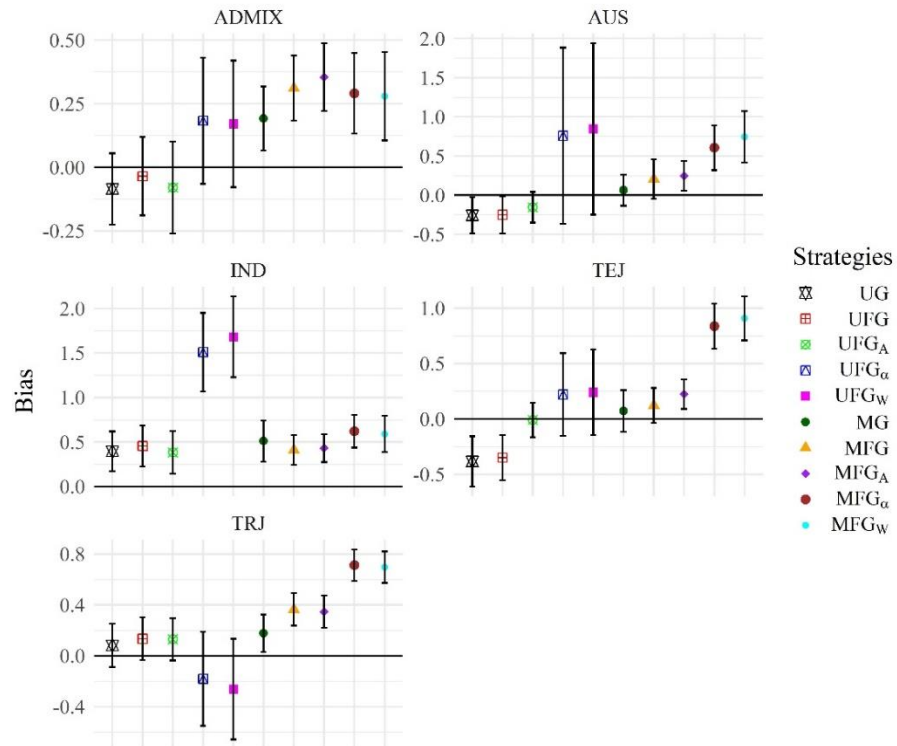


Fig. 3.S14 Bias of panicle length in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% confidence intervals.

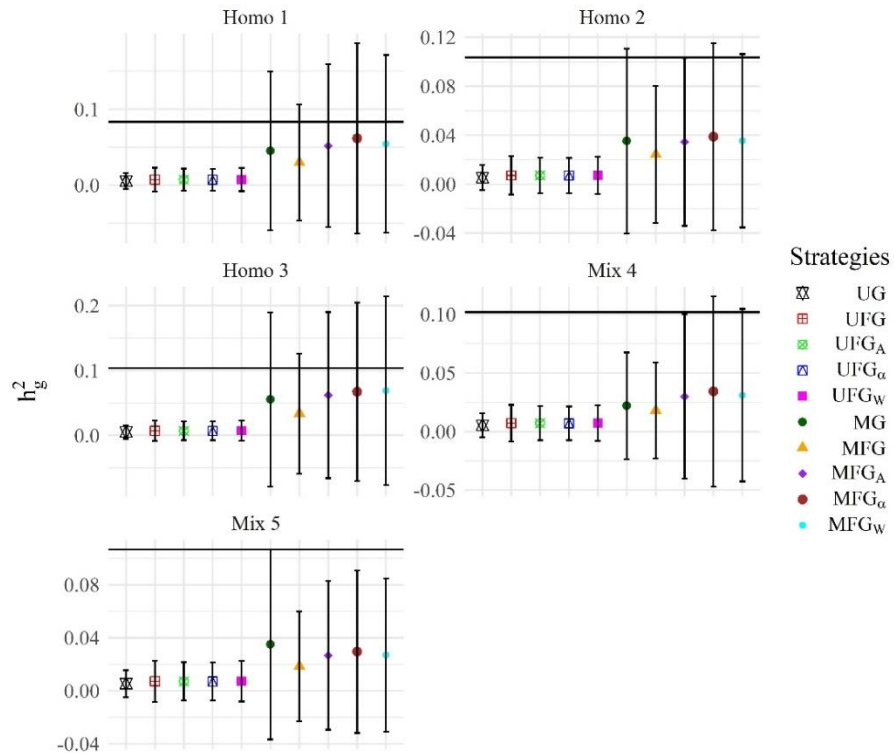


Fig. 3.S15 Genomic heritability (h_g^2) of DBH-01 ($h^2 = 0.1$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = ancestry proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% credible intervals. The horizontal line is the simulated heritability.

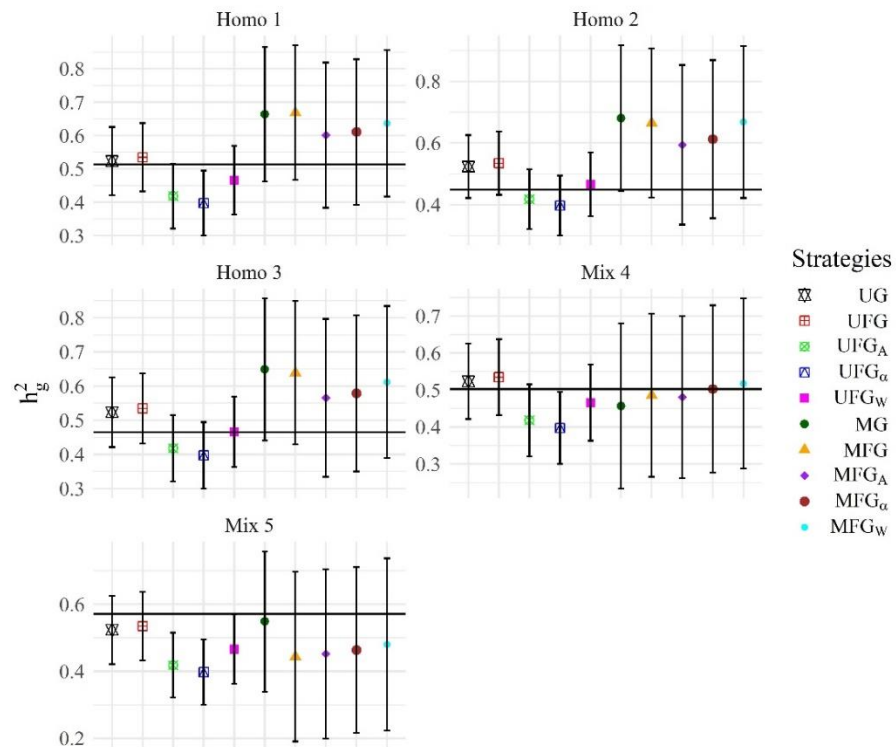


Fig. 3.S16 Genomic heritability (h_g^2) of DBH-05 ($h^2 = 0.5$) in homogeneous (Homo 1, 2, and 3) and admixed populations (Mix 4 and 5) for all strategies (U = uni-population model, M = multi-population model, F = ancestry proportions used as fixed effects, G = genomic matrix type) with simulated data. The bars are the 95% credible intervals. The horizontal line is the simulated heritability.

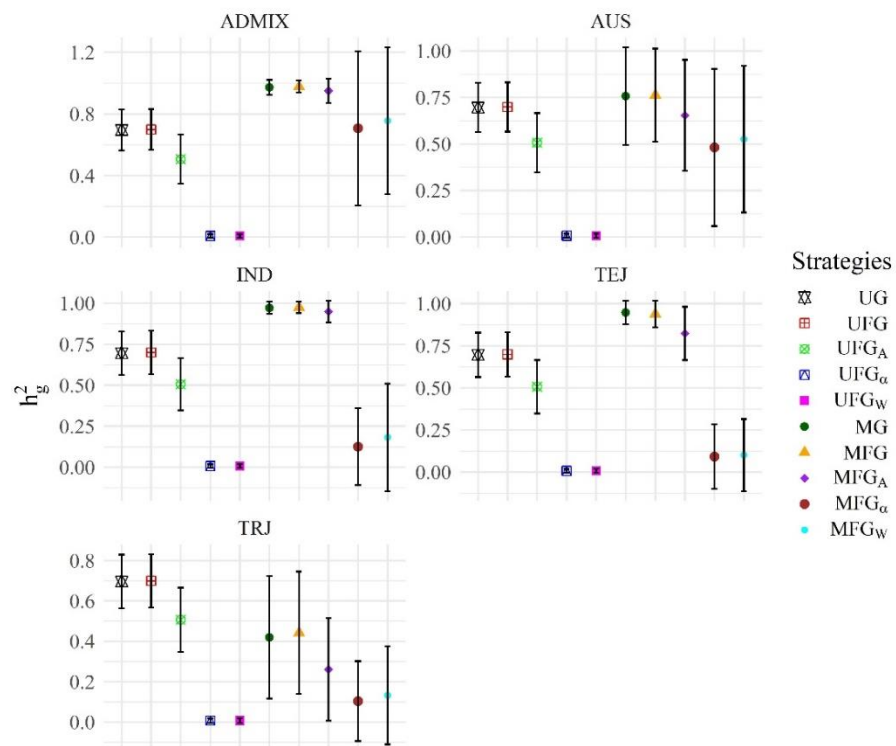


Fig. 3.S17 Genomic heritability (h_g^2) of flowering time in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% credible intervals.

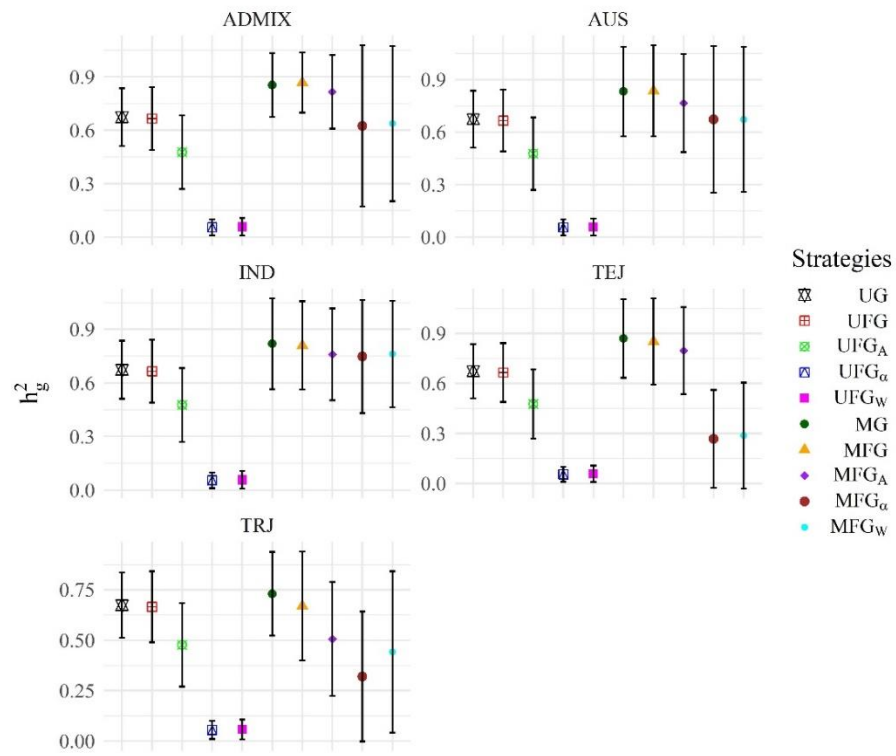


Fig. 3.S18 Genomic heritability (h_g^2) of panicle length in all populations (ADMIX, AUS, IND, TEJ, and TRJ) for all strategies (U = uni-population model, M = multi-population model, F = admixed proportions used as fixed effects, G = genomic matrix type) with real data. The bars are the 95% credible intervals.

4. CHAPTER 2: Entropy and mutual information in genome-wide selection: the splitting of k-fold cross-validation sets and implications for tree breeding

This chapter was published in *Tree Genetics & Genomes* (2020 – DOI: 10.1007/s11295-020-01430-6).

Abstract

Random k-fold cross-validation in genome-wide selection (GWS) can help to estimate predictive ability ($r_{y\hat{y}}$). Predictive ability tends to be higher when training, and validation sets present a high degree of kinship. However, many tree breeding populations are less genetically related to the training sets and have different levels of phenotypic diversity. Therefore, this study proposes methods of splitting k-fold cross-validation sets to optimize $r_{y\hat{y}}$ estimates that are consistent with the breeding population and verify the impact of phenotypic and genotypic distribution on GWS. Using a simulated *Eucalyptus* trait ($h^2 = 0.5$) and *Pinus taeda* L. data for diameter at breast height ($h^2 = 0.31$), six methods were developed based on mutual information (I) and entropy (H) for measuring genetic similarity and phenotypic dissimilarity, respectively. All methods were evaluated for $r_{y\hat{y}}$, bias, minimum squared error of prediction, and genomic heritability. The Pearson correlations of these parameters with the kinship coefficient, and I and H between and within training and validation sets were also estimated. Our results show that closer genetic similarity did not significantly increase $r_{y\hat{y}}$ and that a lower H reduced $r_{y\hat{y}}$ and overestimated genomic breeding values. Consequently, phenotypic diversity (high H) should be added to tree breeding populations to increase genetic gain and reduce bias. The new methods accurately fitted models according to the entropy of tree breeding populations and their genetic relationship to the training sets. Therefore, these methods provided usable estimates of genetic gain to produce consistent success of long-term tree breeding programs.

Keywords: Entropy, mutual information, *Pinus*, *Eucalyptus*, genetic gain, genome wide selection.

4.1 Introduction

Genome-wide selection (GWS) has exhibited high genetic gain per unit time for certain traits and has been thoroughly evaluated for its use in genetic breeding of animals (Daetwyler et al. 2012; Hulsman Hanna et al. 2015; Silva et al. 2016), crops (Lehermeier

et al. 2015), and forestry species (Resende et al. 2012, 2017a, b). GWS allows the estimation of breeding values (GEBVs, hereafter), which express the genetic potential of individuals considering their markers. These GEBVs are estimated using statistical models, like explicit regression, that explore the relationship between markers and phenotype. This relationship is due to the existence of linkage disequilibrium between causal loci (i.e., quantitative trait loci - QTLs) with a least some of the markers used in the analysis (Meuwissen et al. 2001).

To verify a model's predictive ability, k-fold cross-validation can be performed, also allowing the estimation of prediction accuracy and genetic gain. For this, the data is divided into k sets of the same size. The model is fitted to $k - 1$ sets (training set) for each of the k sets. The model obtained is then applied to the validation set, and $r_{y\hat{y}}$ can be obtained.

The random assignment of individuals in k-fold cross-validation can result in high genetic affinity between the training and validation sets, and consequently high $r_{y\hat{y}}$ values (Habier et al. 2010; Daetwyler et al. 2013; Wray et al. 2013). These high values can be due to the high degree of relationship shared between sets (Clark et al. 2012; Pszczola et al. 2012), shared relatives (Pérez-Cabal et al. 2012), and a weak population structure effect generated between sets (low Wright fixation index - F_{ST}) (Scutari et al. 2016). In this scenario, the use of models based on k-fold cross-validation can overestimate $r_{y\hat{y}}$ when applied to tree breeding populations that are less related to the training sets used.

Many strategies for composing training sets have been proposed to produce more accurate $r_{y\hat{y}}$. These strategies split the training and validation sets according to family (Pszczola et al. 2012; Hulsman Hanna et al. 2015; Resende et al. 2017a), population structure (Guo et al. 2014), generation (Pérez-Cabal et al. 2012; Pszczola et al. 2012; Saatchi et al. 2013; Silva et al. 2016), maximum kinship coefficient (Habier et al. 2010), Wright kinship coefficients (Saatchi et al. 2011, 2013; Clark et al. 2012; Pérez-Cabal et al. 2012; Boddhireddy et al. 2014), identity by state clustering methods (Boddhireddy et al. 2014), and unrelated individuals (Pérez-Cabal et al. 2012; Silva et al. 2016). Tree breeding populations are normally composed of a highly genetically diverse set of individuals. Thus, many strategies have been proposed in order to optimize and increase the genetic variability present in training populations.

An alternative to increasing genetic variability in training sets would be the inclusion of individuals from different populations (de Roos et al. 2009; Chen et al. 2013;

Saatchi et al. 2013). This inclusion can avoid a high degree of kinship between the training and validation sets and enhance the model's applicability across distinct populations. However, this is not always possible due to the population structure effect, which requires correction because it can affect the prediction of GEBVs (de los Campos and Sorensen 2014; Lehermeier et al. 2015).

However, while adding more individuals to the training sets ensures higher accuracy, this relationship is not necessarily linear (Hayes et al. 2009; Pérez-Cabal et al. 2012; Boddhireddy et al. 2014) and depends on quantity and quality of the newly added information (Hoffstetter et al. 2016; Rincent et al. 2017). This information comes from genetic (Pszczola et al. 2012; Rincent et al. 2012) and possibly phenotypic differences between individuals (Isidro et al. 2015). Consequently, the distribution of both genotypic and phenotypic information among the training and validation sets can affect the fit of the model, as well as its applicability. Thus, genotypic and phenotypic information could be used to optimize k-fold cross-validation composition, positively affecting model fit and model perpetuation.

Information theory can be used to measure phenotypic and genotypic information in the training and validation sets. This theory includes parameters such as entropy and mutual information, which were proposed by Shannon in his paper titled "*A mathematical theory of communication*" (Shannon 1948). Entropy (H) measures information by the amount of uncertainty regarding the value of a random variable; thus, if the random variable is a constant, there is no entropy or information. Mutual information (I) is a symmetric, non-negative, and nonlinear measurement of the amount of information shared between random variables (Shannon 1948; Cover and Thomas 2012).

This theory has been applied to the study of physics to show how information is incorporated into the second law of thermodynamics by its influence on non-equilibrium free energy (Parrondo et al. 2015). In ecology, H can be applied to understanding movement patterns of sheep with neurodegenerative disease and I can be applied to understanding the flight dynamics of pigeons (Owoeye et al. 2018). In statistical inference, H is used as an index of diversity and as a measurement of information distance between two probability density functions (Kullback-Leibler distance). For example, the least square method can be given by using the minimization of Kullback-Leibler divergence of the Gaussian model (Pardo 2006; Basu et al. 2011; Resende 2015). In economy, information theory can be applied to deriving a concentration index for

estimating the level of competition between the firms of a given sector (Pérez et al. 2018). In population genetics, with the use of linear approximation of I and H , one can obtain equations similar to the ones used for selection, gene flow, random drift, mutation, and LD estimates. For example, if a linear approximation is applied to the entropy of allelic frequencies, this entropy will obtain the same magnitude as heterozygosity frequency in a population in Hardy-Weinberg equilibrium (Smith 2012; Resende 2015).

Furthermore, methods based on I and H have been developed for finding genomic regions associated with animal production in genome-wide association studies (Borowska et al. 2017; Graczyk et al. 2017) and to obtain a representative set of single-nucleotide polymorphism (SNPs) in GWS. As an example, H was applied to reducing the number of SNPs used in genomic predictions for broiler breeding and resulted in higher accuracies than using all SNPs (Long et al. 2007). A similar approach is the Mutual Information Based Transductive Feature Selection (MINT) (He et al. 2016). This method uses I to obtain a representative set of SNPs for genomic prediction and showed high predictive accuracy with rr-BLUP (He et al. 2016), Bayes $C\pi$, BLASSO, Elastic net, Bayes A, and Bayes B (Haws et al. 2015). These examples show that H and I are outstanding tools for GWS and are potentially applicable to genetic tree breeding.

