Multigenerational prediction of genetic values using genome-enabled prediction

Isabela de Castro Sant' Anna[ID][1]*, **Ricardo Augusto Diniz Cabral Ferreira**[2], **Moysés Nascimento**[1], **Gabi Nunes Silva**[3], **Vinicius Quintão Carneiro**[4], **Cosme Damião Cruz**[4], **Marciane Silva Oliveira**[4], **Francyse Edith Chagas**[4]

1 Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **2** Syngenta—Seeds Production Research, Uberlândia, Minas Gerais, Brazil, **3** Department of Statistics, Federal University of Rondônia, Ji-Paraná, Rondônia, Brazil, **4** Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

* isabelacsantanna@gmail.com

Abstract

The identification of elite individuals is a critical component of most breeding programs. However, the achievement of this goal is limited by the high cost of phenotyping and experimental research. A significant benefit of genomic selection (GS) to plant breeding is the identification of elite individuals without the need for phenotyping. This study aimed to propose different calibration strategies using combinations between generations from different genetic backgrounds to improve the reliability of GS and to investigate the effects of LD in different types of mating systems: outcrossing (A_n) self-pollination (S_n) and hybridization (H_n). For this purpose, we simulated a genome with 10 linkage groups. In each group, two QTL were simulated. Subsequently, an F_2 population was created, followed by four generations of inbreeding (S_1 to S_4, H_1 to H_4, A_1, to A_4). Quantitative traits were simulated in three scenarios considering three degrees of dominance (d/a = 0, 0.5 and 1) and two broad sense heritabilities (h^2 = 0.30 and 0.70), totaling six genetic architectures. To evaluate prediction reliability, a model (RR-BLUP) was trained in one generation and used to predict the following generations of mating systems. For example, the marker effects estimated in the F_2 population were used to estimate the expected genomic breeding value (GEBV) in populations S_1 through A_4. The squared correlation between the GEBV and the true genetic value were used to measure the reliability of the predictions. Independently of the population used to estimate the marker effect, reliability showed the lowest values in the scenario where d = 1. For any scenario, the use of the multigenerational prediction methodology improved the reliability of GS.

Introduction

Genomic selection (GS) was proposed by [1] as a means to increase selective efficiency, reduce frequency and costs of phenotyping, and also increase annual gains by reducing selection cycle time [2]. This approach, mainly based on the existence of linkage disequilibrium between markers and genes that control traits, employs simultaneous estimation of genetic marker

effects that are distributed over the genome in order to explain much of the genetic variation and predict the genetic value of individuals [3].

According to [4] the success of the GS approach depends on the heritability and the genetic architecture of the traits (number and effect of Quantitative trait loci—QTL), as well as on the availability of linkage disequilibrium (LD) between markers and QTL in the training and validation populations [4, 5]. The response to GS relies on linkage disequilibrium (LD), thereby the stronger the LD, the higher the reliability of genomic predictions [6,7,8], and higher the long-term gains[9,10,11,12].

According to [13], the LD between QTL and Single Nucleotide polymorphism (SNP) will decrease over generations and the reliability of genomic prediction is expected to decrease without reestimating the SNP effects in more recent generations. Therefore, estimates obtained to different mating systems, such as outcrossing (allogamous plants), self-pollination (autogamous plants) and hybridization may be affected differently.

In this context, we aimed to (1) propose different calibration strategies using combinations between one or more generations from different genetic backgrounds, here called multigenerational sets, to improve the reliability of GS predictions; (2) to investigate the effects of LD in different types of mating systems (outcrossing, self-pollination and hybridization) on the reliability of GS predictions.

Material and methods

Origin of populations

In order to assess the reliability of GS predictions, data were simulated by considering a diploid species with $2n = 2x = 20$ chromosomes as the reference, and the total length of the genome was stipulated in 1,000 cM. Genomes were generated with a saturation level of 101 molecular markers spaced by 1cM per linkage group, totaling 1010 markers. Divergent parental line genomes were simulated, as well as genomes from the base population (F_2), which was used as the reference for obtaining four generations advanced by random mating A_n (n = 1,2, 3, 4), self-pollination S_n (n = 1,2, 3, 4) and hybridization, represented here by backcrossing involving the F_2 parent as Bc_n (n = 1,2, 3, 4). The effective size of the base population is the size of F_2 itself, since the base population (F_2) was derived from two contrasting homozygous parents. This population is in Hardy-Weinberg equilibrium and therefore all disequilibrium is caused by factorial linkage.

Simulation of quantitative traits

The genotypic value for the monogenic model is defined by u + a, u + d, u–a for the genotypes AA, Aa e aa, respectively. In a polygenic model, the total genotypic value expressed by a given individual belonging to the population was the sum of each additive effects of individual locus estimated by the following expression

$$G_i = \mu + a_i + d_i \tag{1}$$

where the additive effect (a) of each locus is one half the difference in mean phenotype between the two homozygous genotypes (for each individual i). The dominance effect (d) is the difference between the mean phenotype of the heterozygous genotype and the average phenotype of the two homozygous genotypes. In our simulation we defined 20 loci to control the trait. Therefore, the additive effect is given by:

$$a_i = \sum_{j=1}^{20} p_j \alpha_j \tag{2}$$

With α_j being the effect of the favorable allele in locus j, considered equal to 1, 0 or -1 for the genotypic classes AA, Aa and aa, respectively, and p_j being the contribution of locus j to the manifestation of the trait under consideration. In this study, it was established as being equivalent to the probability of the set generated by the binomial distribution $(a+b)^s$, where $a = b = 0.5$ and $s = 19$. The value of d_i was defined according to the average degree of dominance expressed in each trait. The quantitative traits were simulated in three scenarios considering three degrees of dominance ($d/a = 0$, 0.5 and 1) and two broad sense heritability ($h^2 = 0.30$ and 0.70), totaling six genetic architectures.

The phenotypic values of the i$^{\text{th}}$ individuals were obtained according to the model: $P_i = G_i + E_i$, where G_i is the genetic effect given by the sum of the genetic effects in each locus, and E_i is the environmental effect, generated according to a normal distribution with means equal to zero and variance given by the equation bellow:

$$\sigma_e^2 = \frac{\sigma_g^2(1 - h^2)}{h^2} \tag{3}$$

where σ_e^2 is the variance given by the environmental values, σ_g^2 is the variance of the genetic values, and h^2 is the heritability defined for the trait. The genetic variance is defined for each population from the information of the genetic control and the importance of each locus in the polygenic model.

$$\sigma_g^2 = \frac{1}{2}\bar{a}^2 + \frac{1}{4}\bar{d}^2, \tag{4}$$

where $\bar{a}^2$, $\bar{d}^2$ were defined by the mean values of the effects associated with the homozygote and heterozygous genotypes for each one of the 20 loci, respectively.

