Genotyping marker density reduction is not an effective approach in long-term prediction-based breeding of cross-pollinated crops

Julio Cesar DoVale (juliodovale@ufc.br)
Federal University of Ceara: Universidade Federal do Ceara https://orcid.org/0000-0002-3497-9793
Humberto Fanelli Carvalho
Universidad Politecnica de Madrid
Felipe Sabadin
Virginia Tech: Virginia Polytechnic Institute and State University
Roberto Fritsche-Neto
International Rice Research Institute

Research Article

Keywords: Linkage disequilibrium, genomic recurrent selection, marker subsets, additive-dominant model, SCA-GBLUP

Posted Date: November 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1005929/v1

Genotyping marker density reduction is not an effective approach in long-term prediction-based breeding of cross-pollinated crops

Júlio César DoVale[a]*, Humberto Fanelli Carvalho[b], Felipe Sabadin[c], Roberto Fritsche-Neto[d]

[a] Department of Crop Science, Federal University of Ceará, Fortaleza, Ceará, Brazil

[b] Universidad Politecnica de Madrid, Madrid, Spain

[c] Virginia Tech: Virginia Polytechnic Institute and State University, Blacksburg, United States

[d] International Rice Research Institute (IRRI), Los Baños, Philippines

*Corresponding author

Júlio César DoVale

email: juliodovale@ufc.br

ORCID of the authors:

Julio César DoVale (JCD): 0000-0002-3497-9793

Humberto Fanelli Carvalho (HFC): 0000-0003-0745-7583

Felipe Sabadin (FC): 0000-0003-2937-1465

Roberto Fritsche-Neto (RFN): 0000-0003-4310-0047

ABSTRACT

Reductions of genotyping marker density have been extensively evaluated as potential strategies to reduce the genotyping costs of genomic selection (GS). Low-density marker panels are appealing in GS because they entail lower multicollinearity and computational time-consumption and allow more individuals to be genotyped for the same cost. However, statistical models used in GS are usually evaluated with empirical data, using "static" training sets and populations. This may be adequate for making predictions during a breeding program's initial cycles, but not for the long term. Moreover, to the best of our knowledge, no GS models consider the effect of dominance, which is particularly important for breeding outcomes in cross-pollinated crops. Hence, dominance effects are an important and unexplored issue in GS for long-term programs involving allogamous species. To address it, we employed two approaches: analysis of empirical maize datasets and simulations of long-term breeding applying phenotypic and genomic recurrent selection (intrapopulation and reciprocal schemes). In both schemes, we simulated twenty breeding cycles and assessed the effect of marker density reduction on the population mean, the best crosses, additive variance, selective accuracy, and response to selection with models (additive, additive-dominant, general (GCA), and specific combining ability (SCA)). Our results indicate that marker reduction based on linkage disequilibrium levels provides useful predictions only within a cycle, as accuracy significantly decreases over cycles. In the long-term, high-marker density provides the best responses to selection. The model to be used depends on the breeding scheme: additive for intrapopulation and additive-dominant or SCA for reciprocal.

Keywords: Linkage disequilibrium; genomic recurrent selection; marker subsets; additive-dominant model; SCA-GBLUP;

DECLARATIONS

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Process 104371/2019-6.

Conflicts of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Codes and datas availability: codes and supporting data is available in:

https://data.mendeley.com/datasets/96p3khhnzj/1

Ethical standards: Not applicable.

AUTHOR CONTRIBUTION STATEMENT

JCD wrote the manuscript, discussion, contributing to ideas and graphs. HFC and FS contributed to the analysis of empirical data, in writing, mainly discussion. RFN elaborated on the hypothesis, conducted the analysis of the simulated experiments, interpreted the results, and contributed to the writing. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

To the Queiroz College of Agriculture (University of São Paulo) and the Federal University of Ceará.

KEY MESSAGE

In genomic recurrent selection, the more markers, the better because they buffer the linkage disequilibrium losses caused by recombination over cycles, and consequently, provide higher responses to selection.

INTRODUCTION

Genomic selection (GS), using high-density, single-nucleotide polymorphism (SNP) chips, has been widely adopted in breeding programs to improve predictive ability and response to selection (Hayes *et al.*, 2009; Crossa *et al.*, 2010). However, these chips' genotyping cost is currently prohibitive, especially for low profitability species (Hou et al. 2020), and programs in poor regions of the world, such as those in Latin America, Africa, and Southwest Asia.