Therefore, the present study was aimed at developing novel strategies to acquire k-fold cross-validation sets based on phenotypic entropy and genetic mutual information. These strategies should produce trustworthy fitted models that can be applied to tree breeding, and other species populations, according to their phenotypic diversity and genetic relatedness to the training set. We also verified how the distribution of information affected parameters commonly used in GWS, such as $r_{y\hat{y}}$, bias, and genomic heritability.

4.2 Material and methods

4.2.1 Simulated data

All data simulations were performed using the HaploSim package (Coster and Bastiaansen 2009) in R software (R Core Team 2019). For this, a *Eucalyptus grandis*-like genome ($2n = 22$ chromosomes) was simulated with a total length of 13 Morgans and 10,000 equally distributed loci. Two hundred gametes were randomly combined into one hundred (N) non-inbred genotypes. Then, two thousand generations were simulated considering a higher mutation rate (10^{-5}) than those found for *E. grandis* ($4.96 \cdot 10^{-9}$ to $4.8 \cdot 10^{-7}$ per generation and base pairs) (Silva-Junior and Grattapaglia 2015). This mutational

rate was high enough to reach mutation-drift equilibrium, create linkage disequilibrium, and ensure polymorphic loci at the end of the simulation (Daetwyler et al. 2013). The recombination number followed a Poisson distribution and there was no interference at any of the recombination positions. Each of the 100 genotypes contributed 2 descendants to the next generation ($\bar{k} = 2$) with an equal number of descendants for each parent ($\sigma_k^2 = 0$). The effective population size (N_e) was obtained using the following: $(2N - 1)/(\bar{k} - 1 + \sigma_k^2/\bar{k})$, where N is the total number of genotypes, \bar{k} is the mean number of descendants per parent and σ_k^2 is the variance in the number of progeny per parent (Crow and Kimura 1970); thus, N_e was 199. The N_e of elite parents is usually 30 to 100 in genomic selection programs and over 300 in conservative tree breeding programs (Grattapaglia 2017). Consequently, the N_e of 199 was chosen to represent an average between these programs.

Next, individuals were crossed based on a pre-established pedigree of full-sib and half-sib families (partial diallel cross). The pedigree consisted of 20 different genitors (10x10) with 10 individuals per crossing (1,000 individuals in total). The 20 genitors were obtained after 2,001 generations by randomly combining gametes, and 1,000 individuals were obtained in generation 2,002 according to their pedigree. The mutation rate was null in the 2,001 and 2,002 generations to avoid low-frequency markers. A total of 4,316 markers were obtained by the end of the simulation. The LD was estimated by r^2 , which is equivalent to the correlation between alleles from different loci. The LD was corrected for family structure because there was no independence between gametes in the population. In this case, the covariance between all pairs of gametes can be used to decorrelate all observations and obtain a good estimate of LD (Mangin et al. 2012). The correction of LD for the family structure was implemented with the R software (R Core Team 2019) package LDcorSV (Desrousseaux et al. 2017). No population structure effect was evident, and thus no correction was performed. A nonlinear regression between megabase distance (Mb) and r^2 was fitted based on the drift-recombination model (Hill and Weir 1988) (Fig 4.1). In this study, the genome-wide average recombination rate was $3 cM Mb^{-1}$, which is a close value found for *E. grandis* ($3.18 cM Mb^{-1}$) (Silva-Junior and Grattapaglia 2015).

Two hundred QTLs were randomly allocated to loci with a minor allelic frequency (MAF) over 0.01. The true breeding values (TBVs) were simulated by summing QTL effects, which followed a Gaussian distribution ($(QTL \sim N(0, \sigma_{QTL}^2 = 0.25))$). Thereafter, a mean

equal to 510 and random residuals ($e \sim N(0, \sigma_e^2)$) were added. σ_e^2 was obtained considering $\sigma_{TBV}^2(1 - h^2)/h^2$, where σ_{TBV}^2 is the TBV variance and h^2 was equal to 0.5, producing a heritability of 0.5078 at the end of phenotype simulations. The trait was simulated to represent diameter at breast height in millimeters. Each QTL explained the additive variance (σ_a^2) in equal quantities in a polygenic model. Consequently, σ_a^2 can be given by $\sigma_{QTL}^2 \sum_{i=1}^Q 2p_i(1 - p_i)$, where σ_{QTL}^2 is the variance of QTLs effects, p_i is the allele frequency of i^{th} QTLs, and Q is the total of QTLs. Then, the variance of TBV explained by QTLs was obtained by $\sigma_a^2/\sigma_{TBV}^2$, which was equal to 21.19% in this study.

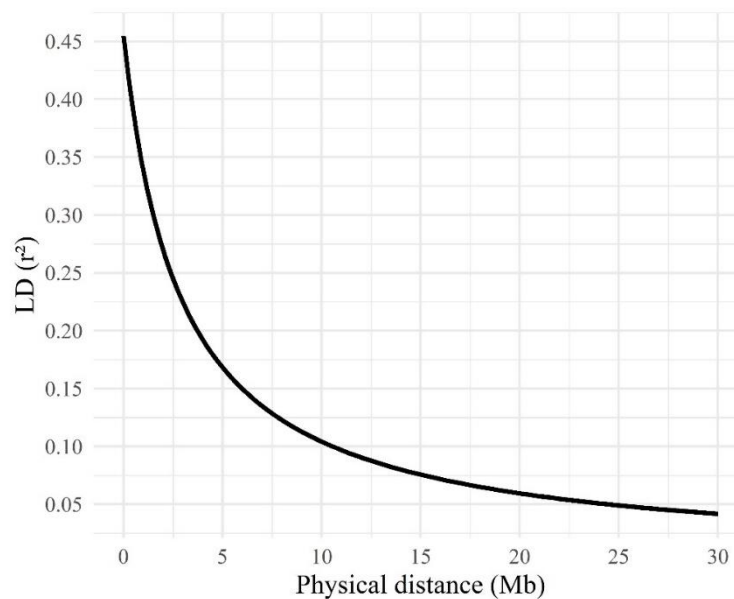


Fig. 4.1 Linkage disequilibrium (r^2) as a function of physical distance (in megabases pairs, *Mb*).

4.2.2 Real data

Pinus taeda L. (loblolly pine) data provided by Resende et al. (2012a) was used. This data came from a population of 32 parents crossed in a circular mating design, resulting in 70 full-sib families with an average of 13.5 trees per family. This population is from the lower Gulf of the United States and Atlantic coastal plain, Florida (USA). These families were genotyped using an Illumina Infinium assay with 7,216 SNPs. A subset of 4,853 polymorphic SNPs was selected by Resende et al. (2012a) and used in this study. None of the markers were excluded based on minor allelic frequency, and the minimum allelic frequency was higher than 0.005. Diameter at breast height with a heritability equal to 0.31 (Resende et al. 2012) was evaluated. Missing data was excluded, leaving 861 remaining individuals.

4.2.3 K-fold cross-validation methods

Eight k-fold cross-validation methods that split data into k sets in different ways were evaluated. These methods were the random method (method 1), the method described by Saatchi et al. (2011) (method 2), and 6 new methods based on phenotypic entropy (H) and genotypic mutual information (I) between individuals (methods 3 to 8). All methods were evaluated by applying 5, 10, 15, and 20-fold cross-validation schemes to verify the influence of different sizes of training and validation sets.

Method 1 (random method) randomly selected individuals to compose the k sets in the k-fold cross-validation. Method 2, described by Saatchi et al. (2011), uses a dissimilarity matrix (DS_{ij}) obtained from Wright kinship coefficients that is given as: $DS_{ij} = 1 - a_{ij}/\sqrt{a_{ii}a_{jj}}$, where a_{ij} is the kinship coefficient between individuals i and j , and a_{ii} or a_{jj} are the self-kinship coefficients ($1 + F$), where F is the inbreeding coefficient. When i was equal to j , DS_{ij} is null; thus, this matrix removes inbreeding effects (null diagonal) (Saatchi et al. 2011). Thereafter, validation sets were established using DS_{ij} with the K-means method (Hartigan and Wong 1979) in R software (R Core Team 2019). Consequently, method 2 produced higher kinship coefficients within validation sets and lower kinship coefficients between sets (Saatchi et al. 2011).

The new methods (3 to 8) used indices based on the relationship between phenotypic entropy (H) and genetic mutual information (I). These indices were applied to select and cluster individuals in k sets. I was used as a similarity measurement between individuals based on all markers, and H was used as a dissimilarity measurement based on phenotype values. A higher H and lower I means that the individuals are more phenotypically and genotypically different, respectively.

Method 3 applied the $1/I$ index, and thus, selected genotypically distinct individuals (lower I), generating genetically heterogenous validation sets. Method 4 used the I index, generating genetically homogenous validation sets. Method 5 applied the H/I index, clustering genetically and phenotypically heterogeneous individuals in the same validation set. Method 6 used the I/H index, producing validation sets that were genetically and phenotypically homogenous. Method 7 applied the $1/(HI)$ index and thus, generated genetically heterogeneous and phenotypically homogeneous validation sets. Method 8 used the HI index to produce genetically homogeneous and phenotypically heterogeneous validation sets.

Table 4.1 summarizes all methods, new indices, and genetic and phenotypic homogeneity or heterogeneity within validation sets. The genotypic control was performed by using Wright kinship coefficients or I , depending on the method. The phenotypic control was performed using H . Methods without any genotypic or phenotypic control had a higher probability of generating heterogenous validation sets. The even methods (2, 4, 6, and 8) generated genetically homogenous validation sets.

Table 4.1 Methods, indices, and genetic and phenotypic characteristics of the validation sets.

Methods	Indices	Genetic		Phenotypic	
		Heterogeneity	Homogeneity	Heterogeneity	Homogeneity
1	Random	-- ^a		--	
2	K-means		x	--	
3	$1/I$	x ^b		--	
4	I		x	--	
5	H/I	x		x	
6	I/H		x		x
7	$1/(HI)$	x			x
8	HI		x	x	

^a Methods without genetic or phenotypic control; ^b Methods with genetic heterogeneity or homogeneity and phenotypic heterogeneity or homogeneity.

In sequence, all measurements and the splitting procedure used in the new methods are explained below, to obtain the k validation sets (folds). Entropy (H) is a measure of uncertainty about the value of a random variable and can be employed as an estimate of diversity. H (Eq. 1) (Cover and Thomas 2012) was obtained for individual phenotypes ($y \sim N(\mu, \sigma^2)$).

$$H = - \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \right] dy = \ln \sqrt{2\pi e \sigma^2} \text{ nat} \quad \text{Eq. (1)}$$

Where μ and σ^2 are the phenotypes mean and variance, respectively. When the Napierian logarithm is used instead of logarithm base 2, the entropy unit is called *nat*. The phenotypes were standardized and transformed to avoid negative entropy (Supplemental Appendix A).

The mutual information (I) between individuals or groups based on markers is given by Eq. 2 (Meyer et al. 2008; Cover and Thomas 2012).

$$I = \sum_{x \in X} \sum_{z \in Z} p(x, z) \ln \left[\frac{p(x, z)}{p(x)p(z)} \right] \quad \text{Eq. (2)}$$

Where X and Z are individuals considered random variables that follow a multinomial distribution ($X \sim Multin(n; p_1, p_2, \dots, p_i)$), $p(x)$ and $p(z)$ are the probabilities of X and Z genotypes, respectively, $p(x, z)$ is the joint probability that follows a multinomial distribution ($X, Z \sim Multin(n; p_1, p_2, \dots, p_i)$), and the notation $x \in X$ indicates the x marker value (0, 1 or 2 in this study) that is possible in individual X .

Mutual information can also be estimated from the entropies of individuals or groups and is given by $I = H_X + H_Z - H_{XZ}$, where H_X and H_Z are the entropies of the individuals X and Z , respectively, and H_{XZ} is the joint entropy (Meyer et al. 2008; Cover and Thomas 2012). H_X is estimated as:

$$H_X = \sum_{x \in X} p(x) \ln \frac{1}{p(x)} \quad \text{Eq. (3)}$$

Where X is the individual or group of individuals considered random variables that follows a multinomial distribution ($X \sim Multin(n; p_1, p_2, \dots, p_i)$) and $p(x)$ is the probability of X . H_{XZ} is estimated by Eq. 3, with the joint probability $p(x, z)$, which also follows a multinomial distribution, instead of $p(x)$. When X is a single individual the distribution is trinomial, and $p(x)$ is the probability of M markers being declared recessive, heterozygous, or dominant.

The mutual information of the same group is also called redundancy (R) (Meyer et al. 2008). R measures the common information between all individuals of the same group. R (Eq. 4) is a symmetric non-linear, non-negative, measurement that does not decrease with the variable number, and is equal to I in the case of two individuals who belong to the group (Meyer et al. 2008; Cover and Thomas 2012).

$$R_{X_1, \dots, X_k} = \sum_{i=1}^k H_{X_i} - H_{X_1, \dots, X_k} \quad \text{Eq. (4)}$$

Where X_1, \dots, X_k are the K individuals in the group, H_{X_i} is the entropy of the i^{th} individual, H_{X_1, \dots, X_k} is the joint entropy of all individuals based on markers. Each individual and group follows a multinomial distribution. In the case of a three-individual group, R is given as: $R_{X_1, X_2, X_3} = H_{X_1} + H_{X_2} + H_{X_3} - H_{X_1, X_2, X_3}$. The ‘‘infotheo’’ (Meyer 2009) R software (R Core Team 2019) package was used to estimate I and R with the *mutinformation* and *multiinformation* functions, respectively.

The steps used for splitting data into k validation sets are:

- A. Obtain a matrix of I and H between all individuals.

- B. Select the maximum value of I and H for everyone.
- C. Apply the index according to the method. For example, H is divided by I in method 5.
- D. Put index values in decreasing order and select the first two individuals with a higher index value.
- E. Estimate R and H between the set of individuals from the previous step with the other individuals (Here, R was used instead of I , because the mutual information was estimated between more than three individuals).
- F. Apply the index, select the individual with a higher index value, and include it in the validation set.
- G. Repeat steps E and F until a validation set is formed. For example, a complete validation set is composed of 20 individuals in 5-fold cross-validation of a hundred individuals.
- H. Save the formed validation set and repeat the steps A – G with the rest of the data until all k validation sets are obtained.

The splitting procedure for all methods formed k validation sets at the end. All methods were then evaluated by fitting a linear model in a k -fold cross-validation scheme, using k validation sets of the respective methods. Fig 4.2 summarizes the splitting procedure of data into k validation sets through the indices (Table 4.1), the fit of the genomic model and estimation of parameters used to evaluate the models. The fitted linear model was given by $y = X\beta + Zg + e$, where y is the data vector, β is the mean vector, g is the additive genetic effect vector ($g|\sigma_g^2 \sim N(0, G\sigma_g^2)$), and e is the residual effect vector ($e \sim N(0, I\sigma_e^2)$). X and Z are the incidence of matrices for β and g effects, respectively. The parameter σ_g^2 follows a scaled inverse chi-squared distribution with 4 degrees of freedom (df_g) and a scale parameter (S_g) equal to 3 ($\sigma_g^2 \sim X^{-2}(df_g, S_g)$). The same distribution was adopted for σ_e^2 ($\sigma_e^2 \sim X^{-2}(df_e = 4, S_e = 3)$). The hyperparameters df_g, S_g, df_e and S_e were the same as those used by Lehermeier et al. (2015). The G matrix was estimated by WW'/m , where W is the marker matrix and $m = \sum_{i=1}^M 2p_i(1 - p_i)$, where p_i is the frequency of the i^{th} marker and M is the total number of markers. The G matrix is the realized genomic relationship matrix and was obtained according to the first method described by VanRaden (2008).

The model was fitted via Bayesian Genomic Best Linear Unbiased Predictor (GBLUP) using the BGLR package (Pérez and de los Campos 2014). Of note, 120,000

iterations were used, with a burn-in (iterations removed) of 40,000 and a thinning of 5. The convergence of the Markov chains was evaluated using graphics and the Raftery and Lewis (1992) and Geweke (1992) methods in the “coda” (Plummer et al. 2010) R software (R Core Team 2019) package.

4.2.4 Evaluation of k-fold cross-validation methods

The methods were compared based on their predictive ability and bias between GEBVs and the phenotypes of the validation set. The predictive ability ($r_{y\hat{y}}$), which is directly related to the accuracy, was estimated by Pearson correlation. Bias (β_1) was estimated using the linear regression coefficient. The minimum squared error of prediction (MSEP) was estimated for the simulated data. MSEP was given by $MSEP = V(\hat{y} - y) + [E(\hat{y} - y)]^2 = PEV + BIAS^2$, where \hat{y} is the GEBVs, y is the validation set phenotypes, PEV is the variance of prediction errors, and $BIAS^2$ is the squared bias. Genomic heritability (h_g^2) was estimated by $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$, where σ_g^2 and σ_e^2 are the genomic and residual variance, respectively. Kinship coefficient means and mutual information between training and validation sets ($I_{Tr.\times Val.}$) were also estimated.

Pearson correlations of $r_{y\hat{y}}$ and β_1 with phenotypic entropy within training ($H_{Tr.}$) and validation ($H_{Val.}$) sets, the mutual information between training and validation sets ($I_{Tr.\times Val.}$), and Wright kinship coefficient means between training and validation sets ($A_{Tr.\times Val.}$) were obtained. All correlations were tested by Student’s t-test with a 5% probability of type I error. The differences between methods for each k-fold cross-validation scheme were verified by applying the Scott-Knott test (Scott and Knott 1974), with 1% probability of type I error for $I_{Tr.\times Val.}$, $A_{Tr.\times Val.}$, $H_{Val.}$, $r_{y\hat{y}}$, β_1 , MSEP, and h_g^2 .

Online resource 1 (.pdf) contains supplementary material regarding the derivation of the avoidance of negative entropy. Online resource 2 (.pdf) contains scripts for the novel methods for splitting k-fold cross-validation sets and an example of 100 individuals and 500 markers.