Efficiency of different calibration strategies for GS studies considering different kinds of mating

In order to assess GS reliability over generations, three calibration strategies were performed. In the first one, we used genotypic and phenotypic information from generation F_2 for training the model, and validation was performed in each advanced population defined according to the different mating systems. In the second one, the contemporary or outdated genotypic and phenotypic values from each population itself (S_n, A_n and Bc_n; n = 1,2,3,4) were used, and validation was performed in the same population (S_n, A_n and Bc_n) and in each advanced population (S_{n+1}, A_{n+1} and Bc_{n+1}). The last strategy used three different kinds of calibration data sets (D_{1t}, D_{2t}, D_{3t}), considering multigenerational sets for prediction. In the first training data set (D1t), the total of 1000 individuals, 500 from the first generation of allogamous or autogamous and 500 individuals from the F_2 population ($D_{1t} = A_1/S_1 + F_2$), were used to estimate the marker effect. Validation was performed in 500 individuals (D_{1v}) from each outdated population ($D_{1v} = A_1/S_1$, A_2/S_2, A_3/S_3 and A_4/S_4). In the second and third kinds of calibration data sets (D_{2t} and D3t), the sets were added using generations two ($D_{2t} = A_1/S_1 + A_2/S_2 + F_2$) and three ($D_{3t} = A_1/S_1 + A_2/S_2 + A_3/S_3 + F_2$). These models were, respectively, validated using the true breeding values (TBV) of the population itself (S_n, A_n) and in each advanced population (S_{n+1} and A_{n+1}) (Fig 1).

The additive dominance model for the REML/G-BLUP method is given by [1]:

$$y = X_b + Zu_a + Zu_d + e, \tag{5}$$

where y is the vector of phenotypic observations, b is the vector of fixed effects, u_a is the vector of random of additive marker effects, u_d is the vector of random of dominance marker effects
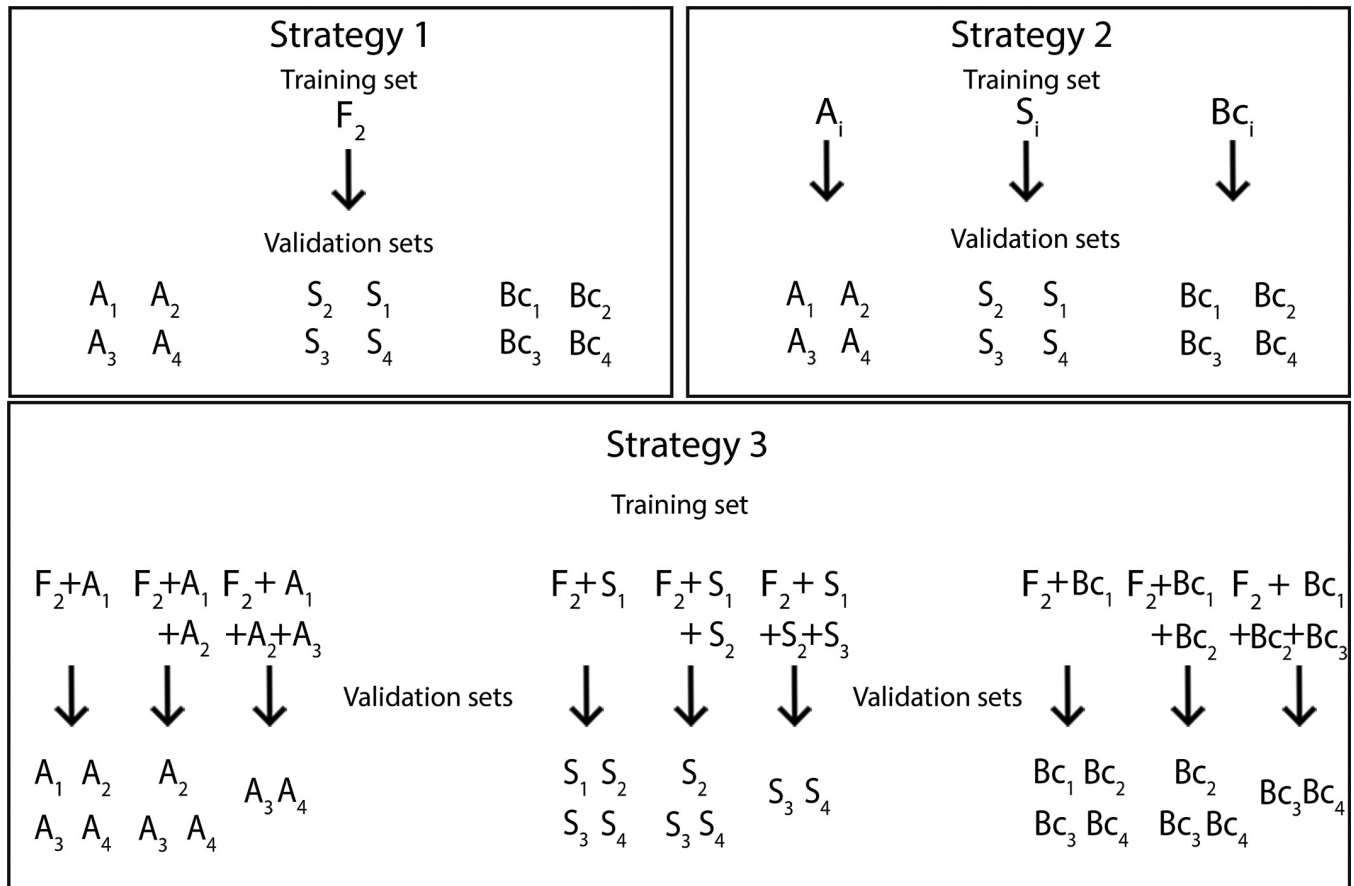
Fig 1. Illustrative scheme of three strategies used in calibration sets of genomic selection training analyses. Letters F_2, A, S and B_C represent the base population of allogamous, self-pollinated and backcrossing species. The index from 1 to 4 represents breeding cycles. In the right square, strategy 2 is the training set composed of species that used the contemporary or outdated genotypic and phenotypic values from the population itself (S_i, A_i and Bc_i; i = 1,2,3,4), and validation was performed in the same population (S_n, A_n and Bc_n) and in each advanced population (S_{n+1}, A_{n+1} and Bc_{n+1}). The last strategy used three different kinds of calibration data sets (D_{1t}, D_{2t}, D_{3t}), which considered multigenerational sets for prediction. In the first training data set (D_{1t}), the total of 1000 individuals, 500 from the first generation of allogamous or autogamous species and 500 individuals from the F_2 population ($D_{1t} = A_1/S_1 + F_2$) were used to estimate the marker effect. Validation was performed on 500 individuals (D_{1v}) from each outdated population ($D_{1v} = A_1/S_1$, A_2/S_2, A_3/S_3 and A_4/S_4). In the second and third kinds of calibration data sets (D_{2t} and D_{3t}), data sets were added using generations two ($D_{2t} = A_1/S_1 + A_2/S_2 + F_2$) and three ($D_{3t} = A_1/S_1 + A_2/S_2 + A_3/S_3 + F_2$).