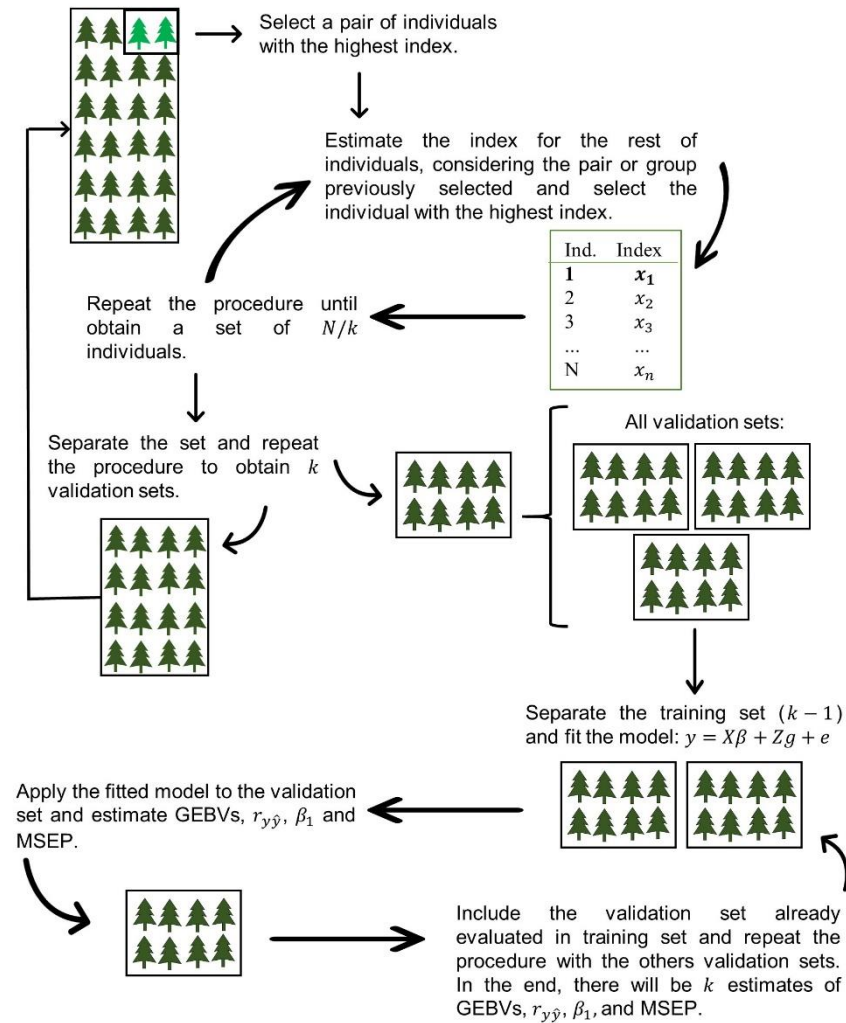


Fig. 4.2 Procedure for splitting the total individuals (N) into k validation sets according to the proposed indices, the model's fit and the estimate of genomic estimated breeding values (GEBVs), predictive ability ($r_{y\hat{y}}$), bias (β_1), and minimum squared error of prediction (MSEP) for the k validation sets.

4.3 Results

4.3.1 Genotypic and phenotypic similarity

Mutual information between the training and validation sets ($I_{Tr.\times Val.}$) was higher ($p < 0.01$) for the odd methods (1, 3, 5, and 7), which formed genetically heterogenous validation sets (Table 4.1) for simulated (Fig. 4.3) and *P. taeda* data (Supplemental Fig 4.S1). The even methods had greater genotypic homogeneity within the validation sets and lower ($p < 0.01$) genetic similarity between the training and validation sets (Fig 4.3 and 4.S1). The same observations made for $I_{Tr.\times Val.}$ can also be applied to the mean

Wright kinship coefficient between the training and validation sets ($A_{Tr.\times Val.}$) for both the simulated (Fig 4.4) and *P. taeda* data (Supplemental Fig 4.S2).

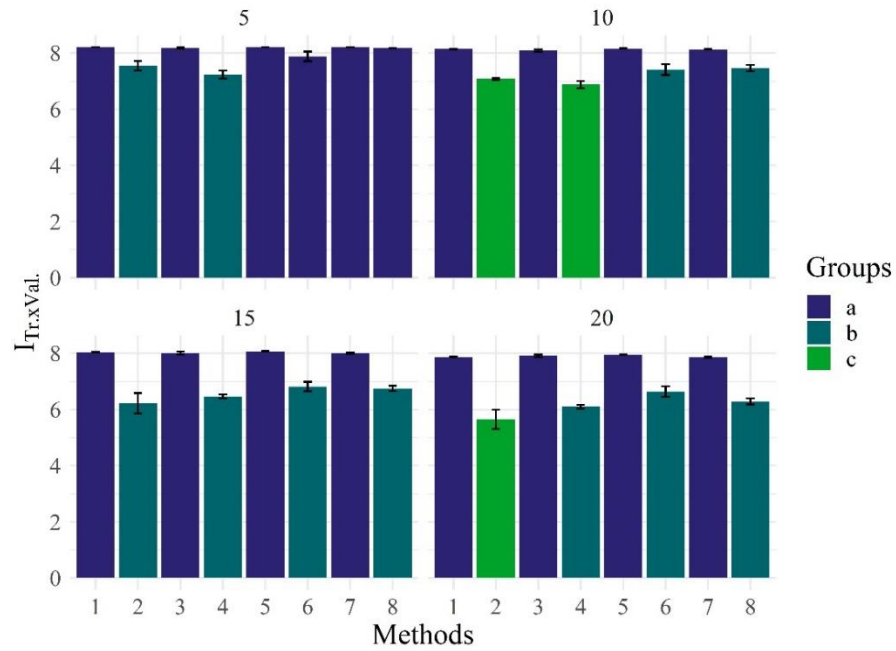


Fig. 4.3 Average of mutual information between the training and validation sets ($I_{Tr.\times Val.}$) for all the methods for 5 to 20-fold cross-validation schemes with simulated data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

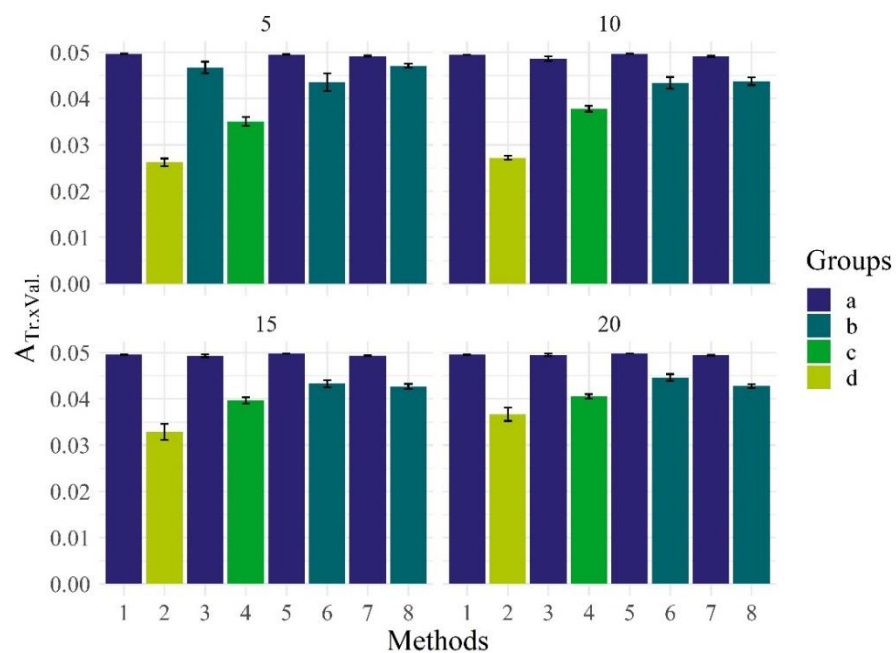


Fig. 4.4 Average of Wright kinship coefficients between training and validation sets ($A_{Tr.\times val.}$) for all the methods for 5 to 20-fold cross-validation schemes with simulated data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), and 8 (HI) showed higher ($p < 0.01$) phenotypic variability in the validation sets ($H_{Val.}$) than other methods for 15 and 20-fold cross-validation schemes (Fig 4.5). Method 7 showed lower ($p < 0.01$) $H_{Val.}$ values than any other method for 10 to 20-fold cross-validation schemes using simulated and *P. taeda* data (Fig 4.5 and Supplemental Fig 4.S3, respectively).

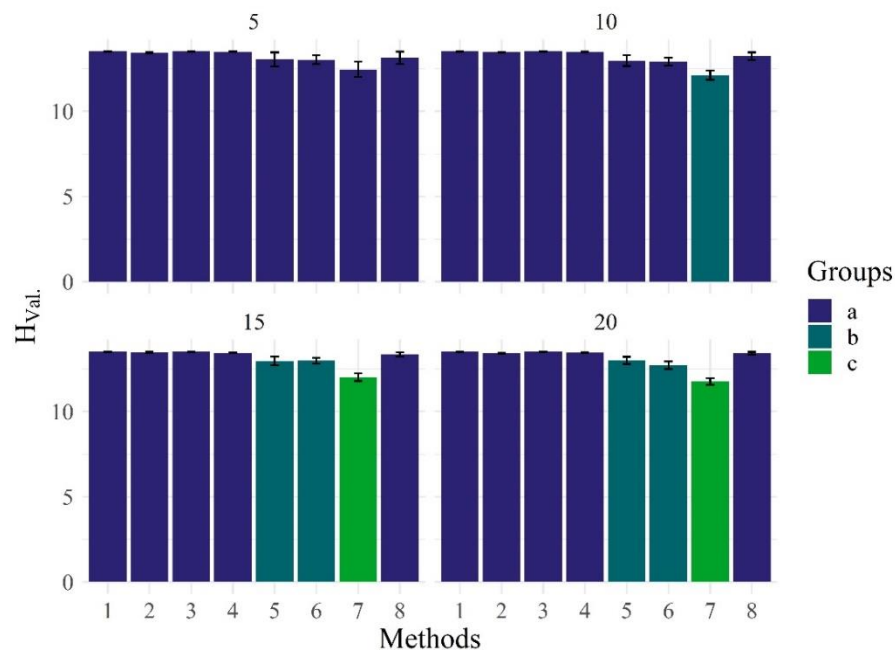


Fig. 4.5 Average of phenotypic entropy within validation sets ($H_{Val.}$) for all the methods for 5 to 20-fold cross-validation schemes with simulated data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

4.3.2 Predictive ability and bias

Methods 6 (I/H) and 7 (1/(HI)) showed lower ($p < 0.01$) $r_{y\hat{y}}$ values than other methods for 10 to 20-fold cross-validation schemes (except method 2) (Fig 4.6). Lower ($p < 0.01$) $r_{y\hat{y}}$ values were also found when using these methods and methods 2 (K-

means) and 4 (*I*) for 15 and 20-fold cross-validation schemes with *P. taeda* data (Supplemental Fig 4.S4).

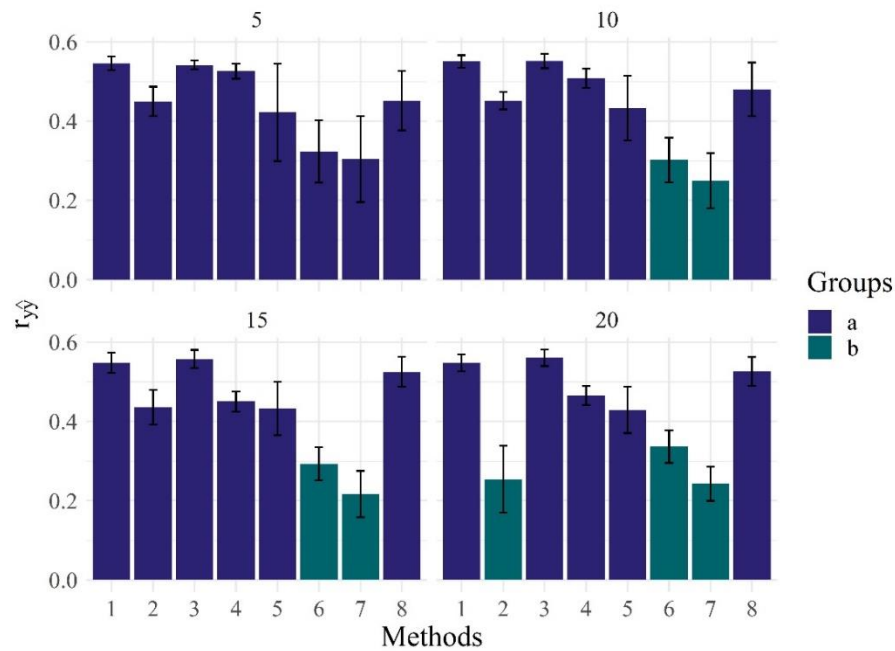


Fig. 4.6 Average of predictive ability ($r_{\hat{y}}$) for all the methods for 5 to 20-fold cross-validation schemes with simulated data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/*I*), 4 (*I*), 5 (*H/I*), 6 (*I/H*), 7 (1/(*HI*)), and 8 (*HI*).

Methods 1 (random), 2 (K-means), 3 (1/*I*), 4 (*I*), 5 (*H/I*), and 8 (*HI*) showed β_1 values close to one (Fig 4.7). However, methods 5 and 8 showed higher standard error values. Methods 6 (*I/H*) and 7 (1/(*HI*)) showed lower ($p < 0.01$) β_1 values than any other method for 15 and 20-fold cross-validation schemes (Fig 4.7). These observations were also applicable to *P. taeda* data (Supplemental Fig 4.S5). A higher number of individuals in the validation sets (5-fold cross-validation) did not highlight the differences between methods for most of parameters (Fig 4.5 to 4.7).

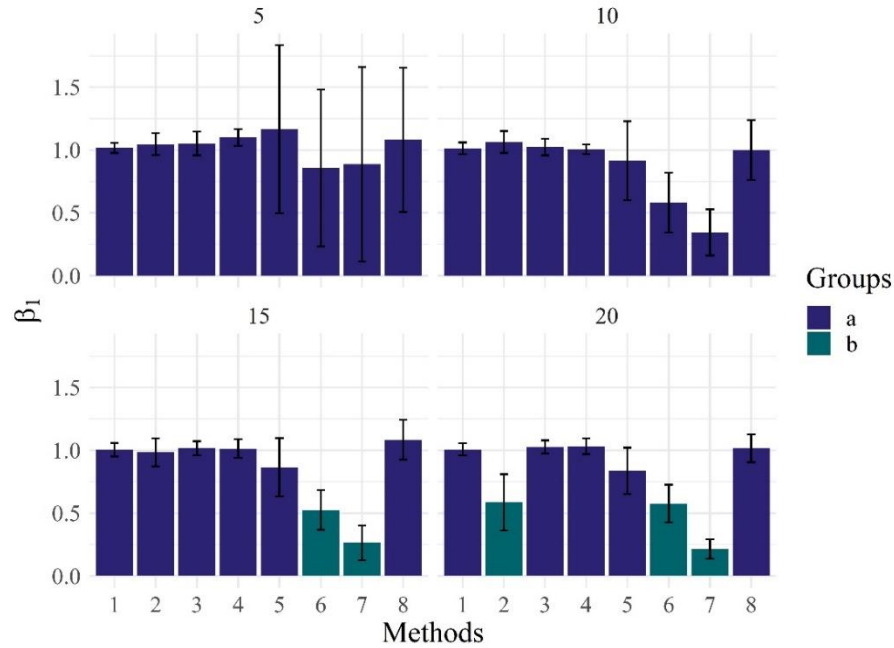


Fig. 4.7 Average of bias (β_1) for all the methods for 5 to 20-fold cross-validation schemes with simulated data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 ($1/I$), 4 (I), 5 (H/I), 6 (I/H), 7 ($1/(HI)$), and 8 (HI).

4.3.3 Correlations between model parameters with phenotypic and genotypic information

Table 4.2 shows the Pearson correlations between $r_{y\hat{y}}$, β_1 , and MSEP with the mutual information between the training and validation sets ($I_{Tr.\times Val.}$), kinship coefficient means between training and validation sets ($A_{Tr.\times Val.}$), and phenotypic entropy within validation ($H_{Val.}$) and training sets ($H_{Tr.}$) for 20-fold cross-validation. These correlations showed the linear influence of these parameters on $r_{y\hat{y}}$, β_1 , and MSEP when using different methods.

Table 4.2 Linear correlations of parameters minimum squared error of prediction ($MSEP$), predictive ability ($r_{y\hat{y}}$), and bias (β_1) with mutual information and kinship coefficient means between training and validation sets ($I_{Tr.\times Val.}$ and $A_{Tr.\times Val.}$, respectively), entropy within validation and training sets ($H_{Val.}$ and $H_{Tr.}$, respectively) for 20-fold cross-validation for all methods with simulated data. Methods 1 (random), 2 (K-means), 3 ($1/I$), 4 (I), 5 (H/I), 6 (I/H), 7 ($1/(HI)$) and 8 (HI).

Methods	Parameters	$I_{Tr.\times Val.}$	$A_{Tr.\times Val.}$	$H_{Val.}$	$H_{Tr.}$
---------	------------	----------------------	----------------------	------------	-----------

1	<i>MSEP</i>	-0.12	-0.21	0.54	-0.6*
1	$r_{y\hat{y}}$	0.15	0.2	0.4	-0.34
1	β_1	0.19	0.2	0.7*	-0.64*
2	<i>MSEP</i>	-0.44	0.38	0.43	-0.35
2	$r_{y\hat{y}}$	0.68*	-0.6*	0.53	-0.13
2	β_1	0.69*	-0.62*	0.5	-0.14
3	<i>MSEP</i>	0.38	0.38	0.76*	-0.78*
3	$r_{y\hat{y}}$	-0.26	-0.33	0.2	-0.18
3	β_1	0.02	-0.05	0.54	-0.51
4	<i>MSEP</i>	0.05	0.11	0.82*	-0.51
4	$r_{y\hat{y}}$	0.36	0.23	0.51	-0.39
4	β_1	0	-0.05	0.68*	-0.37
5	<i>MSEP</i>	0.26	0.11	0.7*	-1*
5	$r_{y\hat{y}}$	0.55	0.43	0.89*	-0.79*
5	β_1	0.38	0.22	0.81*	-0.98*
6	<i>MSEP</i>	0.34	0.32	0.43	-0.96*
6	$r_{y\hat{y}}$	-0.16	-0.18	0.58*	-0.61*
6	β_1	-0.1	-0.11	0.69*	-0.81*
7	<i>MSEP</i>	-0.58*	-0.64*	0.47	-0.99*
7	$r_{y\hat{y}}$	-0.24	-0.39	0.77*	-0.46
7	β_1	-0.49	-0.68*	0.84*	-0.67*
8	<i>MSEP</i>	-0.66*	-0.68*	0.88*	-0.91*
8	$r_{y\hat{y}}$	-0.32	-0.35	0.37	-0.51
8	β_1	-0.68*	-0.64*	0.78*	-0.87*

* $p < 0.05$

The increase of phenotypic diversity in the training sets ($H_{Tr.}$) decreased $r_{y\hat{y}}$, β_1 and *MSEP* values for all methods (negative correlations), while the increase in the diversity of the validation sets ($H_{Val.}$) increased them (positive correlations) (Table 4.2). The same occurred for $r_{y\hat{y}}$ and β_1 with *P. taeda* data, except in method 1 (random) (Supplemental Table 4.S1). However, the methods that included *H* in the splitting process (5 to 8) highlighted these correlations ($p < 0.05$), especially method 5, for simulated and *P. taeda* data. For example, methods 5 to 7 showed positive correlations ($p < 0.05$) for $H_{Val.}$ on $r_{y\hat{y}}$ and β_1 (Table 4.2). Methods 5 to 8 showed negative correlations ($p < 0.05$) for $H_{Tr.}$ with *MSEP* and β_1 (Table 4.2), and positive correlations were found ($p <$

0.05) between MSEP and β_1 with simulated data (0.96, 0.73, 0.6, and 0.79 for methods 5 (H/I), 6 (I/H), 7 ($1/(HI)$), and 8 (HI), respectively).

A higher positive correlation was expected between $r_{y\hat{y}}$ and $I_{Tr.\times Val.}$, but was shown ($p < 0.05$) only for methods 1 (random) and 5 (H/I) with *P. taeda* data (Table 4.S1) and method 2 (K-means) with simulated data (Table 4.2). Furthermore, no correlation ($p \geq 0.05$) was found between $r_{y\hat{y}}$ and $A_{Tr.\times Val.}$, except when using method 2 with simulated data (Table 4.2 and Table 4.S1). Method 2 was expected to show a high positive influence for kinship coefficients on $r_{y\hat{y}}$. However, this method showed a negative correlation (-0.6; $p < 0.05$) between $r_{y\hat{y}}$ and $A_{Tr.\times Val.}$ for simulated data (Table 4.2) and no correlation for *P. taeda* data (-0.26; $p \geq 0.05$) (Table 4.S1).

Mean genomic heritability (h_g^2) and MSEP were equal for all methods. Supplemental Table 4.S2 shows the mean values for $r_{y\hat{y}}$, β_1 , MSEP, and h_g^2 for all methods and k-fold cross-validation schemes with simulated data. Supplemental Table 4.S3 shows the same parameters as Table 4.S2 for *P. taeda* data, except MSEP. Online resource 1 (.pdf) contains all supplemental tables and figures for the simulated and *P. taeda* data.

4.4 Discussion

Tree breeding populations are usually phenotypically and genotypically diverse, but this diversity can decrease over time due to evolutionary factors (i.e., selection and genetic drift). Hence, these populations present different levels of genetic and phenotypic diversity. Even though these factors can interfere with predictions of genomic-based models, they are not typically considered in analysis. Here, we addressed this issue by proposing the use of mutual information and entropy as measurements for genotypic and phenotypic information, respectively. The use of this information in the validation process helped avoid overestimations of predictive ability, as well as generated models that can be applied to more or less diverse populations and produce strong results.

4.4.1 The effect of diversity on $r_{y\hat{y}}$ values

Method 7 ($1/(HI)$) demonstrated that lower ($p < 0.01$) phenotypic diversity (Fig 4.5) produces lower ($p < 0.01$) $r_{y\hat{y}}$ values (Fig 4.6) for higher ($p < 0.01$) genetic relationships between training and validation sets (Fig 4.3 and 4.4). The same evidence was observed in the positive correlations between phenotypic diversity of validation sets ($H_{Val.}$) and $r_{y\hat{y}}$ with simulated data (Table 4.2) and *P. taeda* data (Supplemental Table

4.S1). Hence, tree breeding populations that are genetically closely related to the training set can exhibit lower $r_{y\hat{y}}$ values if they are less phenotypically diverse. Other studies concluded that the genetic relationship between training and validation sets was positively correlated to $r_{y\hat{y}}$ (Saatchi et al. 2011, 2013; Pérez-Cabal et al. 2012; Hulsman Hanna et al. 2015). For example, high correlations (> 0.7) were estimated between $r_{y\hat{y}}$ and the mean of 10 or 100 higher kinship coefficients (Clark et al. 2012). However, none of these studies explored whether low accuracies could be due to low phenotypic diversity in the validation sets. In our study, this was possible because most of the proposed methods included phenotypic entropy (method 5 to 7) and highlighted the positive correlation ($p < 0.05$) between $H_{Val.}$ and $r_{y\hat{y}}$ (Table 4.2). To our knowledge, this is the first study that investigated the influence of phenotypic diversity of validation sets on $r_{y\hat{y}}$.

Some studies have reported low accuracies for method 2 (K-means), which were related to low Wright kinship coefficients between the training and validation sets ($A_{Tr.\times Val.}$) (Saatchi et al. 2011, 2013; Boddhireddy et al. 2014). However, $A_{Tr.\times Val.}$ had a negative correlation ($p < 0.05$) with $r_{y\hat{y}}$ in this method, despite $I_{Tr.\times Val.}$ being positively correlated (Table 4.2). Furthermore, method 2 (K-means) and method 1 (random) showed no differences ($p \geq 0.01$) in $r_{y\hat{y}}$ values for 5 to 15-fold cross-validation (Fig 4.6), but exhibited significantly different $A_{Tr.\times Val.}$ values ($p < 0.01$) (Fig 4.4). Therefore, the relationship between $r_{y\hat{y}}$ and the mean of kinship coefficient between the training and validation sets is not direct (Pérez-Cabal et al. 2012), and they are not always strongly correlated (Hulsman Hanna et al. 2015).

The lower correlation between $r_{y\hat{y}}$ and $A_{Tr.\times Val.}$ could be due to pedigree errors, cryptic genetic relationships, or different kinship relationships with the same kinship coefficient (i.e., half-sib and uncle/aunt-nephew/niece) (Pérez-Cabal et al. 2012; Speed and Balding 2015). Moreover, other factors also influence $r_{y\hat{y}}$, like the number of close relatives shared between the training and validation sets, mainly for traits with low heritability (Pérez-Cabal et al. 2012), fixation index (Scutari et al. 2016), and LD phase persistence between training and validation sets (Chen et al. 2013).

4.4.2 The effect of diversity on β_1 and MSE

β_1 values higher than one indicate an underestimate of GEBVs, while values lower than one show an overestimate, and β_1 values close to one are ideal. The β_1 values were lower than one ($p < 0.01$) in methods 6 (I/H) and 7 ($1/(HI)$) in 15 and 20-fold cross-

validation schemes (Fig 4.7), and showed positive correlations with phenotypic diversity of validation sets (Table 4.2). Consequently, the GEBVs were overestimated ($\beta_1 < 1$) in populations with lower phenotypic diversity, limiting the applications for genomic prediction to only highly diverse populations.

MSEP showed negative correlations ($p < 0.05$) with phenotypic diversity of training set for methods 5 to 8 (Table 4.2). Consequently, a higher diversity of the training set can provide lower MSEP and better prediction of GEBVs. A strategy to improve the phenotypic diversity of the training set would be the mixing of different populations. However, this is not always possible, and in the case of the population structure effect, would require correction for accurate prediction of GEBVs (de los Campos and Sorensen 2014; Lehermeier et al. 2015). Some studies have linked multiple populations in the training set, producing higher $r_{y\hat{y}}$ values (de Roos et al. 2009; Chen et al. 2013; Saatchi et al. 2013), but $r_{y\hat{y}}$ was in fact lower in some cases (Saatchi et al. 2013). Therefore, fitting genomic models to a mixture of populations does not necessary result in a better fit.

4.4.3 The choice of k-fold cross-validation method for 20-fold cross-validation scheme

The validation sets in a k-fold cross-validation scheme mimic populations where genomic selection would be applied, and where genomic information would be present and phenotypic information would not. Even in cases with unknown phenotypic information, it can be measured using a few samples from the population. Moreover, populations used at the onset of tree breeding programs are more phenotypically diverse. Consequently, methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), and 8 (HI) are ideal for these populations, because their validation sets showed higher ($p < 0.01$) phenotypic diversity for both datasets (Fig. 4.5 and 4.S3, respectively). However, only methods 1, 3, and 8 showed higher $r_{y\hat{y}}$ and β_1 close to 1 (Fig 4.6 and 4.7 and Fig. 4.S4 and 4.S5). Therefore, to obtain trustworthy fitted models in more diverse populations, methods 1 or 3 are recommended in cases of closer genetic relationships between the training set and tree breeding population, and method 8 should be used otherwise (Fig. 4.3 and 4.4). Method 5 (H/I) showed values with high amplitude for H_{Val} . and can be used to study the influence of phenotypic diversity on $r_{y\hat{y}}$ and β_1 values.

Method 7 (1/(HI)) showed lower ($p < 0.01$) phenotypic diversity for both datasets (Fig. 4.5 and 4.S3, respectively). Therefore, this method is recommended for

estimating an applicable genetic gain in populations with lower diversity, but only when the training and tree breeding populations are closely genetically related (Fig. 4.3 and 4.4). The methods (1, 3, 4, 5, and 8) that showed higher ($p < 0.01$) $r_{y\hat{y}}$ for both datasets should not be used to evaluate less phenotypically diverse populations. If these methods were applied in these populations, they would overestimate $r_{y\hat{y}}$, and consequently estimate a misleading genetic gain.

4.5 Conclusions

The validation methods evaluated here can help in the decision process for applying genome selections in tree breeding populations. Therefore, k-fold cross-validation methods should be considered than a simple validation and employed to produce fitted models that account for the current phenotypic diversity of the tree breeding populations and their genetic relatedness to the training set.

The new validation methods highlighted the fact that predictive ability is more linearly related to the phenotypic diversity of the validation set than to the kinship between the training and validation sets. Therefore, an increase in the phenotypic diversity of tree breeding populations is essential for obtaining higher genetic gains, and consequently the consistent success of long-term tree breeding programs.

4.6 Data archiving statement

Simulated datasets generated during and/or analyzed during the current study are available in the *Figshare* repository, <https://doi.org/10.6084/m9.figshare.9619571.v1>.

4.7 Acknowledgments

We are thankful to the Federal University of Viçosa, the National Council for Scientific and Technological Development, and the Coordination for the Improvement of Higher Education Personnel for their financial support. We are grateful to professors Camila F. Azevedo, Fabyano F. da Silva, Moysés Nascimento, and Rodrigo O. de Lima for their important comments, suggestions, and criticisms, which improved the manuscript.

4.8 References

- Basu A, Shioya H, Park C (2011) *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC
- Bodhireddy P, Kelly MJ, Northcutt S, et al (2014) Genomic predictions in Angus cattle:

- comparisons of sample size, response variables, and clustering methods for cross-validation. *J Anim Sci* 92:485–497. doi: 10.2527/jas.2013-6757
- Borowska A, Reyer H, Wimmers K, et al (2017) Detection of pig genome regions determining production traits using an information theory approach. *Livest Sci* 205:31–35. doi: 10.1016/j.livsci.2017.09.012
- Chen L, Schenkel F, Vinsky M, et al (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *J Anim Sci* 91:4669–4678. doi: 10.2527/ja4.S2013-5715
- Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol* 44:4. doi: 10.1186/1297-9686-44-4
- Coster A, Bastiaansen JWM (2009) HaploSim. R Packag. version 1.8
- Cover TM, Thomas JA (2012) Elements of information theory. John Wiley & Sons
- Crow JF, Kimura M (1970) An introduction to population genetics theory. New York, Evanston and London: Harper & Row, Publishers
- Daetwyler HD, Calus MPL, Pong-Wong R, et al (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193:347–365. doi: 10.1534/genetics.112.147983
- Daetwyler HD, Kemper KE, Werf JHJ van der, Hayes BJ (2012) Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci* 90:3375–3384. doi: doi:10.2527/ja4.S2011-4557
- de los Campos G, Sorensen D (2014) On the genomic analysis of data from structured populations. *J Anim Breed Genet* 131:163–164. doi: 10.1111/jbg.12091
- de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553. doi: 10.1534/genetics.109.104935
- Desrousseaux D, Sandron F, Siberchicot A, et al (2017) Package ‘LDcorSV’
- Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Stat* 4:641–649
- Graczyk M, Reyer H, Wimmers K, Szwaczkowski T (2017) Detection of the important chromosomal regions determining production traits in meat-type chicken using

- entropy analysis. *Br Poult Sci* 58:358–365. doi: 10.1080/00071668.2017.1324944
- Grattapaglia D (2017) Status and perspectives of genomic selection in forest tree breeding. In: *Genomic selection for crop improvement*. Springer, pp 199–249
- Guo Z, Tucker DM, Basten CJ, et al (2014) The impact of population structure on genomic prediction in stratified populations. *Theor Appl Genet* 127:749–762. doi: 10.1007/s00122-013-2255-x
- Habier D, Tetens J, Seefried FR, et al (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol* 42:5. doi: 10.1186/1297-9686-42-5
- Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C* 28:100–108
- Haws DC, Rish I, Teysseire S, et al (2015) Variable-selection emerges on top in empirical comparison of whole-genome complex-trait prediction methods. *PLoS One* 10:1–22. doi: 10.1371/journal.pone.0138903
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443
- He D, Rish I, Haws D, Parida L (2016) MINT: mutual information based transductive feature selection for genetic trait prediction. *IEEE/ACM Trans Comput Biol Bioinforma* 13:578–583. doi: 10.1109/TCBB.2015.2448071
- Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* 33:54–78
- Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. *G3 Genes|Genomes|Genetics* 6:2919–2928. doi: 10.1534/g3.116.032532
- Hulsman Hanna LL, Garrick DJ, Gill CA, et al (2015) Cross-validation of genetic and genomic predictions of temperament in Nellore-Angus crossbreds. *Livest Sci* 182:28–33. doi: 10.1016/j.livsci.2015.10.020
- Isidro J, Jannink JL, Akdemir D, et al (2015) Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128:145–158. doi: 10.1007/s00122-014-2418-4
- Lehermeier C, Schon CC, de los Campos G (2015) Assessment of genetic heterogeneity

- in structured plant populations using multivariate whole-genome regression models. *Genetics* 201:323–337. doi: 10.1534/genetics.115.177394
- Long N, Gianola D, Rosa GJM, et al (2007) Machine learning procedure for selecting single nucleotide polymorphisms in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 124:377–389
- Mangin B, Siberchicot A, Nicolas S, et al (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)* 108:285–291. doi: 10.1038/hdy.2011.73
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829. doi: 11290733
- Meyer PE (2009) Package ‘infotheo.’ R Packag. version 1
- Meyer PE, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. *IEEE J Sel Top Signal Process* 2:261–274. doi: 10.1109/JSTSP.2008.923858
- Owoeye K, Musolesi M, Hailes S (2018) Characterizing animal movement patterns across different scales and habitats using information theory. *bioRxiv* 1–20. doi: 10.1101/311241
- Pardo L (2006) *Statistical inference based on divergence measures*. Chapman and Hall/CRC
- Parrondo JMR, Horowitz JM, Sagawa T (2015) Thermodynamics of information. *Nat Phys* 11:131–139. doi: 10.1038/nphy4.S3230
- Pérez-Cabal MA, Vazquez AI, Gianola D, et al (2012) Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front Genet* 3:1–7. doi: 10.3389/fgene.2012.00027
- Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198:483–495. doi: 10.1534/genetics.114.164442
- Pérez R, López AJ, Caso C, et al (2018) On economic applications of information theory. In: *The Mathematics of the Uncertain*. Springer, pp 515–525
- Plummer M, Best N, Cowles K, Vines K (2010) coda: output analysis and diagnostics for MCMC. R package version 0.14-2
- Pszczola M, Strabel T, Mulder HA, Calus MPL (2012) Reliability of direct genomic

- values for animals with different relationships within and to the reference population. *J Dairy Sci* 95:389–400. doi: 10.3168/jds.2011-4338
- R Core Team (2019) R: a language and environment for statistical computing
- Raftery AE, Lewis SM (1992) [Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Stat Sci* 7:493–497
- Resende MFR, Muñoz P, Resende MDV, et al (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190:1503–1510
- Resende RT, Resende MDV, Silva FF, et al (2017a) Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. *Heredity (Edinb)* 119:245–255. doi: 10.1038/hdy.2017.37
- Resende RT, Resende MDV, Silva FF, et al (2017b) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytol* 213:1287–1300. doi: 10.1111/nph.14266
- Resende MD V (2015) Genética quantitativa e de populações. Suprema, Visconde do Rio Branco 452
- Rincent R, Charcosset A, Moreau L (2017) Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theor Appl Genet* 130:2231–2247. doi: 10.1007/s00122-017-2956-7
- Rincent R, Laloë D, Nicolas S, et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192:715–728. doi: 10.1534/genetics.112.141473
- Saatchi M, McClure MC, McKay SD, et al (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet Sel Evol* 43:1–16. doi: 10.1186/1297-9686-43-40
- Saatchi M, Ward J, Garrick DJ (2013) Accuracies of direct genomic breeding values in Hereford beef cattle using national or international training populations. *J Anim Sci* 91:1538–1551. doi: 10.2527/jas.2012-5593
- Scott AJ, Knott M (1974) A cluster analysis method for grouping means in the analysis

of variance. *Biometrics* 507–512

Scutari M, Mackay I, Balding D (2016) Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet* 12:e1006288

Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423. doi: 10.1145/584091.584093

Silva-Junior OB, Grattapaglia D (2015) Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. *New Phytol* 208:830–845. doi: 10.1111/nph.13505

Silva RMO, Fragomeni BO, Lourenco DAL, et al (2016) Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. *J Anim Sci* 94:3613–3623. doi: 10.2527/ja4.S2016-0401

Smith RD (2012) Information theory and population genetics. arXiv:11035625v2[q-bioPE] (Quantitative Biology)

Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* 16:33–44

VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–4423. doi: 10.3168/jds.2007-0980

Wray NR, Yang J, Hayes BJ, et al (2013) Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 14:507–515. doi: 10.1038/nrg3457

4.9 Appendix A: Derivation of avoidance of negative entropy

This is a part of Online resource 1.

The standard deviation between two values (σ_{xy}) is given by:

$$\sigma_{xy} = \sqrt{\frac{1}{2-1} \left[\left(x - \frac{x+y}{2} \right)^2 + \left(y - \frac{x+y}{2} \right)^2 \right]} = \sqrt{\frac{1}{4} (x^2 - 2xy + y^2 + x^2 - 2xy + y^2)} = |x - y| \sqrt{\frac{1}{2}}$$

Eq. A1

Where x and y are values of a random variable. By applying Eq. A1 into Eq. 1, we get Eq. A2.

$$H_{x,y} = 0.5 \ln 2\pi e + \ln \sigma_{xy} = 0.5 \ln 2\pi e + \ln[|x - y| \sqrt{0.5}]$$

Eq. A2

Where $H_{x,y}$ is the entropy between two phenotypic values (x and y).

When Eq. A2 equals zero ($H_{x,y} = 0$), the difference in values that causes negative entropy is obtained. Thus,

$$-0.5 \ln 2\pi e = 0.5 \ln 0.5 + \ln|x - y|$$

$$-0.5 \ln (\pi e) = \ln|x - y|$$

$$x - y = e^{-0.5 \ln (\pi e)} = 1/\sqrt{\pi e} \cong 0.3422$$

Therefore, differences between x and y lower than 0.3422 causes negative entropy. The transformation of standardized values was performed by multiplying the reciprocal number of lower amplitudes. The pair of individuals with lower difference had an amplitude of 1 and an entropy of 1.07 *nat* after transforming the data.

4.10 Attachments

This part contains supplementary material of Online resource 1 and 2.