and e refers to the vector of random errors; The variance structure is given by $u_a \sim N(0, G_a\sigma^2_{u_a})$; $u_d \sim N(0, G_d\sigma^2_{u_d})$; by $e \sim N(0, I\sigma^2_e)$.

An equivalent model at the marker level is given by

$$y = Xb + ZWm_a + ZSm_d + e, \tag{6}$$

where: $u_a = Wm_a; Var(Wm_a) = WI\sigma^2_{m_a} W' = WW'\sigma^2_{m_a}; u_d = Sm_d, Var(Sm_d) = SI\sigma^2_{m_d} S' = SS'\sigma^2_{m_a}$;

X and Z are matrices of incidence for the vectors additive (ma) and dominance (m_d) marker genetic effects. The variance components associated to these effects are $\sigma^2_{m_a}$ and $\sigma^2_{m_d}$, respectively. G_a and G_d are the genomic relationship matrices for the additive and dominance effects. The quantity m_a in one locus is the allele substitution effect and is given by $m_a = \alpha_i = a_i + (q_i - p_i)di$, where p_i and q_i are allelic frequencies and a_i and d_i are the genotypic values for one homozygote and heterozygote, respectively, at locus i. In turn, the quantity m_d can be directly defined as $m_{di} = d_i$. The matrices W and S are defined based on the values 0, 1 and 2 for the

number of one of the alleles at the *i* marker locus in a diploid individual. The correct parameterization of **W** and **S** is as follows, according to the marker genotypes at a locus m.

$$W = \begin{cases} MM : 2 - 2p \rightarrow 2q \\ Mm : 1 - 2p \rightarrow q - p \\ mm : 2 - 2p \rightarrow 2q \end{cases}$$

$$S = \begin{cases} MM : 0 \rightarrow -2q^2 \\ Mm : 1 \rightarrow 2pq \\ mm : 0 \rightarrow -2p^2 \end{cases}$$

The covariance matrix for the additive effects is given by $G_a \sigma_a^2 = Var(Wm_a) = WW' \sigma_{m_a}^2 =) = ,$ which leads to: $G_a = WW'/(\sigma_{m_a}^2/\sigma_a^2) = WW'/\sum_{i=1}^{n}[2p_i(1-p_i)]$, as $\sigma_a^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]\sigma_{m_a}^2$. The covariance matrix for the dominance effects is given by $G_d \sigma_d^2 = Var(Sm_d) \ SS' \sigma_{m_d}^2$. Thus , $G_d = SS'/(\sigma_{m_d}^2/\sigma_d^2) = SS'/\sum_{i=1}^{n}[2p_i(1-p_i)]^2$ as $\sigma_d^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]^2\sigma_{m_d}^2$

The additive-dominance G-BLUP method was fitted using Genomic Land software [18] via REML through mixed model equations.

After obtaining the GEBVs, reliability, which is defined by the square correlation between the genomic estimated breeding values and the true breeding values (TBV), was estimated [9].

Finally, linkage disequilibrium was estimated using the maximum likelihood method [14,15]. It was chosen to represent the disequilibrium squared correlation coefficient between two loci r^2 [16] for pairs of loci. The LD measure, r^2, corresponds to the ratio of genetic variance controlled by the QTL whose associated marker is able to explain.

Computational applications for data analysis

The simulations were implemented with software GENES [17]) and the statistical analyses were performed with software Genomic Land [18].

Results

Studying LD in all populations

Linkage disequilibrium (r^2) was calculated for all pairwise physical distances among all the SNPs (in each linkage group separately. The average genome-wide LD for pairs of SNPs within a 1 centimorgan distance from each population group was 0.006 for allogamous species, 0.051 for autogamous species, 0.037 for hybridization, and 0.189 for F_2. These results are represented in scatter plots of r^2 values versus the genetic or physical distance between all pair of alleles (Fig 2). The linkage disequilibrium was almost entirely dissipated in allogamous population since the first generation but continues in autogamous and backcrossing sets.

Calibration using the F_2 population

We contrasted the reliability of the linear models (G-BLUP) considering scenarios with and without dominance in different levels of heritability over generations. Overall, the reliability values decreased over generations for all the scenarios considered. Specifically, for the scenarios without dominance (d = 0) and low heritability ($h^2 = 0.30$), the averages of reliability values were 0.55, 0.20 and 0.50 for, respectively, outcrossing (allogamous species), selfing-pollination (autogamous species) and hybridization mating systems (Table 1). In the scenarios with
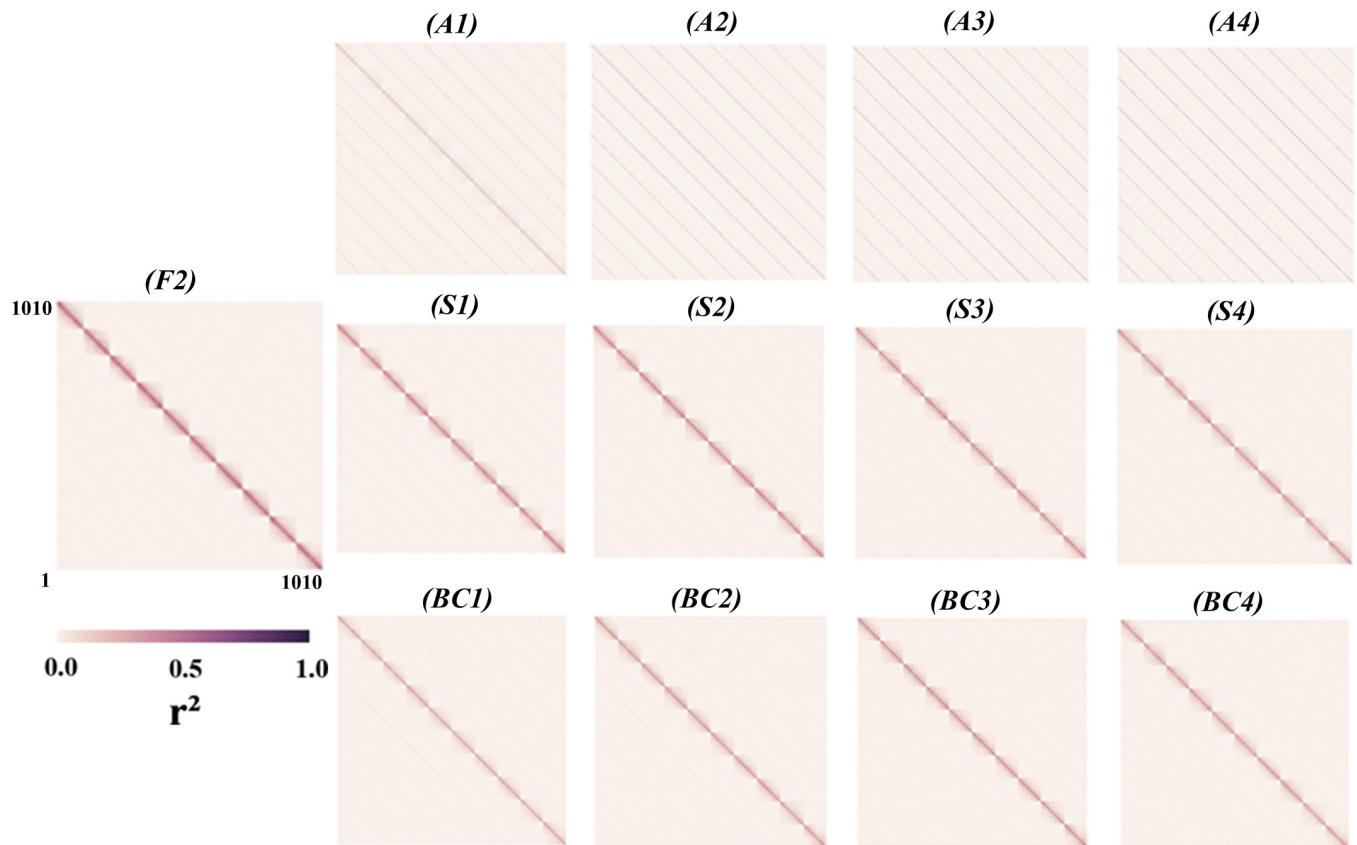
Fig 2. From left to right, estimation of Linkage Disequilibrium measured by r^2 statistic in F_2, allogamous set (top), autogamous set (center), and backcrossing set (bottom). X and Y axes represent the (1010) markers distributed in 10 linkage groups with 101 markers each. The intensite of correlation inside each linkage block decreased the as we observe advanced generations. These results are represented in scatter plots of r^2 values versus the genetic or physical distance between all pair of alleles (Fig 2).

https://doi.org/10.1371/journal.pone.0210531.g002

dominance (d = 0.5 and d = 1) and low heritability, the averages of reliability values changed from 0.49, 0.15, 0.44 to 0.48, 0.17 and 0.43, respectively, for outcrossing (allogamous species), self-pollination (autogamous species) and hybridization mating systems. The GS reliability considering the F_2 population as training and test population was equal to 0.81, 0.72 and 0.70 for the scenarios with low heritability and, respectively, no dominance, partial (d = 0.5) and complete (d = 1) dominance. The results observed for the high heritability scenarios ($h^2 = 0.70$) were similar when compared with those obtained for the low heritability scenarios (Table 1).

Table 1. Reliability values of selection of generations advanced by self-pollination (S_n), outcrossing (A_n) and hybridization (Bc_n) (obtained from phenotyping and genotyping of the F_2 population for six traits (heritability equal to 0.30 and 0.70 repeated in scenarios with an average degree of dominance(d) level equal to 0, 0.5 and 1). The scale of colors changes from 0 (red) to 0.5 (yellow) to 1(green).

Low Heritability																
d	F_2	S_1	S_2	S_3	S_4	Mean	A_1	A_2	A_3	A_4	Mean	Bc_1	Bc_2	Bc_3	Bc_4	Mean
0	0.81± 0.03	0.63 ± 0.07	0.56 ± 0.05	0.53 ± 0.06	0.51 ± 0.04	0.55	0.31 ± 0.06	0.22 ± 0.03	0.16 ± 0.02	0.11 ± 0.01	0.20	0.45 ± 0.01	0.50 ± 0.05	0.53 ± 0.05	0.52 ± 0.04	0.50
0.5	0.72 ± 0.03	0.55 ± 0.01	0.49 ± 0.00	0.47 ± 0.00	0.46 ± 0.01	0.49	0.23 ± 0.03	0.18 ± 0.01	0.12 ± 0.00	0.08 ± 0.01	0.15	0.40 ± 0.01	0.42 ± 0.02	0.47 ± 0.01	0.46 ± 0.01	0.44
1	0.70 ± 0.10	0.51 ± 0.08	0.48 ± 0.06	0.47 ± 0.07	0.47 ± 0.04	0.48	0.20 ± 0.03	0.20 ± 0.04	0.15 ± 0.03	0.12 ± 0.04	0.17	0.36 ± 0.07	0.41 ± 0.05	0.49 ± 0.07	0.47 ± 0.09	0.43
High Heritability																
0	0.94 ± 0.01	0.73 ± 0.01	0.65 ± 0.01	0.62 ± 0.01	0.60 ± 0.00	0.65	0.39 ± 0.01	0.28 ± 0.02	0.19 ± 0.02	0.14 ± 0.02	0.25	0.52 ± 0.00	0.57 ± 0.01	0.62 ± 0.01	0.62 ± 0.00	0.58
0.5	0.89 ± 0.05	0.69 ± 0.04	0.62 ± 0.04	0.60 ± 0.04	0.59 ± 0.04	0.68	0.35 ± 0.05	0.28 ± 0.07	0.21 ± 0.05	0.15 ± 0.08	0.25	0.48 ± 0.03	0.51 ± 0.04	0.59 ± 0.04	0.57 ± 0.05	0.54
1	0.89 ± 0.01	0.65 ± 0.02	0.58 ± 0.00	0.58 ± 0.01	0.57 ± 0.00	0.65	0.23 ± 0.02	0.22 ± 0.01	0.16 ± 0.03	0.11 ± 0.00	0.18	0.46 ± 0.01	0.51 ± 0.01	0.58 ± 0.01	0.57 ± 0.00	0.52

https://doi.org/10.1371/journal.pone.0210531.t001

Calibration using contemporary or outdated populations

Overall, the second calibration strategy outperforms the results obtained with the first calibration strategy, which used only the F_2 population for calibration. For example, population A_4 had the worst reliability prediction value, 0.08 (Table 1), using training strategy 1, and improved to 0.49 (Table 2) using training strategy 2. The reliability values for the different mating systems (outcrossing—allogamous, self-pollination-autogamous and hybridization-backcrossing) obtained using contemporary or outdated populations are displayed in Table 2.