4.10.1 Online resource 1

Table 4.S1 Linear correlations of parameters predictive ability ($r_{y\hat{y}}$) and bias (β_1) with mutual information and kinship coefficient means between training and validation sets ($I_{Tr.\times Val.}$ and $A_{Tr.\times Val.}$, respectively), entropy within validation and training sets ($H_{Val.}$ and $H_{Tr.}$, respectively) for 20-fold cross-validation for all methods with *P. taeda* data. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

Methods	Parameters	$I_{Tr.\times Val.}$	$A_{Tr.\times Val.}$	$H_{Val.}$	$H_{Tr.}$
1	$r_{y\hat{y}}$	0.58*	-0.31	-0.04	0.05
1	β_1	0.4	-0.23	0.39	-0.36
2	$r_{y\hat{y}}$	0.4	-0.26	0.68*	-0.25
2	β_1	0.43	-0.29	0.64*	-0.18
3	$r_{y\hat{y}}$	0.15	-0.42	0.22	-0.17
3	β_1	0.21	-0.35	0.45	-0.4
4	$r_{y\hat{y}}$	0.31	0.3	0.55	-0.33
4	β_1	0.48	0.35	0.68*	-0.36
5	$r_{y\hat{y}}$	0.63*	-0.42	0.79*	-0.63*
5	β_1	0.9*	-0.66*	0.88*	-0.92*
6	$r_{y\hat{y}}$	0.06	-0.15	0.33	-0.12
6	β_1	0.3	0.04	0.7*	-0.53
7	$r_{y\hat{y}}$	-0.04	-0.11	0.8*	-0.42
7	β_1	0	0.05	0.7*	-0.62*
8	$r_{y\hat{y}}$	-0.15	-0.16	0.55	-0.47
8	β_1	-0.31	-0.31	0.75*	-0.66*

* $p < 0.05$

Table 4.S2 Predictive ability ($r_{y\hat{y}}$), bias (β_1), minimum squared error of prediction ($MSEP$), and genomic heritability (h_g^2) for the methods and k-fold cross-validation schemes (Folds) with simulated data. Values in parentheses show standard error due to k-fold cross-validation. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

Methods	Folds	$r_{y\hat{y}}$	β_1	$MSEP$	h_g^2
1	5	0.55 (0.02)	1.02 (0.04)	78.31 (1.72)	0.47 (0.0075)
	10	0.55 (0.02)	1.01 (0.05)	77.58 (3.9)	0.47 (0.0036)
	15	0.55 (0.03)	1 (0.05)	77.4 (3.19)	0.47 (0.0026)
	20	0.55 (0.02)	1.01 (0.05)	78.08 (2.94)	0.47 (0.0026)
2	5	0.45 (0.04)	1.05 (0.09)	82.1 (6.2)	0.45 (0.0379)
	10	0.45 (0.02)	1.06 (0.09)	84.17 (3.93)	0.47 (0.0045)
	15	0.44 (0.04)	0.98 (0.11)	83.81 (2.89)	0.47 (0.0034)
	20	0.25 (0.08)	0.59 (0.22)	87.1 (4.12)	0.47 (0.0032)
3	5	0.54 (0.01)	1.05 (0.1)	79.98 (2.88)	0.47 (0.0136)
	10	0.55 (0.02)	1.02 (0.07)	78.22 (2.71)	0.47 (0.0068)
	15	0.56 (0.02)	1.02 (0.05)	77.48 (3.47)	0.47 (0.0042)
	20	0.56 (0.02)	1.03 (0.05)	77.13 (3.91)	0.47 (0.0028)
4	5	0.53 (0.02)	1.1 (0.07)	78.61 (1.53)	0.46 (0.0086)
	10	0.51 (0.02)	1 (0.04)	79.01 (2.16)	0.47 (0.0042)
	15	0.45 (0.03)	1.01 (0.08)	81.2 (3.66)	0.47 (0.004)
	20	0.47 (0.02)	1.03 (0.06)	79.99 (3.38)	0.47 (0.0025)
5	5	0.42 (0.12)	1.17 (0.67)	95.23 (43.07)	0.46 (0.0472)
	10	0.43 (0.08)	0.91 (0.31)	85.77 (26.9)	0.47 (0.0115)
	15	0.43 (0.07)	0.86 (0.23)	81.87 (20.59)	0.47 (0.0059)
	20	0.43 (0.06)	0.84 (0.18)	80.96 (16.6)	0.47 (0.0039)
6	5	0.32 (0.08)	0.86 (0.62)	110.11 (42.3)	0.45 (0.0596)
	10	0.3 (0.06)	0.58 (0.24)	94.92 (22.79)	0.47 (0.0118)
	15	0.29 (0.04)	0.52 (0.16)	90.35 (17.82)	0.47 (0.0065)
	20	0.34 (0.04)	0.58 (0.15)	85.98 (14.96)	0.47 (0.0046)
7	5	0.3 (0.11)	0.89 (0.77)	109.77 (45.72)	0.46 (0.0493)
	10	0.25 (0.07)	0.34 (0.18)	93.62 (20.93)	0.47 (0.0124)
	15	0.22 (0.06)	0.26 (0.14)	86.75 (16.82)	0.47 (0.0066)
	20	0.24 (0.04)	0.21 (0.08)	84.93 (15.04)	0.47 (0.0044)
8	5	0.45 (0.08)	1.08 (0.57)	94.93 (38.31)	0.46 (0.0383)
	10	0.48 (0.07)	1 (0.24)	83.14 (14.83)	0.47 (0.0093)
	15	0.53 (0.04)	1.08 (0.16)	81.61 (10.16)	0.47 (0.0036)
	20	0.53 (0.04)	1.01 (0.11)	80.65 (7.22)	0.47 (0.0021)

Table 4.S3 Predictive ability ($r_{y\hat{y}}$), bias (β_1), and genomic heritability (h_g^2) for the methods and k-fold cross-validation schemes (Folds) with *P. taeda* data. Values in parentheses show standard error due to k-fold cross-validation. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

Methods	Folds	$r_{y\hat{y}}$	β_1	h_g^2
1	5	0.45 (0.0333)	1.06 (0.0816)	0.42 (0.0201)
	10	0.45 (0.0149)	1.02 (0.0444)	0.42 (0.0065)
	15	0.45 (0.0294)	1.03 (0.073)	0.42 (0.0039)
	20	0.44 (0.0278)	0.99 (0.0646)	0.42 (0.0034)
2	5	0.32 (0.0678)	1.01 (0.2284)	0.43 (0.006)
	10	0.34 (0.0427)	1.04 (0.131)	0.42 (0.0063)
	15	0.27 (0.0596)	0.89 (0.2011)	0.42 (0.0038)
	20	0.25 (0.0599)	0.83 (0.2097)	0.41 (0.0036)
3	5	0.43 (0.0372)	1 (0.0676)	0.42 (0.0156)
	10	0.45 (0.0296)	1.02 (0.0599)	0.42 (0.007)
	15	0.45 (0.0202)	1.03 (0.0531)	0.41 (0.0038)
	20	0.45 (0.0259)	1.03 (0.0705)	0.42 (0.0036)
4	5	0.39 (0.0211)	1.04 (0.0933)	0.42 (0.0132)
	10	0.31 (0.0458)	0.91 (0.1161)	0.42 (0.006)
	15	0.29 (0.0452)	0.87 (0.1393)	0.42 (0.0037)
	20	0.24 (0.047)	0.8 (0.1684)	0.41 (0.0029)
5	5	0.41 (0.0604)	1.06 (0.4373)	0.43 (0.024)
	10	0.42 (0.0471)	0.97 (0.2497)	0.42 (0.0157)
	15	0.42 (0.0442)	0.95 (0.199)	0.42 (0.0083)
	20	0.42 (0.0394)	0.94 (0.1593)	0.41 (0.0041)
6	5	0.21 (0.0662)	0.67 (0.5041)	0.43 (0.0244)
	10	0.23 (0.039)	0.55 (0.1424)	0.42 (0.0118)
	15	0.25 (0.0352)	0.65 (0.127)	0.42 (0.009)
	20	0.23 (0.0416)	0.57 (0.1268)	0.42 (0.0042)
7	5	0.28 (0.0652)	0.56 (0.3526)	0.41 (0.025)
	10	0.23 (0.0445)	0.38 (0.1815)	0.41 (0.0078)
	15	0.25 (0.0471)	0.41 (0.1338)	0.42 (0.0049)
	20	0.26 (0.0395)	0.4 (0.1291)	0.41 (0.0025)
8	5	0.44 (0.0656)	1.08 (0.3546)	0.41 (0.0143)
	10	0.38 (0.0415)	0.98 (0.2035)	0.42 (0.0081)
	15	0.42 (0.0425)	1.17 (0.1614)	0.42 (0.0045)
	20	0.42 (0.0365)	1.2 (0.1143)	0.42 (0.0037)

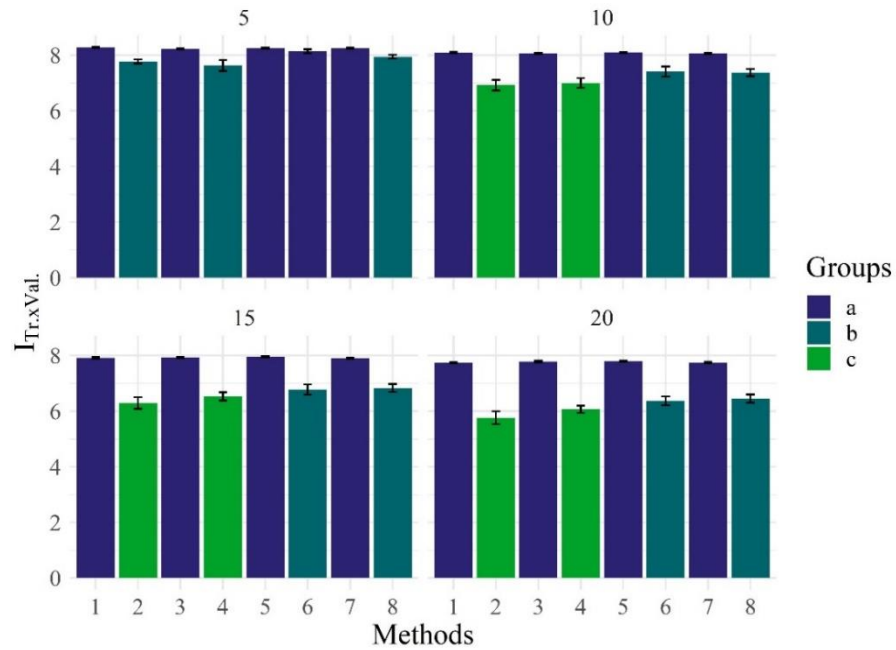


Fig. 4.S1 Average of mutual information between the training and validation sets ($I_{Tr,xVal.}$) for all the methods for 5 to 20-fold cross-validation schemes with *P. taeda* data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

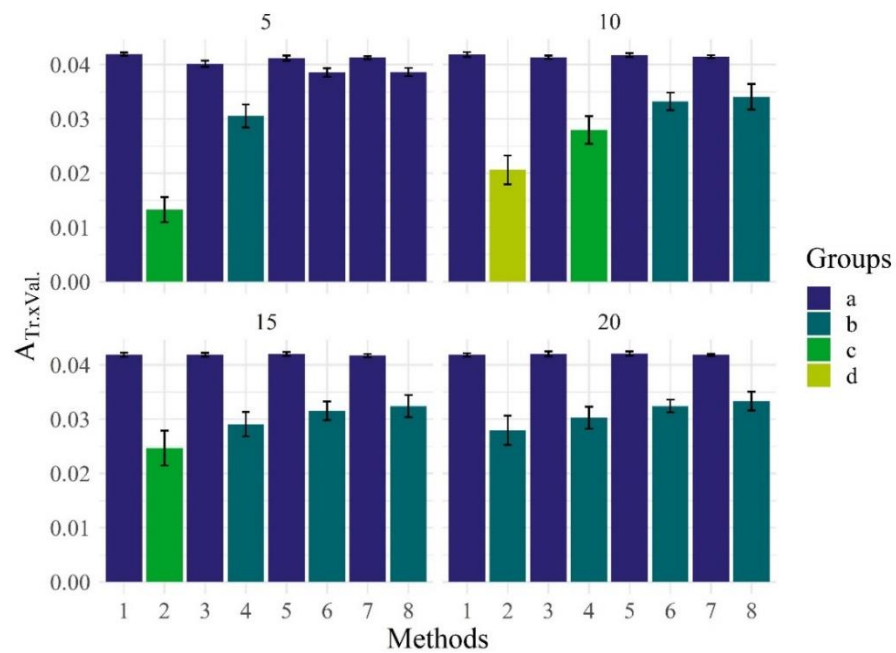


Fig. 4.S2 Average of Wright kinship coefficients between training and validation sets ($A_{Tr,xVal.}$) for all the methods for 5 to 20-fold cross-validation schemes with *P. taeda* data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

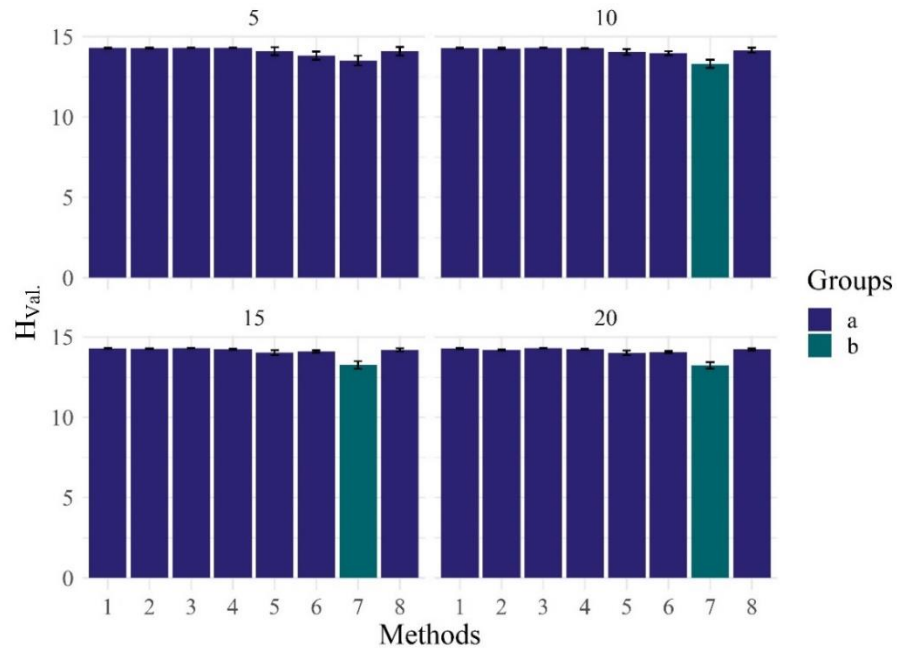


Fig. 4.S3 Average of phenotypic entropy within validation sets (H_{val}) for all the methods for 5 to 20-fold cross-validation schemes with *P. taeda* data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

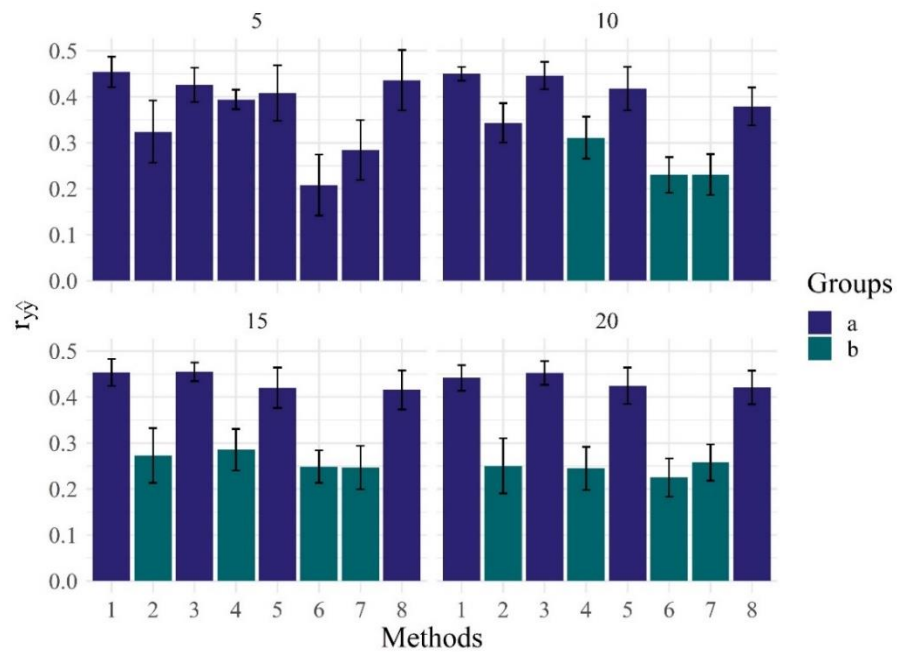


Fig. 4.S4 Average of predictive ability ($r_{\hat{y}}$) for all the methods for 5 to 20-fold cross-validation schemes with *P. taeda* data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

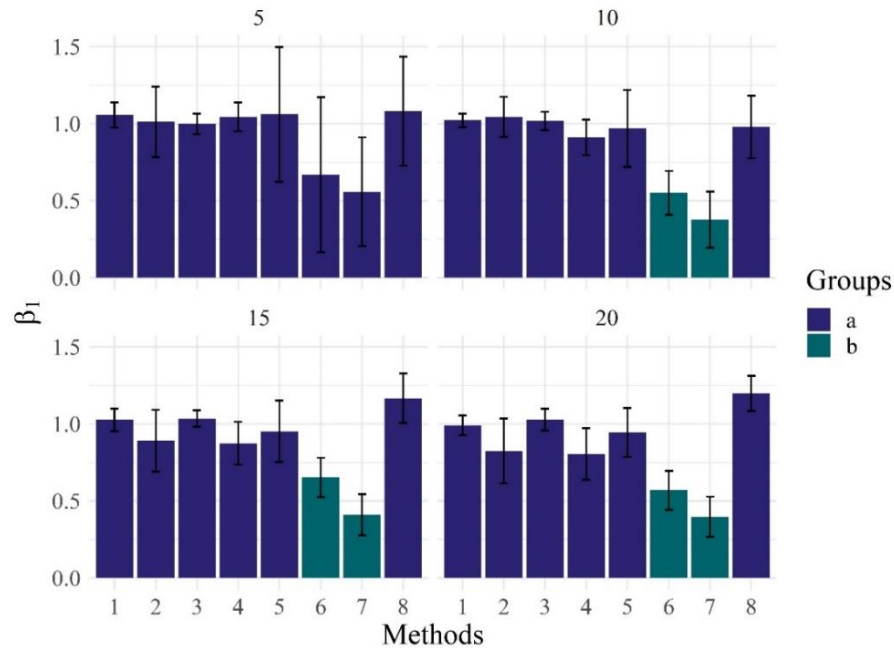


Fig. 4.S5 Average of bias (β_1) for all the methods for 5 to 20-fold cross-validation schemes with *P. taeda* data. Methods of the same group did not differ according to a Scott-Knott test with 1% probability of type I error for each k-fold cross-validation scheme. Bars represent standard error. Methods 1 (random), 2 (K-means), 3 (1/I), 4 (I), 5 (H/I), 6 (I/H), 7 (1/(HI)), and 8 (HI).

4.10.2 Online resource 2

This Online resource 2 contains all the scripts of the novel methods and an example of 100 individuals and 500 markers.