Table 2. Reliability values of selection of generations advanced by self-pollination(S_n), random mating (A_n) and hybridization (Bc_n) obtained from phenotyping and genotyping of the same generation (diagonally contemporary) or only from previous genotyping and phenotyping (outdated off-diagonally, horizontal reading) for the heritability trait equal to 0.30 in the scenario with an average degree of dominance equal to 0, 0.5 and 1.

Traits	Training set	S_1	S_2	S_3	S_4	A_1	A_2	A_3	A_4	Bc_1	Bc_2	Bc_3	Bc_4
							Validation set						
d0	S_1	0.61 ± 0.06	0.55± 0.10	0.53 ± 0.10	0.52 ± 0.10								
	S_2		0.54 ± 0.09	0.53 ± 0.10	0.54 ± 0.10								
	S_3			0.54 ± 0.10	0.60 ± 0.10								
	S_4				0.59 ± 0.02								
	A_1					0.42 ± 0.02	0.37 ± 0.03	0.32 ± 0.01	0.33 ± 0.03				
	A_2						0.44 ± 0.12	0.40 ± 0.10	0.40 ± 0.12				
	A_3							0.50 ± 0.12	0.53 ± 0.09				
	A_4								0.46 ± 0.11				
	Bc_1									0.30 ± 0.09	0.34 ± 0.01	0.36 ± 0.02	0.36 ± 0.02
	Bc_2										0.55 ± 0.05	0.55 ± 0.06	0.54 ± 0.08
	Bc_3											0.55 ± 0.05	0.51 ± 0.04
	Bc_4												0.52 ± 0.07
	Average set	0.54 ± 0.07				0.42 ± 0.08				0.46 ± 0.05			
							Validation set						
d0.5	S_1	0.61 ± 0.11	0.52 ± 0.03	0.48 ± 0.00	0.46 ± 0.00								
	S_2		0.55 ± 0.13	0.54 ± 0.08	0.53 ± 0.08								
	S_3			0.60 ± 0.10	0.54± 0.07								
	S_4				0.55± 0.07								
	A_1					0.48± 0.10	0.36 ± 0.03	0.34 ± 0.07	0.33± 0.11				
	A_2						0.44 ± 0.10	0.35 ± 0.11	0.34 ± 0.11				
	A_3							0.48 ± 0.10	0.40 ± 0.01				
	A_4								0.49 ± 0.12				
	Bc_1									0.53 ± 0.14	0.47 ± 0.15	0.49 ± 0.15	0.47 ± 0.11
	Bc_2										0.50 ± 0.14	0.51 ± 0.09	0.48 ± 0.11
	Bc_3											0.55 ± 0.10	0.48 ± 0.06
	Bc_4												0.55 ± 0.10
	Average set	0.53 ± 0.06				0.40± 0.10				0.50 ± 0.10			
							Validation set						
d1	S_1	0.51 ± 0.08	0.49± 0.09	0.49 ± 0.10	0.49± 0.08								
	S_2		0.53 ± 0.10	0.53 ± 0.09	0.52 ± 0.08								
	S_3			0.56 ± 0.07	0.56± 0.11								
	S_4				0.55± 0.08								
	A_1					0.41 ± 0.11	0.27 ± 0.10	0.27 ± 0.10	0.26 ± 0.11				
	A_2						0.39 ± 0.11	0.29± 0.10	0.24 ± 0.10				
	A_3							0.48 ± 0.11	0.36 ± 0.07				
	A_4								0.47 ± 0.09				
	Bc_1									0.49 ± 0.11	0.39 ± 0.03	0.44 ± 0.05	0.40 ± 0.05
	Bc_2										0.54 ± 0.10	0.46± 0.07	0.45 ± 0.08
	Bc_3											0.45± 0.11	0.38 ± 0.08
	Bc_4												0.53 ± 0.06
	Average set	0.52 ± 0.11				0.34 ± 0.07				0.45 ± 0.09			

Overall, for the scenarios with lower heritability, the inclusion of outdated calibration sets allowed to increase the average of reliability values in allogamous population and maintain in the others mating systems. Specifically, the averages for the allogamous populations (A_1 to A_4), in the scenario with low heritability, increased from 0.20, 0.15, 0.17 to 0.42, 0.40, 0.34, for, respectively, the scenarios without dominance (d = 0), and those with dominance equal to 0.5 and 1.

For the autogamous population (S_1 to S_4), the averages of reliability values previously obtained with strategy 1 were 0.55 (d = 0), 0.49 (d = 0.5) and 0.48 (d = 1) and increased to 0.54 (d = 0), 0.53 (d = 0.5) and 0.52(d = 1) using strategy 2. For the hybridization set (Bc_1 to Bc_4), GS reliability was 0.50(d = 0), 0.41 (d = 0.5) and 0.43 (d = 1) with strategy 1, staying around 0.46 (d = 0), 0.50 (d = 0.5) and 0.45(d = 1) with strategy 2. The results for the scenarios with higher heritability show same pattern that those observed for scenarios with low heritability ($h^2 = 0.30$). For higher heritability scenarios ($h^2 = 0.70$), the calibration strategy 2 outperform the results obtained from strategy 1 for all mating systems (S1 Table).

Calibration using multigenerational populations

The reliability values considering the utilization of combined generations in the calibration set, here called multigenerational sets, for all different scenarios, are presented in Table 3. This strategy increased the reliability values in all populations assessed for the scenarios with low heritability in the three degrees of dominance. For the scenarios with no dominance, the average of reliability values increased from 0.20 (strategy 1) to 0.42 (strategy 2) and 0.63 (strategy 3) in outcrossing (allogamous) populations. In self-pollination (autogamous) and hybridization (backcrossing) populations, the averages of values increased from 0.54 (strategies 1 and 2) to 0.64 (strategy 3), and the reliability values increased from 0.50 (strategy 1) to 0.71 (strategy 3), although decreased to 0.46 (strategy 2).