Script of the novel methods with an example at the end:

Observation: each method returns a list with the individuals' code for each validation set

```
library(infotheo)
```

```
marc<-read.table("example_markers.txt", header = T)
```

```
dat<-read.table("example_pheno.txt", header = T)
```

```
ind<-as.character(dat[,1]) ## Individuals' code
```

```
fenx<-scale(dat[,2]) ## Phenotype data
```

```
marc<-as.matrix(marc)
```

Transformation of phenotypic data and avoidance of negative entropy:

```
fen1<-fenx[order(fenx)]
```

```
tte<-NULL
```

```
for(i in 1:length(fen1)){tte[i]<-fen1[i+1]-fen1[i]}
```

```
tte[tte[]==0]<-NA
```

```
f<-1/min(abs(tte), na.rm = T)
```

```
fen<-f*fenx
```

```
row.names(fen)<-ind
```

```

## Phenotypic entropy matrix between all individuals:
Hf<-matrix(ncol = length(fen), nrow = length(fen))
system.time(for(i in 1:nrow(Hf)){
  for(j in 1:ncol(Hf)){
    Hf[i,j]<-0.5*log(2*pi*exp(1)) + log(sd(c(fen[i],fen[j])))
    if(is.na(Hf[i,j])){Hf[i,j]<-NA }else{if(Hf[i,j]<0){Hf[i,j]<-0}}
  }
})
colnames(Hf)<-ind; row.names(Hf)<-ind

## Mutual information matrix between all individuals:
gen<-as.data.frame(t(marc))
colnames(gen)<-ind
Ix<-mutinformation(gen)
row.names(Ix)<-ind; colnames(Ix)<-ind; diag(Ix)<-NA

```

```

# Example for 20-fold cross-validation:
### Method 3 (1/I)
tr_val<-list()
tv<-table(rep_len(1:20, length(ind))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds
  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)
  xhi<-1/irow # Index of method 3: 1/I
  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index
  for(j in 1:tv[k]){
    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals
    res<-matrix(nrow = length(namex), ncol = 2)
    for(i in 1:length(namex)){
      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed
    }
    row.names(res)<-namex
    indice<-as.matrix(1/res[,1]); row.names(indice)<-namex # Index of method 3: 1/I
    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index
    if(length(xval)==tv[k]){break()}
  }
  tr_val[[k]]<-xval
  ind<-subset(ind, subset = !ind%in%xval)
}
save(tr_val,file ="meth_3_tr_val_20fold.Rdata")

```

```

### Method 4 (I)

tr_val<-list()

ind<-as.character(dat[,1])

tv<-table(rep_len(1:20, length(dat[,1]))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds

  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)

  xhi<-irow # Index of method 4: I

  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index

  for(j in 1:tv[k]){

    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals

    res<-matrix(nrow = length(namex), ncol = 2)

    for(i in 1:length(namex)){

      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed

    }

    row.names(res)<-namex

    indice<-as.matrix(res[,1]); row.names(indice)<-namex # Index of method 4: I

    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index

    if(length(xval)==tv[k]){break()}

  }

  tr_val[[k]]<-xval

  ind<-subset(ind, subset = !ind%in%xval)

}

save(tr_val,file ="meth_4_tr_val_20fold.Rdata")

```

```

### Method 5 (H/I)

tr_val<-list()

ind<-as.character(dat[,1])

tv<-table(rep_len(1:20, length(dat[,1]))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds

  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)

  hrow<-apply(Hf[ind,ind],1,max, na.rm = T)

  xhi<-hrow/irow # Index of method 5: H/I

  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index

  for(j in 1:tv[k]){

    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals

    res<-matrix(nrow = length(namex), ncol = 2)

    for(i in 1:length(namex)){

      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed

      res[i,2]<-0.5*log(2*pi*exp(1))+log(sd(c(fen[xval,],fen[namex[i,]],na.rm = T))) # Phenotypic entropy
between each individual and the already group formed

    }

    row.names(res)<-namex

    indice<-as.matrix(res[,2]/res[,1]); row.names(indice)<-namex # Index of method 5: H/I

    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index

    if(length(xval)==tv[k]){break()}

  }

  tr_val[[k]]<-xval

  ind<-subset(ind, subset = !ind%in%xval)

}

save(tr_val,file ="meth_5_tr_val_20fold.Rdata")

```

```

### Method 6 (I/H)

tr_val<-list()

ind<-as.character(dat[,1])

tv<-table(rep_len(1:20, length(dat[,1]))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds
  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)
  hrow<-apply(Hf[ind,ind],1,max, na.rm = T)
  xhi<-irow/hrow # Index of method 6: I/H

  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index
  for(j in 1:tv[k]){
    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals

    res<-matrix(nrow = length(namex), ncol = 2)

    for(i in 1:length(namex)){
      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed

      res[i,2]<-0.5*log(2*pi*exp(1))+log(sd(c(fen[xval,],fen[namex[i,]],na.rm = T))) # Phenotypic entropy
between each individual and the already group formed

    }

    row.names(res)<-namex

    indice<-as.matrix(res[,1]/res[,2]); row.names(indice)<-namex # Index of method 6: I/H

    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index

    if(length(xval)==tv[k]){break()}

  }

  tr_val[[k]]<-xval

  ind<-subset(ind, subset = !ind%in%xval)

}

save(tr_val,file ="meth_6_tr_val_20fold.Rdata")

```

```

### Method 7 (1/(HI))

tr_val<-list()

ind<-as.character(dat[,1])

tv<-table(rep_len(1:20, length(dat[,1]))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds

  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)

  hrow<-apply(Hf[ind,ind],1,max, na.rm = T)

  xhi<-1/(irow*hrow) # Index of method 7: 1/(HI)

  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index

  for(j in 1:tv[k]){

    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals

    res<-matrix(nrow = length(namex), ncol = 2)

    for(i in 1:length(namex)){

      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed

      res[i,2]<-0.5*log(2*pi*exp(1))+log(sd(c(fen[xval,],fen[namex[i,]],na.rm = T))) # Phenotypic entropy
between each individual and the already group formed

    }

    row.names(res)<-namex

    indice<-as.matrix(1/(res[,1]*res[,2])); row.names(indice)<-namex # Index of method 7: 1/(HI)

    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index

    if(length(xval)==tv[k]){break()}

  }

  tr_val[[k]]<-xval

  ind<-subset(ind, subset = !ind%in%xval)

}

save(tr_val,file ="meth_7_tr_val_20fold.Rdata")

```

```

### Method 8 (HI)

tr_val<-list()

ind<-as.character(dat[,1])

tv<-table(rep_len(1:20, length(dat[,1]))) # Number of individuals per validation set

for(k in 1:20){ # Number of folds

  irow<-apply(Ix[ind,ind], 1,max, na.rm = T)

  hrow<-apply(Hf[ind,ind],1,max, na.rm = T)

  xhi<-irow*hrow # Index of method 8: HI

  xval<-names(xhi[order(xhi, decreasing = T)][1:2]) # Selection of a pair of individuals with maximum
index

  for(j in 1:tv[k]){

    namex<-subset(ind, subset = !ind%in%xval) # Unselected individuals

    res<-matrix(nrow = length(namex), ncol = 2)

    for(i in 1:length(namex)){

      res[i,1]<-multiinformation(gen[,c(xval,namex[i])]) # Redundancy between each individual and the
already group formed

      res[i,2]<-0.5*log(2*pi*exp(1))+log(sd(c(fen[xval,],fen[namex[i,]],na.rm = T))) # Phenotypic entropy
between each individual and the already group formed

    }

    row.names(res)<-namex

    indice<-as.matrix(res[,1]*res[,2]); row.names(indice)<-namex # Index of method 8: HI

    xval<-c(xval,names(indice[order(indice, decreasing = T),][1])) # Selection of one individual with
maximum index

    if(length(xval)==tv[k]){break()}

  }

  tr_val[[k]]<-xval

  ind<-subset(ind, subset = !ind%in%xval)

}

save(tr_val,file ="meth_8_tr_val_20fold.Rdata")

```

Data of 100 individual phenotypes ("example_pheno.txt")

ID y

1a1 506.39203182031
2a1 514.415108010212
3a1 506.40685376676
4a1 519.661606643678
5a1 497.899711164983
6a1 499.356261188138
7a1 501.5906527504
8a1 503.817387320537
9a1 519.462469627809
10a1 499.513173439212
11a1 514.065353413041
12a1 495.809784154707
13a1 510.52472870986
14a1 513.192884089667
15a1 508.104867486338
16a1 518.341774458627
17a1 522.510146834495
18a1 502.991958888806
19a1 512.698013731191
20a1 525.200347653088
21a1 505.62283264832
22a1 504.871342725013
23a1 519.193557442468
24a1 532.736961063311
25a1 505.816194390675
26a1 508.572833110366
27a1 502.99625485781
28a1 509.882855732194
29a1 516.187134043057
30a1 487.464685953003
31a1 512.682668058766
32a1 513.022622956799
33a1 504.456956513298
34a1 510.588964597378
35a1 518.721997231853

36a1 512.310689757253
37a1 528.855159940384
38a1 519.296638694361
39a1 525.414076281678
40a1 514.698761153152
41a1 517.923251994135
42a1 514.681088819305
43a1 486.367161007361
44a1 516.275595502001
45a1 506.064642986979
46a1 504.336256564195
47a1 520.227449224521
48a1 523.49565045997
49a1 513.14984915391
50a1 500.545442329666
51a1 507.747335848109
52a1 519.908149888018
53a1 502.241250158568
54a1 503.632336899673
55a1 519.3886046303
56a1 509.925553179667
57a1 510.560295084355
58a1 521.299700606913
59a1 526.182481016663
60a1 509.874475548006
61a1 510.045047698071
62a1 502.889628086925
63a1 518.829576060196
64a1 517.043772555627
65a1 527.327374493183
66a1 505.79411993887
67a1 519.393784245806
68a1 503.632213986815
69a1 514.915295549035
70a1 503.211461785901
71a1 497.341778134059
72a1 504.596487963717

73a1 510.872552430744
74a1 515.935683273935
75a1 501.208418148218
76a1 508.74925978762
77a1 506.366857682307
78a1 512.521432092032
79a1 504.215168382528
80a1 507.486452065121
81a1 490.48293093472
82a1 483.995685073314
83a1 506.365028181779
84a1 511.905387769774
85a1 512.168094663061
86a1 524.118150966584
87a1 517.033607193885
88a1 515.84639609021
89a1 528.355431302474
90a1 512.28295377115
91a1 494.698725406485
92a1 532.635267747011
93a1 518.028984286095
94a1 503.321482322631
95a1 525.056697606507
96a1 521.833796259268
97a1 517.254409677745
98a1 507.098803135579
99a1 522.292603673568
100a1 510.321387860347

Data of 500 markers for 100 individuals ("example_markers.txt").

X51 X54 X56 X63 X95 X157 X219 X224 X231 X237 X263 X314 X348 X365 X371 X387 X391 X422
X443 X537 X540 X545 X549 X560 X563 X618 X769 X830 X852 X865 X866 X894 X914 X931 X948
X950 X973 X999 X1003 X1030 X1091 X1128 X1160 X1190 X1212 X1213 X1242 X1363 X1406 X1415
X1446 X1466 X1553 X1648 X1652 X1699 X1734 X1737 X1745 X1785 X1786 X1804 X1811 X1851
X1892 X1979 X1980 X2015 X2041 X2180 X2186 X2206 X2265 X2338 X2339 X2340 X2343 X2344
X2345 X2347 X2357 X2367 X2380 X2404 X2409 X2433 X2483 X2484 X2494 X2527 X2537 X2540
X2564 X2575 X2610 X2611 X2612 X2613 X2642 X2712 X2729 X2731 X2735 X2737 X2739 X2740
X2741 X2746 X2747 X2749 X2751 X2752 X2754 X2762 X2804 X2870 X2879 X2904 X2924 X2927
X2978 X2979 X2982 X2983 X3001 X3034 X3058 X3072 X3110 X3133 X3152 X3174 X3199 X3292
X3321 X3331 X3355 X3371 X3425 X3456 X3460 X3463 X3464 X3529 X3538 X3618 X3626 X3632
X3633 X3635 X3638 X3640 X3641 X3664 X3666 X3667 X3668 X3670 X3672 X3673 X3674 X3690
X3698 X3706 X3708 X3710 X3724 X3740 X3741 X3816 X3849 X3881 X3887 X3929 X3930 X4039
X4042 X4049 X4071 X4073 X4101 X4102 X4122 X4216 X4266 X4353 X4368 X4371 X4400 X4427

X4435 X4521 X4531 X4552 X4563 X4578 X4600 X4666 X4668 X4763 X4764 X4810 X4819 X4841
 X4868 X4885 X4889 X4940 X5073 X5092 X5096 X5141 X5147 X5184 X5186 X5190 X5191 X5194
 X5197 X5245 X5250 X5251 X5253 X5254 X5256 X5257 X5271 X5273 X5287 X5320 X5324 X5346
 X5347 X5348 X5349 X5350 X5351 X5357 X5358 X5366 X5368 X5369 X5371 X5372 X5373 X5374
 X5375 X5378 X5384 X5385 X5386 X5387 X5390 X5391 X5392 X5393 X5408 X5409 X5429 X5433
 X5434 X5482 X5538 X5568 X5609 X5637 X5668 X5670 X5746 X5749 X5795 X5837 X5854 X5858
 X5899 X5901 X5960 X5989 X5993 X5997 X6044 X6046 X6050 X6109 X6159 X6210 X6219 X6270
 X6294 X6303 X6341 X6364 X6391 X6435 X6469 X6483 X6486 X6495 X6501 X6507 X6546 X6558
 X6594 X6603 X6609 X6647 X6671 X6818 X6852 X6950 X6952 X6976 X7062 X7083 X7097 X7113
 X7256 X7272 X7286 X7294 X7314 X7347 X7355 X7376 X7422 X7427 X7428 X7439 X7487 X7497
 X7508 X7521 X7531 X7540 X7556 X7575 X7596 X7616 X7633 X7698 X7723 X7725 X7763 X7782
 X7788 X7790 X7808 X8010 X8021 X8027 X8058 X8137 X8165 X8197 X8243 X8262 X8277 X8293
 X8297 X8319 X8333 X8339 X8360 X8375 X8376 X8377 X8380 X8381 X8383 X8387 X8389 X8390
 X8391 X8393 X8395 X8396 X8397 X8398 X8400 X8421 X8433 X8441 X8462 X8474 X8501 X8515
 X8558 X8635 X8654 X8727 X8731 X8748 X8760 X8806 X8821 X8869 X8954 X8956 X8996 X9051
 X9052 X9053 X9054 X9058 X9078 X9098 X9099 X9100 X9101 X9103 X9105 X9158 X9285 X9301
 X9441 X9451 X9476 X9481 X9483 X9484 X9488 X9489 X9494 X9559 X9581 X9618 X9662 X9689
 X9695 X9723 X9745 X9765 X9822 X9832 X9865 X9872 X9929 X10054 X10057 X10063 X10092
 X10096 X10105 X10173 X10192 X10201 X10209 X10278 X10334 X10399 X10416 X10428 X10451
 X10494 X10511 X10573 X10574 X10588 X10591 X10624 X10649 X10674 X10699 X10714 X10715
 X10784 X10803 X10815 X10876 X10902 X10942 X10990 X11020 X11034 X11045 X11048 X11104
 X11141 X11159 X11161 X11205 X11209 X11246 X11323 X11387 X11392 X11414 X11444 X11455
 X11483 X11532 X11589 X11614 X11615 X11634 X11680 X11779 X11831 X11861 X11909

002200111100001110002210200200002010201002112200000022202
 122011022022221000202020002001020022222211111200200221
 00122201122112020220201000020202111111210222222000011112
 20001120021020110002001212000002020220112022011101111120
 020020222111222000000222111102122121200000010002201220022
 002100210121011202222110220200222002000202002022122002202
 000020210221022021120000022002000222102222120010012200100
 221022211111021101111111021202202010111112001110100200012
 00000020011021012022201111000100100000101000

002110010111002000002220200210012020100001111200100022212
 0220100111222220002020200020020200222221211220200200220
 00022201122112121220201100020202111111210222222000011112
 211101200210201100020022020000020202100020111020021111120
 0100202222211111111111002211012221200110021111201220021
 002000210020021202222210210210221002000211002022122002202
 000020210221022021120000022002000222102222120000022200000
 220022202000022200200222021201101000220002012100000110002
 00010010010011012022202020000100000000111010

002110010111002000002220200210111120201001112200100022202
 02201102201222210002020200020010101222222211220200210220
 00021102122112121220201000020202111111210222222011102011
 211101200210200010021012120000122202200020220011002022221
 01002022222211111111111002220002222211100020002210110122
 002201010021010112212211221220222001110220002121212002201
 111021110211012111210000022002000222202212220010002110000
 220022211111021200200222011201101010120002001110000110002
 00000020010020011012101111000100100000101110

11110102000001111112220200210012020201001211200010012202
 122010021112222100020202000200101011211212111220200200220
 100222120222120202212020000201020202111210222222000011112
 200011200210201100021012020000121202200020220020020200020
 111111111112222000000222112211012221200010020002200210122
 012000120020011112212110220200221002000202012012022002102
 00002021022202202012000002200200022210222220000012020000
 220022211111011100200222012200111010121112001111000200012
 00000020010020011102100202000000200000101000

002110010111002000002220200200101120201002002100000011202
 12200111112222100020202000210101012211202002111200200221
 10112211122102111121202000020112111122010111112000020222
 20001120021021110002101202000012120220002022001101111121
 01002022222111111111111002211012221210100020002100110121
 002100100022000112222121221220212101110220012012122002201
 10002012022201201012000002200200022210222220000012110000
 22002221111021200200222011201101010120002002200000200002
 00000020020021001012101112000010200000200010

111100111101002000112221100210012020100001111200100022212
 02201001112222200020200020020200222221211220200200220
 0002201122112121222020110002020211111121022222000011112
 211101200210201100020022020000020202100020111020021111120
 01002022222111111111111002211012221200110021111201220021
 002100100021010112222210210210221002000211002022122002202
 000020210221022021120000022002000222102222120000022200000
 220022202000022200200222021201101000220002012100000110002
 00000020010021002012212020000200100000101000

002200020000002000002220200210012010200001111201001022202
 02201001111222210002020200020010101222222111220200200220
 000222120222120202212020101111020202111210222222000011112
 200011100210200000011012120011020202210020220111020200020
 01002022222111111111111002202022221200000020002100110121
 002000210020011112212110220200222002000211102022122002202
 000020210221022021120000022002000222102222120010012200011
 220021102000022200200222011202102010111112002100000110002
 00100010111121012022201111000100000001102000

111101101211002000002220200210011120202002102200000012202
 122001111122222100020202000210101012222212111220200200220
 100211121221021212222020101111011111220210222222000010112
 111122200201212200022002110000121212200010220102002022221
 000020222222000222222000002220002222210100020002100110121
 1022000000221000222222221212202120011102100020222222002211
 100020110222012011220000022002000222202222220000002120000
 110111220222020200200222012200111020020002002200000200002
 00000021020020011012101112000010200000111001

111100111100012110112221110210012020100001111200000022202
 122001011122121100020202000210101011111202102220200200220
 100211121221021212212020001111121111220200222222000020222
 211122100200212201022002020000122202200020220102112022221
 01002022222111111111111002211012222210100020002200220021
 102100110020021212212110220111221002000211002022122002202
 000021210211022021110000022002000222102221220000012110000
 22102221111102110111111021202202001221112001100000200002
 00101010010010000002000202000100200000101010

111100111100012000102220200220021120200001112200000021202
 021011121112222111111111011200102001211211211220200200220
 000211120222120202212020100201020202002220222222011102011
 211101200210201100021012120000021202200020220111020200020
 0200202222222000000222002020222020200000020002100110122
 002001211200111122121112202102210020002002022122002202
 001021210221022020120000022002000222102222220000012110000
 220022202000021200200222021202202010111112002100000200002
 00000020020021112021202020000100100000102000

002200111100001110002210200200002010201002002200000012202
 222001012022222000020202000210001012222212102111200200221
 101222111221020202212010000201021111220200222222000020222
 200022200200212200021002120000121202201120220111011111120

020020222111222000000222111102122121200000010002201220022
002100210121011202222110220200222002000202002022122002202
000020210221022021120000022002000222102222120010012200000
11011121111021200200222011201212110121102101111000200012
00000020010021012112201111000100100000102000

111100202200002000112221100200002020101002001200100012212
122000001122222100020202000210101012222211102220200200220
100222111221021212212011000201021111220200222222000020222
211112200200212200021012020000121202100020111011012022221
00002022222000222222000002220002222210210021111201220021
002100100021010112222221211220221002010220002022222002201
10002011022101201012000002200200222202222220000012110000
22002121022202021111111002200111020011112002100000110002
0000001011120011012100211000010100000110010

002110010111002000002220200210111120201001112200100022202
022011022022222100020202000211112101211112220220200200210
000211021221121212211120000202120202111210222222000011112
200022200200211100022002020000112202200020220011002022221
01002022222211111111111002220002222211100020002210110122
002201010021010112212211221220222001110220002121212002201
111021110211012111210000022002000222202212220010002110000
220022211111021200200222011201101010120002001110000110002
0000002001002001101210202000020000000102110