In the scenarios with dominance (d = 0.5), the average of reliability values increased from 0.15(strategy 1) to 0.40 (strategy 2) and 0.48(strategy 3) in allogamous populations. In autogamous and hybridization populations, the average of values increased from 0.49(strategies 1), and 2) to 0.53 (strategy 2) to 0.64 (strategy 3), and the reliability values increased from 0.41 (strategy 1) to 0.50 (strategy 2) to 0.62 (strategy 3).

Table 3. Reliability values of selection of generations advanced by self-pollination (S_n) random mating (A_n) and hibridization (Bc_n) obtained from phenotyping and genotyping of combined previous generations (multigenerational) or only from previous genotyping and phenotyping for heritability traits equal to 0.30, repeated in scenarios with an average degree of dominance level equal to 0, 0.5 and 1.

dominance	0				0.5				1			
Multigerational Training	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
F_2S_1	0.64 ± 0.10	0.62 ± 0.10	0.60 ± 0.10	0.59 ± 0.08	0.65 ± 0.06	0.58 ± 0.06	0.57 ± 0.06	0.55 ± 0.06	0.61 ± 0.07	0.57 ± 0.06	0.56 ± 0.07	0.56 ± 0.04
$F_2S_1S_2$		0.66 ± 0.11	0.65 ± 0.10	0.64 ± 0.09		0.65 ± 0.07	0.63 ± 0.06	0.63 ± 0.05		0.64 ± 0.08	0.63 ± 0.07	0.62 ± 0.05
$F_2S_1S_2S_3$			0.71 ± 0.08	0.69 ± 0.08			0.71 ± 0.08	0.70 ± 0.07			0.69 ± 0.11	0.68 ± 0.09
Average Set	0.64 ± 0.07				0.64 ± 0.06				0.63 ± 0.07			
Multigerational Training	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
F_2A_1	0.61 ± 0.10	0.64 ± 0.10	0.57 ± 0.10	0.62 ± 0.10	0.53 ± 0.09	0.39 ± 0.03	0.38 ± 0.08	0.36 ± 0.11	0.48 ± 0.10	0.37 ± 0.06	0.36± 0.06	0.36 ± 0.10
$F_2A_1A_2$		0.62 ± 0.10	0.63 ± 0.10	0.64 ± 0.10		0.55 ± 0.10	0.50 ± 0.10	0.47 ± 0.10		0.51 ± 0.10	0.44 ± 0.10	0.43 ± 0.10
$F_2A_1A_2A_3$			0.68 ± 0.08	0.68 ± 0.10			0.63 ± 0.10	0.58 ± 0.08			0.60 ± 0.09	0.53± 0.11
Average Set	0.63 ± 0.10				0.48 ± 0.11				0.45 ± 0.10			
Multigerational Training	Bc_1	Bc_2	Bc_3	Bc_4	Bc_1	Bc_2	Bc_3	Bc_4	Bc_1	Bc_2	Bc_3	Bc_4
F_2Bc_1	0.71 ± 0.04	0.67 ± 0.01	0.68 ± 0.01	0.67 ± 0.00	0.61 ± 0.05	0.55 ± 0.02	0.58 ± 0.03	0.57 ± 0.00	0.54 ± 0.10	0.47 ± 0.10	0.54 ± 0.09	0.51 ± 0.10
$F_2Bc_1Bc_2$		0.75 ± 0.02	0.73 ± 0.01	0.70 ± 0.01		0.66 ± 0.07	0.65 ± 0.03	0.62 ± 0.01		0.60 ± 0.10	0.59 ± 0.06	0.55 ± 0.08
$F_2Bc_1Bc_2Bc_3$			0.79 ± 0.02	0.73 ± 0.01			0.71 ± 0.06	0.66 ± 0.02			0.65 ± 0.07	0.58 ± 0.07
Average Set	0.71 ± 0.02				0.62 ± 0.02				0.55 ± 0.02			

In the scenarios with dominance (d = 1), the averages of reliability values increased from 0.17 (strategy 1) to 0.34 (strategy 2) and 0.45 (strategy 3) in allogamous populations. In autogamous and hybridization populations, the averages of values were 0.48(strategy 1), 0.52 (strategy 2) and 0.63 (strategy 3), and the reliability values increased from 0.43 (strategy 1) to 0.45 (strategy 2) 0.55 (strategy 3). The results for the higher heritability scenarios (0.7) follow the same pattern of reliability values. It can be seen in supplementary material (S2 Table).

For the group of allogamous populations, the inclusion of only the first random-mating generation enabled the averages of reliability values to around 0.50 within d = 0.5 or without dominance. The inclusion of phenotypic data with advanced generations as 2 or 3 previous generations (F_2, A_1, A_2, for example) improved even more the reliability values in the advanced generations A_3 and A_4. In the autogamous populations, the inclusion of two or more previous generations in the training allowed obtaining reliability above 0.61 in all three scenarios, as showed in Table 3. For the hybridization strategy 3 increased the reliability values to above 0.45 within or without dominance. Overall, the inclusion of only one more generation with the F_2 population sufficed to increase the prediction process over all scenarios.

Discussion

Our study was designed to maximize GS reliability in terms of the available genetic material. In contrast to previous studies on GS, we do not compare different prediction methodologies, such as G-BLUP, and different Bayesian approaches. In this study, we investigate different calibration strategies using combinations between one or more generations from different genetic backgrounds to improve the reliability of GS predictions. Furthermore, we investigated the effects of LD in different types of mating systems on the reliability of GS predictions. We used the G-BLUP on simulated data sets to access the reliability values for all evaluated scenarios.

Studying LD in all populations

In the F_2 population, the linkage disequilibrium was, solely, provided by the factorial linkage with the formation of 10 blocks of disequilibrium corresponding to the 10 linkage groups simulated for the genome (Fig 2). These factorial linkage effects on disequilibrium occur in different ways in successive generations of outcrossing or self-pollination. For example, in allogamous populations and with the progression of generations, it was possible to notice that disequilibrium was almost entirely dissipated, probably leading to low efficiency of procedures, such as GS, whose principle was the existence of disequilibrium between the marker and genes of interest.

According to the study of linkage disequilibrium carried out by Sorkheh and Malysheva-Otto (2008), random mating is an impactful factor in the decrease of linkage disequilibrium over time. Thus, conservation of linkage disequilibrium values in autogamous and backcrossing populations is justified by the fact that, for an effective recombination, double heterozygotes are required, and these are much more common in allogamous than in autogamous species. In addition, for autogamous species, linkage disequilibrium extends over large physical distances in comparison to outcrossing species, as verified by [19,20,21,22,23,24] which justifies the fact that the means of LD among all pairs of loci is 0.006 for allogamous and 0.051 for autogamous species. We believed that the impact of random mating on linkage disequilibrium can justify the importance of outdated calibration sets in allogamous populations.

Calibration using the F_2 population

We analyzed the reliability values for all the populations trained using F_2 in scenarios in which heritability were equal to 0.3 and 0.7 with and without dominance. In this analysis, we sought

to emphasize some important issues. The first one refers to the extent to which disturbing factors, such as dominance and environmental noise (influencing character heritability), can affect the reliability of the selection process via GS within the F_2 generation itself. However, the most important issue is the extent to which the GS efficacy is compromised when one goes from one generation to another where the loss of disequilibrium will be manifested, as such a loss is more intense in random-mating generations than in those from hybridization and self-pollination. These results suggest that the type of crossing shows a major contribution to the low reliability values.

Overall the presence of dominance also contributes to decrease in the prediction process, which is aggravated in the outcrossing populations in all three strategies. Almeida Filho [25], who simulated a genetically similar population of loblolly pine to assess the predicting ability of polygenic and oligogenic traits controlled by different degrees of dominance, also observed a drop in the prediction process reliability across generations. Their study showed that prediction in subsequent progeny population improved only in dominance superior to 0.2 and for oligogenic traits. In our study, genomic prediction reliability the assessment encompassed traits measured in the 12 populations that were previously genotyped and phenotyped for each trait. As displayed in Table 1, the reliability of the predictions increased when we changed heritability from 0.3 to 0.7. However, when we increased dominance from 0 to 1, the values decreased in the majority of scenarios. For example, in particular cases with allogamous species, the scenario with heritability 0.7 and dominance equal to 1 the prediction values were three times lower than those obtained for self-pollination and hybridization (0.65 and 0.52). This is expected because we know that dominance is a disturbing factor, since even when dominance is including in the model the reliability depends on considering heterosis and future combinations between parents. In general, the results showed a decrease in the reliability of total genomic predictions when the dominance increased. Although dominance was contemplated in the linear model used, according to some authors [4, 25, 26, 27], its incorporation does not lead to an improvement in the prediction process reliability of complex traits. In addition, according to Toro and Varona [28] when working with polygenic traits in simulated populations, the additive-dominant models exceeded the additive models only in the first predictive generation. Therefore, all these authors concluded that dominance increases in terms of complexity of the models, but it does not increase in terms of reliability.

Efficiency of different calibration strategies for GS studies considering different kinds of mating

The drop in reliability observed for the predictions of successive generations of random mating, hybridization or self-pollination is expected not only because of the drop in the disequilibrium of subsequent generations, but also because of the lack of re-estimation of markers, i.e., a new genotyping. According to [29], if genomic selection is practiced for many generations the effect of the markers does not change but the proportion of the genetic variance explained by them declines. This will cause the rate of response to selection to decline, as found by [13] using simulation. Thus, the inclusion of the latest genotyping, i.e., of genotyped and phenotyped individuals in their own generation, leads to an increase in the reliability estimates and can even be used in the prediction of subsequent generations with the same population structure, as displayed in Table 3. According to [30], this occurs because, despite the useful feature of genomic selection that long term response is predictable as the marker allele frequencies are known, it ignores non-additive effects of the QTL which may cause a change in the gene substitution effect of the QTL and, therefore, in the apparent effect of the marker, as selection changes gene frequencies. Several authors change the index weights as selection proceeds

[29,30,31]. Naturally, it would be possible to continually re-estimate the marker effect and include additional markers–those that had been excluded in the initial index. In our case, we demonstrated, by using a heritability of 0.3, that investment in genotyping of a new generation is important to ensure the efficiency of the prediction process, especially in populations whose linkage disequilibrium had already disappeared, as observed in our study, as well as for the fourth generation of population structures used here. For instance [11]predicted the breeding value of descendants 10 generations after the generation in which the phenotypes were recorded. Consequently, the reliability of GEBVs, calculated using the relationship matrix derived from the pedigree, was close to zero, so the reliability of the GEBVs using marker data was independent of the pedigree relationships.

The best strategy for prediction in allogamous populations. The importance of population structures (allogamous, autogamous, and hybridization systems) was also investigated to allow inferring whether training and validation could be performed with different genetic backgrounds. Thus, considering a random-mating population, we observed that reliability is reduced with the progress of generations by loss of disequilibrium. However, this reduction may be reverted if information from other generations of populations with the same structure is added to the training process, thus increasing reliability we used here in strategy 3. Therefore, especially for allogamous species, population structure is more impactful than linkage disequilibrium [32].

These results were also observed by [33] who conducted their study using one set of 73,147 marker in rice to predict five traits in the 2012 dry and wet seasons. Their study showed that the level of family relationship between selection candidates and the reference population had a higher effect on the reliability of genomic values than on linkage disequilibrium *per se*. Similar results were also found by [11], who, by using a simulation with genes in equilibrium, found that the prediction effects of selection were not zero because the family relationships were important, even in the absence of linkage disequilibrium.

Depending on the family structure of the data, there may also be information on the within-family marker effect, as obtained from a linkage analysis, and this may also add to the reliability of the GEBVs. Hence, in practical terms, one can assume that, in allogamous species, the prediction of advanced generations from early training may not be advantageous due to the loss of disequilibrium, which according to [5], for any pair of linked polymorphic loci, LD decreases over generations because of accumulation of recombination. Consequently, the most interesting solution is a new, more contemporary genotyping combined with the information from a previous genotyping (Strategy 3). A more evident example is that in scenarios with a high dominance degree (d/a = 1), the inclusion of F_2 and the three generations can better predict A_4 than only F_2 and A_1. Thus, it seems that the population structure becomes even more important in scenarios of high dominance.

Conclusion

As a result, we improved the reliability values of all the scenarios tested with high complexity. We highlight the importance of calibrating the training model used in GS in order to improve prediction in scenarios with low heritability and high dominance.

Supporting information

S1 Table. Reliability values of selection of generations advanced by self-pollination(S_n), random mating (A_n) and hybridization (Bc_n) obtained from phenotyping and genotyping of the same generation (diagonally contemporary) or only from previous genotyping and phenotyping (outdated off-diagonally, horizontal reading) for the heritability trait equal

to 0.30 in the scenario with an average degree of dominance equal to 0, 0.5 and 1.
(DOCX)

S2 Table. Reliability values of selection of generations advanced by self-pollination (S_n) random mating (A_n) and hibridization (Bc_n) obtained from phenotyping and genotyping of combined previous generations (multigenerational) or only from previous genotyping and phenotyping for heritability traits equal to 0.70, repeated in scenarios with an average degree of dominance level equal to 0, 0.5 and 1.
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: Isabela de Castro Sant' Anna, Ricardo Augusto Diniz Cabral Ferreira, Moysés Nascimento.