002101101211001000112221100200002020101002101200010002202
222000011112222000020202000210000021211202002220200200220
20022220222020202222020000200020202220200222222000020222
20002220020021220002200202000022220220002022000200202222
00002022222000222222000002220002222210100020002200220022
002200010021000112222121221210221002010211012012122002101
10002011022201201022000002200200022220222220000012020000
220022211111021201111111002201001020011112002100000110002
0000002001120000002000202100000100000110000

002110010111002000002220200210111120200001112100000021202
022011121112222200020202000200202002211212111111200200221
00112201122112111120202000020212111111121022222011102012
211111101110200000021012020000020202200020220111020200020
010020222222111111111111002211012221201000020002200220122
012000120010011202212210220210221002000211002022122001202
000020210222022021120000022002000222102222220000012211000
220022202000022200200222020201101010120002002200000110002
00000020020021001012101112000010200000200010

111100202200002000112221100200002020100001111200100022212
02201001112222220002020200020020200222221211220200200220
000222011221121212202011000202021111111210222222000011112
211101200210201100020022020000020202100020111020021111120
010020222222111111111111002211012221200110021111201220021
002000210020021202222210210210221002000211002022122002202
000020210221022021120000022002000222102222220000022200000
220022202000022200200222021201101000220002012100000110002
00000020010021012112201111000100100001102001

002200020000002000002220200210012010200001111201001022202
02201001111222210002020200020010101222222111220200200220
000222120222120202212020101111020202111210222222000011112
200011100210200000011012120011020202210020220111020200020
010020222222111111111111002202022221200000020002100110121
002100100021000022212121221210222002010220102022222002201
100020110221012011220000022002000222202222120010002110011

22002111111021200200222002201001020011112002100000110002
00100010110010011012100202000100100000201000

002101010111002000002220200110012020201002002200010002202
222000001122222000020202000210001011211201102220200200220
1002110212211212122120201011120111111122022222000001002
111111200211201100021012110000020212200010220111011111120
01002022222111111111111002211012221200000020002100110121
10210011002111111222211120210212001100201002022122002212
00002021022102202112000002200200022210222220000012210000
110111211111021200200222021201212010120002002200000200002
00000021020020011012101112000010200000111000

002101101211002000002220200211011120201002101200010002202
22200101202222000020202000221011111211102111220200200210
100211121122020202211120000201021202111101111112000021122
200022200200212200022002120100121202200020220102112022221
0000202222200022222000002211012222210110020002200210122
002000220020021202222200220200222002000211012012022002202
00002021022202202012000002200200022210222220000112211000
220022202000021101111111012201101020011112001110100110002
00010010010010011012101112000010200000101010

002110010100012000102220200221021120200001211200010012202
12201102202222100020202000211112101211112220220200200210
00021102122212020220112000020212111111121022222011102012
200011200210200010021011120000020212201120220111020200020
02002022222222000000222002202022220200000020002100110122
012000120010011202212210221220212001110220012012122002201
1001201202220010102201111110111111120222220000012020000
220022211111020200200222002200111010121112001110000110002
00000010111120011012100202000010100000200000

111101010111002000101120200220012020100001112200000022212
022010011122122100020202000200101011211212111220200200220
00022202121212020220112000020212111111121022222011102012
200011200210200000021012020011020202200020220111121111120
02002022222222000000222002202022220210100121102200220012
00210011002200011122222122122012100201022002022222002201
1000201102210120112200000220020002220222120000011020000
110111211111021200200222011201212110121102101111000200012
00000020010021012112201111000100100000102000

111100111100002000112221100210012020100001111200100022212
122000001122222100020202000210101012222211102220200200220
100222111221021212212011000201021111220200222222000020222
211112200200212200021012020000121202100020111011012022221
000020222222000222222000002220002222210210021111201220021
002100100021021202222210210210221002000211002022122002202
000020210221022021120000022002000222102222120000022200000
220022202000021200200222012200000010120002012100000110002
00010010010010011012101111000000100000100010

111100111100012000102220200220011120210101112200000022202
22200001201222200002020200021000002222212102220200210220
10021112122112121220201000020202111111121022222011102011
211101200210200010021012120000021202200020220020011111120
02002022222222000000222002211012221201000020002210110122
002101120020021202212200220210222001100211002121112002202
011021210211022121110000022002000222102212220010012200000
220022202000022200200222020202202000220002001110000110002
00000020010021012022202020000200000000102110

002101010111001000112221100210012020100001111210000012202
02211102201222220002020210020020200222221211220200200220
000222011221121212202020101112121111111111112000011112
201111100210200000011012120011020202200020220020011111121
010020222221111111111111002211012221200000020002200220022
002100120020011202222110220210212101100211012012022002202
000020210222022121110000022002000222102222120000022211000
220022202000022201111111011202102010111112002100000110002
00000020011121001012101111100100000000111000

002110010111002000102220201220022010200001112200000022202
02202002202222220002020200020020200222222220220200200210
00022201122112020120202000020212111111210222222011102012
211111101110200000021022020000020202200020220020020200020
02002022222222000000222002202022220200000020002100110121
002000210021011202222110220210212101100211012012022002202
000020220222022020020000022002000222002222220000022200000
220022202000022200200222020202202000220002002200000200002
00000020020022002022202021000110100000201010

111100111100002000112221100210012020100001111200100022212
02201001112222220002020200020020200222221211220200200220
000222011221121212202011000202021111111210222222000011112
211101200210201100020022020000020202100020111020021111120
010020222221111111111111002211012221200110021111201220021
00200021002002120222210210210221002000211002022122002202
000020110221012011220000022002000222202222120000012110000
220022211111020200200222002200001020011112001110000200002
10101020011020001002111111000100200000100000

002200111100002000002220200200002010201001112200000012202
1220100120222210002020200021010101222212111220200200220
100222120222120202212020000201020202111210222222000011112
200011200210211100021012020000121202200020220011011111121
010020222221111111111111002211012221210100020002100110121
002100100021000022212121221210222002010220102022222002201
100020110221012011220000022002000222202222120010002110011
220021111111021200200222002201001020011112002100000110002
00100010111120011012100202000000100001101000

111101010111002000002220200220021120201001212200000022202
022011121122222200020202000200102001211211211220200200220
000211021221121212212020101112011111111220222222000001002
111111200211201100021012110000020212200010220111011111120
010020222221111111111111002211012221200000020002100110121
10210011002111111222211120210212001000211002022122002202
000020210221022021120000022002000222102222120010012200000
221022211111021201111111021201102010120002012100000110012
00000010020021012022201120000100100000101001

111100111100012110112221110210012020100001111200000022202
022011021122121200020202000200202001111212211220200200220
00021102122112121220202000111212111111121022222000011112
211111100210201101021012020000021202200020220111121111120
02002022222222000000222002202022221200000020002200220021
102000110020021212212110220111221002000211002022122002202
000021210211022021110000022002000222102221220000012110000
221022211111021101111111022201101011121112001100000200002
00101010010010000002000202000100200000110000

002110101200012000102220200211011120201002101200010002202
222001012022222000020202000221011111211102111220200200210
100211121222020202211120100200020202111210222222011111121
211112200200212200022002120000122202200020111012002022212

0000202222200022222000002220002222211100020002200220122
012100010011000112212221222220212001110220012012122002201
1001201202220010102201111110111111120222220000012020000
220022211111020200200222002200111010121112001110000110002
00000010111120011012100202000010100000200000

111101101211002000101120200210002020101002002200000012212
122000001122122000020202000210000021211202002220200200220
100222121212020202211120000201121111220200222222011111122
20002220020021110002200202001112120220002022010211202221
01002022222111111111111002211012221210100121102201220022
00220010012200011222212122121022200201021100202222002201
10002011022101201122000002200200022202222120010002200100
22102221111102110111111102120220201011112001110100200012
00000020011021012022201111000100100000102000

002110010111002000002220200210111120201001112200000021202
02201112111222211111111011200102001211211211220200200220
00021112022212020220202000020202111200211111112000002002
200001200110200000020122020100020212200020111020121111120
02002022222222000000222002202022220200000121102200220012
001100110021011112222211221210221002000202002022122001202
000010220222022020020000022002000222102222220000022200000
220021201111021211111111011201212010111112002100000110002
00000010111121012022201120000110000000111010

111100111100012000102220200220011120210102002200000012202
222000012012222000020202000210000022222212102220200210220
100211121221021212212010000201021111220200222222011111121
211112200200211110022002120000122202200020220011002022221
010020222222111111111111002220002222211100020002210110122
002201010021010112212211221220222001110220002121212002201
111021110211012111210000022002000222202212220010012200000
220022202000022200200222020202202000120002001110000110002
01010010011020111012101111000100100000101000

111101111100011111112221100200002020101002001210000002202
122101012012222100020202100210101012222211102220200200220
10022211122102121221202010111112111122010111112000020222
201122100200211100012002120011121202200020220020020200020
111111111112222000000222112211012221200010020002200210122
012000120020011112212110220200221002000202012012022002102
000020210222022020120000022002000222102222220000022110000
220022202000012100200222021201212000221112001111000200012
00000020010021012112201111000100100000102000

002110101211002000002220200200101120201002002100000011202
122001111112222100020202000210101012211202002111200200221
101122111221021111212020000201121111220200222222011111122
211122101100211100022002020000121202200020220102011111121
00002022222000222222000002220002222211100020002200220122
012100010011000112212221221220221002010220002022222001201
10002011022201201122000002200200022220222220000012211000
220022202000022200200222010202202000220002001110000110002
00000020010021012022201111000100000000101110

111100111101002000002220200200002020101002001200010002202
22200101202222210002020200021010101222222220220200200220
00022202022222020220202000020202111200211111112000002012
200011200210200010021011120000020202200020111020021111120
010020222222111111111111002211012221200110021111201220021
002000210020021202222210210210221002000211002022122002202
000020210221022021120000022002000222102222120000022200000

220022202000022200200222021201101000220002012100000110002
00000020010021012112201111000100100001102001

002200111100002000002220200200002010201002001201001012202
122000001112222000020202000210000022222212002220200200220
100222220222020202222020101110020202220200222222000020222
200022100200211100012002120011121202210020220102011111121
00002022222000222222000002211012222210100020002100110121
002100100021000022212121221210222002010220102022222002201
100020110221012011220000022002000222202222120010002110011
22002111111021200200222002201001020011112002100000110002
0010001011120011012100202000000100001101000

111101101211002000002220200210011120202002102200000012202
122001111122222100020202000210101012222212111220200200220
201211121221021112222020101111121111220210222222000020221
211112210200211100012002020000112202200010220102002022221
00002022222000222122000002211012222210110020002200210122
002100110021010112222211221220221002010220002022222002201
100020110221012011220000022002000222202222120010002110000
221022211111021201111111021201102010120002012100000110012
00000010020021012022201120000100100000101001

002101101211002000002220200211011120201002101200010002202
222001012022222000020202000221011111211102111220200200210
100211121122020202211120000201021202111101111112000021122
200022200200212200022002120100121202200020220102112022221
000020222222000222222000002211012221200010020002200210122
002000220020021202222200220200222002000211012012022002202
000020210222012010220000022002000222202222220000102121000
220022211111021101111111012201101020011112001110100110002
00010010010010011012101112000010200000110000

111100202200012000102220200210011120201002002200000002202
222001012022222000020202000221011111211102111220200200210
100211121222020202211120000201121111220200222222011111122
200022200200211110022001120000121212201120220102011111121
0100202222221111111111110022110122212101000200022002202122
01210001001100011221222122220212001110220012012122002201
100120120222001010220111111011111112022220000012020000
220022211111020200200222002200111010121112001110000110002
00000010111120011012100202000010100000200000

002200111100001110002210200200002010201002002200000012212
122001012022222000020202000210001012222212102111200200221
101222111221020202212010000202021111220200222222000020222
200022200200212200021002120000121202201120220102002022221
01002022211111111111111111111111112122210100010002201220022
002200100122000112222121221210222002010211002022222002201
100020110221012011220000022002000222202222120010002110100
221022220222020101111111012201101020011112001110100200012
00000020011020011012100202000000200000101000

111100111100002000112221100210012020201001112200000021202
022011121112222111111111011200102001211211211220200200220
000211120222120202202020000202021112002111111112000002002
200001200110200000020122020100020212200020111020121111120
020020222222222000000222002202022220200000121102200220012
001100110021011112222211221210221002000202002022122001202
000010120222012010120000022002000222202222220000012110000
220021210222020211111111002200111020011112002100000110002
00000010111120011012100211000010100000110010

002110101211002000002220200200101120202002002200100012202
 1220010120222200002020200022101111211102111220200200210
 10021112122102121222112000020112020222020022222000020222
 200022200200211100022002020000112202200021111011012022221
 00002022222000222222000002220002222110100020002200220022
 00221100002101012122221221220221002000202002022122001202
 000010211221022021120000022002000222112222120000022200000
 220022202000021200200222011201102010120002002100101110002
 01010010011021112022202020000200000000102000

1111011111000111111222020020000202020200210120000002202
 122101012012222100020202100210101011211202002220200200220
 2002222022202020222020000200020002022022020022222000020222
 20002220020021220002200202000022202200020220002002022222
 0000202222200022222000002220002222210100020002200220022
 002100120020011202222110220210212101100211012012022002202
 000020210222022121110000022002000222102222120000012121000
 22002221111102120111111002201001020011112002100000110002
 00000020011120000002000202100000100000110000

002110101211002000102220201210012010201002002200000012202
 122010012022222100020202000210101012222212002111200200221
 101122111221021111212020000201121111220200222222011111122
 211122101100211100022002020000121202200020220102011111121
 00002022222000222222000002220002222211100020002200220122
 012100010011000112212221221220221002010220002022222001201
 100020120222012010120000022002000222102222220000012110000
 220022211111021200200222011201101010120002002200000200002
 00000020020021001012101112000010200000200010

111100111100002000112221100210012020100001111200010012202
 12201102202222210002020200021010101222222220220200200220
 00022202022222020220202000020202111200211111112000001122
 200022200200211110022001120000121202200020111011012022221
 00002022222000222222000002220002222210210021111201220021
 002100100021010112222221211220221002010220002022222002201
 10002011022101201022000002200200022202222220110012110000
 220022211111021200200222012200000000220002012100000110002
 00000020010021012112201111000100100001102001

002101010111002000002220200200002010201002001201001012202
 122000001112222000020202000210000022222212002220200200220
 100222220222020202222020101110020202220200222222000020222
 200022100200211100012002120011121202210020220102011111121
 00002022222000222222000002211012222210100020002100110121
 002100100021000022212121221210222002010220102022222002201
 10002011022012010120000022002000222102222220000012110000
 220022211111021200200222011201101010120002002200000200001
 01010010010010011012100202000100100000201000

002101010111002000002220200120022020200001112200010012202
 122010011122222100020202000200102001211211211220200200220
 0002110212211212122120201011120111111122022222000001002
 11111200211201100021012110000020212200010220111011111120
 01002022222111111111111002211012221200000020002100110121
 10210011002111111222221120210212001100201002022122002212
 000020210222022021120000022002000222102222220000012210000
 110111211111021200200222021201212010120002002200000200002
 00000021020021012022202021000110100000112000

002101101211002000002220200211011120201002101200010002202
 222001012022222000020202000221011111211102111220200200210
 100211121122020202211120000201021202111101111112000021122
 200022200200212200022002120100121202200020220102112022221

00002022222000222222000002202022221200010020002200210122
002000220020021202222200220200222002000211012012022002202
000020210222022020120000022002000222102222220000112211000
220022202000022101111111021202202010111112001110100110002
00010010010011012022202021000110100000110000

111100202200012000102220200210011120201002002200000012202
222001012012222011111111011210001011211201102220200200220
100211220222020202222020100200020202111210222222011111121
211112200200212200022002120000122202200020111012002022212
0000202222200022222200000222000222211100020002200220122
01210001001100011221222122220212001110220012012122002201
1001201202220010102201111110111111120222220000012020000
220022211111020200200222002200111010121112001110000110002
00000020020021112021202020000100100000102000

111101010111002000101120200220012020100001112200000022212
022010011122122100020202000200101011211212111220200200220
000222021212120202201120000202121111111210222222011102012
200011200210200000021012020011020202200020220111121111120
02002022222222000000222002202022220200000121102200220012
002000220021011201222210220210121002000211002022122002202
000020210221022021120000022002000222102222120000021110000
110111211111021200200222002200111120021102101111000200012
00000020010020011102100202000000200000101000

002110101211002000002220200200101120202002002200000011202
122001111112222200020202000200202001211211211220200200220
000211120222120202202020000202021112002111111112000002002
200001200110200000020122020100020212200020111020121111120
02002022222222000000222002202022220200000121102200220012
001100110021011112222211221210221002000202002022122001202
000010220222022020020000022002000222102222220000022200000
220021201111021211111111011201212010111112002100000110002
00000010111120011012100211000010100000111010

002110101211002000002220200200101120202002002200100012202
12200101202222000020202000221011111211102111220200200210
100211121221021212221120000201120202220200222222000020222
200022200200211110022002120000122202200020220011002022221
010020222222111111111111002220002222211100020002210110122
002201010021010112212211221220222001110220002121212002201
11102111021101211121000002200200022202212220010002110000
220022211111021200200222011201101010120002001110001110002
01010010011020111012101111000100100000101000

111101111100011111112220200200002020202002101200010012202
122010021112222100020202000200101011211212111220200200220
100222120222120202212020000201020202111210222222000011112
200011200210201100021012020000121202200020220011011111121
010020222222111111111111002211012221200000020002200220022
002100120020011202222121221220212101110220012012122002101
10002011022201201022000002200200022202222220000012020000
220022211111011100200222012200111010121112001111000200012
00000020010020011102100202000000200000101000

002110010111002000002220200210111120200001112100000021202
02201112111222220002020200020020200222222220220200200210
000222011221120201212020000111121111220101111112000020222
200011200210211100021012020000121202200020220011011111121
010020222222111111111111002211012221210100020002100110121
002100100022000112222121221220212101110220012012122002201
10002012022201201122000002200200022202222220000002121000

220022211111021200200222011201101010120002001110000110002
00000020010021012022201111000100000000101110

111100202200002000112221100200002020101002001200010002202
22200101202222000020202000220000022222212111220200200220
10022212022212020221202000020102111211110111112000011122
200022200200211110022001120000121202200020111011012022221
0000202222200022222000002220002222210210021111201220021
00210010002101011222221211220221002010220002022222002201
10002011022101201022000002200200022220222220110012110000
22102222022020200200222002200000010120002012100000110002
00000020010020011102100202000000200001102001

002200020000002000002220200210012010200001111201001012202
1220000011122220000202020002100000222221200220200200220
1002222022202020222202010111002020222020022222000020222
200022100200211100012002120011020202210020220111020200020
01002022222111111111111002202022221200000020002100110121
002000210021000022212121221210222002010220102022222002201
100020110221012011220000022002000222202222120010002110011
22002111111021200200222002201001020011112002100000110002
00100010111120011012100202000000100001101000

111101010111002000002220200220021120201001212200000022202
0220111211222220000202020002002020022222220220200200220
1012110212211211122120201011121211111122022222000011111
21110121021020000001101202000001120220001022011101111120
01002022222111111011111002202022221200010020002200210122
00200022002002120222200220210221002000211002022122002202
000020210221022021120000022002000222102222120010012200000
22102221111102120111111021201212010120002002200000200002
00000021020021012022202021000110100000112000