Data curation: Isabela de Castro Sant' Anna, Moysés Nascimento.

Formal analysis: Isabela de Castro Sant' Anna, Moysés Nascimento, Gabi Nunes Silva.

Investigation: Isabela de Castro Sant' Anna, Ricardo Augusto Diniz Cabral Ferreira, Moysés Nascimento, Francyse Edith Chagas.

Methodology: Isabela de Castro Sant' Anna, Ricardo Augusto Diniz Cabral Ferreira, Vinicius Quintão Carneiro.

Software: Isabela de Castro Sant' Anna.

Supervision: Moysés Nascimento, Cosme Damião Cruz, Marciane Silva Oliveira.

Validation: Isabela de Castro Sant' Anna, Gabi Nunes Silva, Vinicius Quintão Carneiro.

Visualization: Isabela de Castro Sant' Anna, Vinicius Quintão Carneiro.

Writing – original draft: Isabela de Castro Sant' Anna, Ricardo Augusto Diniz Cabral Ferreira, Vinicius Quintão Carneiro.

Writing – review & editing: Isabela de Castro Sant' Anna, Cosme Damião Cruz, Marciane Silva Oliveira, Francyse Edith Chagas.

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–1829. PMID: 11290733

2. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017; 22 (11): 961–975. https://doi.org/10.1016/j.tplants.2017.08.011 PMID: 28965742

3. Resende MDV,Silva FF, Azevedo CF Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição Sobrevivência. 1st ed. Visconde do Rio Branco: Suprema, 2014.

4. Xu Y, Wang X, Ding X, Zheng X, Yang Z, Xu C, et al. Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice. 2018; 11(1):32/ s12284-018-0223-427

5. Sorkheh K, Malysheva-Otto LV, Wirthensohn MG, Tarkesh-Esfahani S, Martínez-Gómez P. Linkage disequilibrium, genetic association mapping and gene localization in crop plants. Genet Mol Biol. 2008; 31(4):805–814.

6. Calus MP, De Roos AP, Veerkamp RF. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 2008; 178(1):553–561. https://doi.org/10.1534/genetics.107.080838 PMID: 18202394

7. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH. Genomic selection using different marker types and densities. J Anim Sci. 2008; 86(10): 2447–2254. https://doi.org/10.2527/jas.2007-0010 PMID: 18407980

8. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014; 19(9):592–601. https://doi.org/10.1016/j.tplants.2014.05.006 PMID: 24970707

9. Dekkers JC. Marker-assisted selection for commercial crossbred performance. J Anim Sci. 2007; 85 (9):2104–2114. https://doi.org/10.2527/jas.2006-683 PMID: 17504955

10. Daetwyler HD, Villanueva B, Bijma P, Woolliams JA. Inbreeding in genome-wide selection. J Anim Breed and Genet. 2007; 124(6):369–376.

11. Habier D, Fernando RL, Dekkers JC. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007; 177(4):2389–2397. https://doi.org/10.1534/genetics.107.081190 PMID: 18073436

12. Kumar S, Bink MC, Volz RK, Bus VG and Chagné D.Towards genomic selection in apple (malus×domestica borkh.) Breeding programmes: prospects, challenges and strategies. Tree Genet. Genomes. 2012; 8(1):1–14.

13. Muir WM. Comparison of genomic and traditional blup-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet. 2007; 124(6): 342–355. https://doi.org/10.1111/j.1439-0388.2007.00700.x PMID: 18076471

14. Flint-Garcia SA, Thornsberry JM, Buckler IV ES. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003; 54(1):357–374.

15. Zhao H, Nettleton D, Soller M, Dekkers JC. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet Res. 2005; 86(1):77–87. https://doi.org/10.1017/S001667230500769X PMID: 16181525

16. Weir BS, Hill WG. Nonuniform recombination within the human beta-globin gene cluster. Am J Hum Genet.1986; 38(5):776.–778 PMID: 3013006

17. Cruz CD. Genes: a software package for analysis in experimental statistics and quantitative genetics. Acta Sci Agron. 2013; 35(3):271–276.

18. Azevedo CF, Nascimento M, Fontes VC, Silva FF, Resende MDVR, Cruz CD. GenomicLand: Software for genome-wide association studies and genomic prediction. Acta Scientiarum. Forthcoming.

19. Nordborg M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics. 2000; 154: 923–929. PMID: 10655241

20. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, et alStructure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences. 2001; 98(20):11479–84.

21. Goodman MM and Buckler ES. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A. 2001; 98(20): 11479–11484. https://doi.org/10.1073/pnas.201394398 PMID: 11562485

22. Garris AJ, Mccouch SR, and Kresovich S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (Oryza sativa l.). Genetics. 2003; 165(2):759–769. PMID: 14573486

23. Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013; 193(2):347–365. https://doi.org/10.1534/genetics.112.147983 PMID: 23222650

24. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the

25. Almeida Filho JE, Guimarães JF, Silva FF, de Resende MD, Muñoz P, Kirst M, et al. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity; 117(1):33–41. https://doi.org/10.1038/hdy.2016.23 PMID: 27118156

26. Denis M, Bouvet JM. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Theor Appl Genet. 2013; 9(1):37–51.

27. Santos VS, Martins Filho S, de Resende MD, Azevedo CF, Lopes PS, Guimarães SE, et al. Genomic prediction for additive and dominance effects of censored traits in pigs. Genet. Mol. Res., 2016, 15(4)/ gmr15048764.

28. Toro MA, Varona L. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol. 2010; 42(1):33.

29. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009; 136(2): 245–257. https://doi.org/10.1007/s10709-008-9308-0 PMID: 18704696

30. Goddard ME and Hayes BJ. Genomic selection. J Anim Breed Genet. 2007; 124: 323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x PMID: 18076469

31. Dekkers JC, Van Arendonk JA. Optimizing selection for quantitative traits with information on an identified locus in outbred populations. Genet Res. 1998; 71(3):257–275.

32. Wientjes YC, Veerkamp RF, Calus MP. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics. 2013; 193(2):621–631. https://doi.org/10.1534/genetics.112.146290 PMID: 23267052

33. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, et al. 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 lines. PLoS Genet. 2015; 11(2):e1004982. https://doi.org/10.1371/journal.pgen.1004982 PMID: 25689273