002101101211002000002220200211011120201002101200010002202
22200101202222000020202000221011111211102111220200200210
10021112112202020221112000020102120211110111112000021112
20001120021020110002101212010002020220002022011112111120
01002022222111111111111002202022221200010020002200210122
00200022002002120222200220210221002000211002022122002202
000020210221022021120000022002000222102222120010012200000
22102221111102120111111021201212010120002002200000200002
00000021020021012022202021000110100000112000

002101101211002000002220200211011120201002101200010002202
22200101202222000020202000221011111211102111220200200210
10021112112202020221112000020102120211110111112000021112
20001120021020110002101212010002020220002022011112111120
01002022222111111111111002202022221200010020002200210122
002000220020021202222002202022002000211012012022002202
000020210222022020212000002200200022210222220000112211000
220022202000021101111111021202202001221112001100000200002
00101010010011001012101111000200100000111000

111100111100012000102220200220021120200001112200000021202
0210111211122221111111111011200102001211211211220200200220
00021112022212020221202010020102020200222022222011102011
211101200210201100021012120000021202200020111021011111111
01002022222111111111111002211012221201000020002200220122
012000120011000112212221222220212001110220012012122002201
10102012022200101022011111101111111120222220000012020000
220022211111020200200222002200111010121112001110000110002
00000010111120011012100202000010100000200000

111101101211002000002210200200002010201002002200000012202
22200101202222000020202000210001012222212102111200200221
101222111221020202212010000202021111111210222222000011112
200011200210201100020012120000020202201120220111011111120
020020222111222000000222111102122121200000010002201220022
002100210121011202222110220200222002000202002022122002202
000020210221022021120000022002000222102222120010012200100
221022211111021101111111021202202010111112001110100200012
00000020011021012022201111000100100000102000

002110010111002000002220200210111120201001112200000021202
02201112111222211111111011200102001211211211220200200220
00021112022212020220201100020202111111210222222000011112
211101200210201100020022020000020202100020111020021111120
01002022222111111111111002211012221200110021111201220021
002000210020021202222210210210221002000211002022122002202
000020210221022021120000022002000222102222120000022200000
220022202000022200200222021201101000220002012100000110002
00000010111121012022201120000110000000111010

111100111100012000102220200220011120210101112200000022202
12201002201222210002020200020010101222222211220200210220
000211021221121212220201000020202111111210222222011102011
21110120021020001002101212000002120220002022001100202221
010020222222111111111111002220002222110100020002200220022
002211000020021211222210220210221002000202002022122001202
000010211221022011120000022002000222112222120000022200000
220022202000021200200222011201102010120002002100101110002
01010010011021112022202020000200000000102000

111101111100011111112220200200002020202002101200010002202
222000011112222000020202000210000021211202002220200200220
20022220222020202222020000200020202220200222222000020222
20002220020021220002200202000022220220002022000200202222
00002022222000222222000002220002222210100020002200220022
002200010021000112222121221220212101110220012012122002201
100020110222012111210000022002000222202222120000012121000
220022211111021201111111002201001020011112002100000110002
00000020011120000002000202100000100000111000

002110010111002000102220201220022010200001112200000022202
02202002202222220002020200020020200222222220220200200210
00022201122112020121202000011212111111111111112000011112
200000200220200000020022020000020202200020220020020200020
020020222222222000000222002202022220200000020002100110121
002000210021011202222110220210212101100211012012022002202
0000202202202202020020000022002000222002222220000022200000
220022202000022200200222020202202000220002001110000110002
00000020010021012022201111000100000000101110

111100111101002000002220200210012020100001111200010012202
12201102202222210002020200021010101222222220220200200220
000222020222220202202020000202021112002111111112000002012
200011200210200010021011120000020202200020111020021111120
010020222222111111111111002211012221200110021111201220021
002000210020021202222210210210221002000211002022122002202
000020210221022020120000022002000222102222220110022200000
221022211111021200200222011201102010111112001110000200002
10101020011021001002111111000100200000100000

002200020000002000002220200210012010200001111201001022202
022010011112222100020202000200000022222212002220200200220
0002221202221202022220201011100202220200222222000020222
200022100200211100012002120011121202210020220102011111121
0000202222200022222000002211012222210100020002100110121
0021001000210000222121212212102220020102201022222200220
10002011022101201122000002200200022202222120010012110000
220022211111021200200222011201101010120002002200000200001
01010010010010011012100202000100100000201000

111101010111002000002220200220021120201001212200000022202
022011121122222100020202000200102001211211211220200200220
000211021221121212212020101112021111111220222222000011102
111111200211201100021012110000020212200010220111011111120

010020222222111111111111002211012221200000020002100110121
 10210000002210002222222121220212001110210002022222002211
 10002011022201201122000002200200022220222220000002120000
 110111220222020200200222012200111020020002002200000200002
 00000021020020011012101112000010200000111000

002101101211002000002220200211011120201002101200010012202
 122001011122121100020202000210101011111202102220200200220
 100211121221021212212020001111121111220200222222000020222
 211122100200212201022002020000122202200020220102112022221
 010020222222111111111111002211012222210100020002200220021
 102100000021010122212121221121221002010220002022222002201
 10002111021101201121000002200200022202222220000002020000
 221022220222020101111111012201101020011112001110100110002
 00010010010010011012101112000010200000101010

002110101200012000102220200221021120200001211200010012202
 122011022022222100020202000211112101211112220220200200210
 000211021222120202201120000202121111111210222222011102012
 200011200210200010021011120000020212201120220111020200020
 020020222222222000000222002202022220200000020002100110122
 002001211020011112212111220210212001100211012012022002202
 00012022022201102012011111101111111102222220000022110000
 220022202000021200200222011201212000221112001110000110002
 00000010111121012022201111000110000000201000

002200020000001110002210200210012010200001112200000022202
 222001012022222000020202000210001012222212102111200200221
 101222111221020202212010000201021111220200222222000020222
 200022200200212200021002120000121202201120220102002022221
 0100202221111111111111111111111112122210100010002201220022
 002200100122000112222121221210222002010211002022222002201
 100020210221022021120000022002000222102222120010002110100
 221022220222020101111111012201101020011112001110100200012
 00000020011020011012100202000000200000101000

111100202200002000112221100200002020101002001200100012212
 12200000112222220002020200020020200222221211220200200220
 0002220112211212122202011000202021111111210222222000011112
 211101200210201100020022020000020202100020111020021111120
 010020222222111111111111002211012221200110021111201220021
 002000210020021202222210211220221002010220002022222002201
 10002012022201201012000002200200022202222220000012110000
 220021210222020211111111002200111020011112002100000110002
 00000010111120011012100211000010100000110010

111100111100012000102220200220011120210101112200000022202
 12201002201222210002020200020010101222222211220200210220
 0002110212211212122202010000202021111111210222222011102011
 211101200210200010021012120000021202200020220020011111120
 020020222222222000000222002211012221201000020002210110122
 002101120020021202212221221220221002010211002022222001201
 111021210211022021120000022002000222112222120000022200000
 220022202000021200200222011201102010120002002100101110002
 01010010011021112022202020000200000000102000

111101020000011111112220200210012020201001211200010012202
 122010021112222100020202000200101011211212111220200200220
 100222120222120202212020000201020202111210222222000011112
 200011200210201100021012020011020202200020220020020200020
 111111111112222000000222112211012221200010020002200210122
 012000120020011112212110220200221002010211012012122002101
 100020110222012111210000022002000222202222120000012121000

220022211111021201111111002201001020011112002100000110002
00000020011120000002000202100000100000110000

002110101211002000102220201210012010201002002200000011202
12200111112222100020202000210101012211202002111200200221
1011221112210211112120200002021211111121022222011102012
211111101110200000021012020000020202200020220111020200020
01002022222111111111111002211012221201000020002200220122
012000120010011202212210220210221002000211002022122001202
00002021022202202112000002200200022210222220000012211000
220022202000022200200222020202202000220002001110000110002
00000020010021012022201111000100000000101110

111100202200002000112221100200002020101002001200100012212
12200000112222100020202000210101012222211102220200200220
10022211122102121221201100020102111122020022222000020222
211112200200212200021012020000121202100020111011012022221
00002022222000222222000002220002222210210021111201220021
002100100021010112222221211220221002010220002022222002201
100020110221012011220000022002000222202222120000022200000
220022202000022200200222021201101000220002012100000110002
00000020010021012112201111000100100001102001

002101010111002000002220200210012010200000222200000022202
02202002202222220002020200020020200222222220220200200220
0002220202222202022020200002020202002220222222000002002
200000200220200000020022020000020202200020220020020200020
02002022222222000000222002202022220200000020002200220022
00200022002002220222200220200222002000202002022022002202
00002022022022020120000022002000222102222220000012110000
220022211111021200200222011201101010120002002200000200001
01010010010010012022201111000200000000202000

002101101211002000002220200110012020201002002200010002202
222000001122222000020202000210001011211201102220200200220
100211121221021212222020101111011111220210222222000010112
111122200201212200022002110000121212200010220102002022221
00002022222000222222000002220002222210100020002100110121
1022000000221000222222121220212001110210002022222002211
1000201102220120112200000220020002222022220222000002120000
110111220222020201111111012200001020020002012100000110012
00000010020020011012100211000000200000100001

111100202200012110112221110200002020101002001200000012202
122001011122121100020202000210101011111202102220200200220
100211121221021212212020001111121111220200222222000020222
21112210020021220102101202000021202200020220111121111120
02002022222222000000222002202022221200000020002200220021
102000110020021212212110220111221002000211002022122002202
000021210211022021110000022002000222102221220000012110000
221022211111021101111111021202202001221112001100000200002
00110010010011012022202021000110100000102010

111100111100012000102220200220021120200001112200000021202
02101112111222211111111011200102001211211211220200200220
000211120222020202222020100200020202111210222222011111121
211112200200212200022002120000122202200020111012002022212
000020222222000222222000002220002222211100020002200220122
01210001001100011221222122220212001110220012012122002201
10012012022200101022011111101111111202222220000012020000
220022211111020200200222002200111010121112001110000110002
00000010111121012022201111000110000000201000

002200111100001110002210200200002010201002002200000022202
 1220110220222210002020200020010200222222211111200200221
 00122201122112020221201000020102111122020022222000020222
 200022200200212200021002120000121202201120220102002022221
 0100202221111111111111111111111112122210100010002201220022
 002200100122000112222121221210222002010211002022222002201
 100020110221012011220000022002000222202222120000011020000
 110111220222021200200222011201212110121102101111000200012
 00000020010021012112201111000100100000102000

111100202200002000112221100200002020101002001200100012212
 122000001122222100020202000210101012222211102220200200220
 1002221112210212122120110002010211111121022222000011112
 211101200210201100020022020000020202100020111020121111120
 02002022222222000000222002202022220200000121102200220012
 001100110021011112222211221210221002000202002022122001202
 000010220221022021120000022002000222102222120000022200000
 220022202000022200200222021201101000220002012100000110002
 0001001001001101202220202000010000000101010

111100111100012000102220200220011120210101112200000022202
 12201002201222210002020200020010101222222211220200210220
 00021102122112121220201000020202111111210222222011102011
 211101200210211110022002120000122202200020220011002022221
 010020222222111111111111002220002222211100020002210110122
 002201010021010112212211221220222001110220002121212002202
 011021210211022121110000022002000222102212220010012200000
 220022202000022200200222020202202000220002001110000110002
 0000002001002101202220202000020000000102110

0021011012110010001122211002000020201010020012100000022202
 122101012012222100020202100210101012222211102220200200220
 1002221112210212122020201011121211111111111112000011112
 200011200210201100021012020000120202200020220011011111121
 1011111111211111111111111112220002222210110020002200210122
 012100010021000022212121221220212101110220012012122002201
 100020110222012111210000022002000222202222120000012121000
 220022211111021201111111002201001020011112002100000110002
 00000020011120000002000202100000100000110000

002110010111002000102220201220022010200001112200000022202
 02202002202222220002020200020020200222222220220200200210
 00022201122112020121202000011212111111111111112000011112
 200000200220200000020022020000020202200020220020020200020
 02002022222222000000222002202022220200000020002100110121
 002000210021011202222110220210212101100211012012022002202
 0000202202202202002000002200200022200222220000022200000
 22002220200002220020022202020220200022000200220000020002
 00000020020022002022202021000110100000201010

111100202200002000112221100200002020101002001200100012212
 122000001122222100020202000210101012222211102220200200220
 1002221112210212122120110002010211112202002222000020222
 211112200200212200021012020000121202100020111011012022221
 0000202222200022222000002220002222000222210210021111201220021
 00210010002101011222222121122022100201022000202222002201
 100020110221012011220000022002000222202222120000012110000
 220022211111021200200222012200000010120002012100000110002
 00000020010020011102100202000000200001101001

002200111100002000002220200200002010200001111201001022202
 022010011112222100020202000200101012222222111220200200220
 000222120222120202212020101111020202111210222222000011112
 200011100210200000011012120011020202210020220111020200020

0100202222211111111111002202022221200000020002100110121
002000210020011112212110220200222002000211102022122002202
0000202202201201012000002200200022210222220000012110000
22002221111021200200222011201101010120002002200000200001
01010010010010011012100202000100100000201000

111101101211002000002220200210011120202002102200000012202
1220011111222210002020200021020200222222220220200200220
1012110212211211122120201011121211111122022222000011111
211101210210200000011012020000020212200010220111011111120
0100202222211111111111002211012221200000020002100110121
1021001100211111222221120210212001100201002022122002212
00002021022202202112000002200200022210222220000012200000
22102221111102020111111012200001020020002012100000110012
00000010020021012022201120000100100000101001

002101010111002000002220200221021120200001211200010012202
12201102202222100020202000211112101211112220220200200210
00021102112212020220112000020202111111121022222000011112
21111100210201101021012020000021202200020220111121111120
02002022222222000000222002202022221200000020002200220021
102000110020021212212110220111221002000211002022122002202
000020210222022020120000022002000222102222220000112211000
22002220200002210111111021202202010111112001110100110002
00010010010011012022202021000110100000102010

111100202200012000102220200210011120201002101200010002202
22200101202222000020202000221011101211112220220200200210
00021102122212020220112000020212111111121022222011102012
200011200210200010021011120000020212201120220111020200020
02002022222222000000222002202022220200000020002100110122
002001211020011112212111220210221002010211002022222002201
1010211102210120102200000220020002220222220000002020000
220022211111020200200222012201101020011112002100000200002
00000020020020111011101111000010100000200000

002200111100001110002210200200002010201002002200000012212
122000001122122000020202000210000021211202002220200200220
100222121212020221112000020112111122020022222011111122
200022200200211100022002020011121202200020220102112022221
01002022222211111111111002211012221210100121102200220012
002100110022000111222221221220121002010220002022222002201
10002011022101201122000002200200022202222120010002110100
22102221111102110111111021202202010111112001110100200012
00000020011021012022201111000100100000102000

002110010111002000002220200210111120201001112200000021202
0220111211122210002020200020020200222221211220200200220
000222011221121212202011000202021111111210222222000011112
211101200210201100020022020000020202100020111011012022221
000020222220002222200000222000222210110021111201220021
00200021002002120222210210210221002000211002022122002202
000020210221022021120000022002000222102222120000022200000
220022202000021200200222012200000010120002012100000110002
00010010010010011012101111000000100000100010

002110101211002000002220200200101120202002002200100012202
12200101202222100020202000211112101211112211220200200210
10021112122102121222112000020112020222020022222000020222
200022200200211100022002020000112202200021111011012022221
0000202222200022222000002220002222110100020002200220022
002211000021010121222221221220221002010211002022222001201
100010111221012011220000022002000222212222120000012110000

220022211111020200200222002200001020020002002100101110002
01010010011020111012101111000100100000101000

002101101211001000112221100200002020101002001210000002202
122101012012222100020202100210101012222211102220200200220
1002221112210212122120201011112111122010111112000020222
200022200200212200022002020000222202200020220002002022222
00002022222000222222000002220002222210100020002200220022
002200010021000112222121221220212101110220012012122002201
100020110222012111210000022002000222202222120000012020000
220022211111011100200222012200111010121112001111000200012
00000020010020011102100202000000200000101000

002110101211002000102220201210012010201002002200000012202
122010111112222100020202000210101012211202002111200200221
101122111221021111212020000201121111220200222222011111122
211122101100211100022002020000121202200020220102011111121
00002022222000222222000002220002222211100020002200220122
012100010011000112212221221220221002010220002022222001201
100020110222012011220000022002000222202222220000002121000
220022211111021200200222011201101010120002001110000110002
00000020010020011012100202000000100000100110

111100111101002000002220200210012020100001111200010012202
12201102202222220002020200020020200222221211220200200220
000222011221121212202011000202021111111210222222000011112
211101200210201100020022020000020202100020111020021111120
010020222222111111111111002211012221200110021111201220021
002000210020021202222210210210221002000211002022122002202
000020210221022021120000022002000222102222120000022200000
220022202000022200200222021201101000220002012100000110002
00000020010021012112201111000100100001102001

002200020000002000002220200210012010200000222200000022202
022020022022222100020202000200101012222222111220200200220
000222120222120202212020101111020202111210222222000011112
200011100210200000011012120011020202210020220111020200020
010020222222111111111111002202022221200000020002200220022
00200022002002220222210220200222002000211102022122002202
000020210221022021120000022002000222102222120010012200000
220022202000022200200222020202020200022002002200000200001
01010010010011012022201111000200000000202000

111101010111002000002220200220021120201001212200000022202
02201112112222220002020200020020200222222220220200200220
101211021221121112212020101112121111111220222222000011111
211101210210200000011012020000011202200010220111011111120
0100202222221111111011111002202022221200010020002200210122
002000220020021202222200220210221002000211002022122002202
000020210221022021120000022002000222102222120010012200000
221022211111021201111111021201102010120002012100000110012
00000010020021012022201120000100100000101001

111101202200012110112221110200002020101002001200000012202
122001011122121100020202000210101011111202102220200200220
100211121221021212212020001111121111220200222222000020222
211122100200212201022002020000122202200020220102112022221
010020222222111111111111002211012221200000020002200220021
102000110020021212212121221121221002010220002022222002201
100021110211012011210000022002000222202222220000102121000
220022211111021101111111012201101020011112001110100110002
00010010010010011012101112000010200000101010

111100202200012000102220200210011120201002002200000011202
12100111111222201111111011210001011211201102220200200220
10021122022202020222202010020002020211121022222011111121
211112200200212200022002120000122202200020111012002022212
00002022222000222222000002220002222201000020002200220122
012000120010011202212210221210212001100211012012022002202
0001202202220110201201111110111111110222220000022110000
220022202000021200200222011201212000221112001110000110002
00000010111121012022201111000110000000201000

5. GENERAL CONCLUSIONS

The limited resources to phenotype new individuals are scarce and the use of different populations to compose the training set has been applied as a solution. Hence, population structure effect should be accounted to obtain lower bias. However, the reduction of accuracy can be observed, if uni and multi-population model with modified genomic relationship matrices are applied for genomic prediction of admixed populations. Therefore, the strategy of accounting population structure effect should be chosen according to trait heritability, trait architecture, and the admixture level of population for success of genomic prediction.

Furthermore, k-fold cross-validation methods can produce fitted models that account for phenotypic and genotypic diversity and, thus, should be regarded as more than simple validation statistical tool. The novel methods also highlighted that the phenotypic diversity of validation set should be increased to obtaining higher predictive ability. Therefore, increasing the diversity of breeding populations, considering the genotypic and phenotypic diversity, when k-fold cross-validation is applied, and accounting the population structure effect, when combined populations are evaluated, are essentially relevant strategies for a consistent success of long-term genomic breeding programs.