In this context, marker density reduction has been extensively evaluated as an alternative to reduce the genotyping costs of GS. This reduction can be carried out based on different criteria, for instance, marker effects (e Sousa et al. 2019), trait genetic architecture (Zhang et al. 2015), haplotype block analysis (Ma et al. 2016), genomewide association studies (GWAS) (Subedi et al. 2013), and linkage disequilibrium (LD) (Al-Tobasei *et al.*, 2021). Of course, all strategies have advantages and disadvantages. However, the LD criterion seems to be the more suitable strategy because it eliminates markers for redundancy that are highly correlated without interfering with marker effects and thereby reduces multicollinearity (an obstacle for highly dense SNP panels). Moreover, it avoids marker effect noise (Xu 2013), and reduces the probability of overfitting by having the marker reduction process be independent of the phenotype (Vallejo *et al*., 2018).

The use of low-density marker panels in breeding programs is also appealing due to a reduction in computational time-consumption that allows more individuals to be genotyped for the same cost (Gorjanc et al. 2015). Furthermore, with larger datasets of phenotyped and genotyped individuals, GS' predictive ability may be increasingly driven by LD rather than by linkage information (Hickey et al. 2014). Furthermore, several studies claim that it is possible to substantially reduce the number of markers while maintaining high predictive ability (Tayeh et al. 2015; Ma et al. 2016; e Sousa et al. 2019; Al-Tobasei et al. 2020). Thus, the use of genotyping techniques with highly discriminative markers could be a viable alternative for current, low-cost SNP array strategies, with the potential to increase the fraction of the genome captured in a cost-efficient manner (Elshire et al. 2011).

The studies mentioned above made it possible to obtain an informative marker subset with either empirical or simulated data. The results of this were often satisfactory, as they considered the LD within a generation in a fixed scenario, i.e., used fixed training sets (TS) in the early stages of a breeding program to infer across multiple breeding cycles. However, both selection and recombination events that occur during breeding cycles may change LD between markers and Quantitative Trait Loci (QTLs) (in the latter case, due to cross-over occurrences), especially in more complex traits (Xu 2013). As a result, there are changes in allelic frequencies and, consequently, allele substitution effects which may even cause allele fixation (Walsh and Lynch 2018).

Although previous studies, like those mentioned above, evaluated the effect of marker density on the performance of genomic prediction models, there are no reports of its effect on models that incorporate additive as well

as dominance effects in cross-pollinated crops under recurrent selection (RS) schemes. The vast majority of published papers so far have considered only additive models for genomic predictions, which may not represent the genetic complexity of all crops. Even though only additive effects are transmitted over generations, the *per se* performance of a population and its single-crosses depends on dominance deviations (Falconer and Mackay, 1996)., This is the basis of heterosis and is crucial in the expression of agronomic traits of cross-pollinated crops. (Bernardo 2010). Additionally, as estimated by marker effects, Galli *et al.* (2020) observed negative covariance in genetic values of parental lines versus their single-crosses. This reinforces the assertion that poor predictions are made about single-crosses' performance when they are solely based on the additive effects of the parents' genome.

Simulations are a suitable tool to assess the effect of marker density reduction in a long-term breeding program applying GS. Several factors can be controlled to make inferences on given genetic parameters over several breeding cycles in a fast, inexpensive and consistent way (Dai et al. 2020). Simulations are essential for breeding programs of cross-pollinated crops. These usually adopt RS-based methods for obtaining cultivars, either by intrapopulation recurrent selection (IRS) or reciprocal recurrent selection (RRS), and hence genetic progress is achieved very slowly over breeding cycles. Several published studies address data simulation in the conducted long-term breeding programs, but with different objectives. Some sought to evaluate methods for updating of training sets (e.g., Neyhart *et al*., 2016), others using RS to optimize GS (Muleta *et al*., 2019) and even relate the number of parents in the persistence of accuracy and genetic gains (Muller *et al*., 2017). However, a study involving IRS and RRS schemes, which are widely used and allow a closer approach to the daily life of cross-pollinated crops breeding programs, has not yet been carried out with information genomic.

The flexibility of the models and methods used in GS allows for the integration of allelic interaction effects, as the combining abilities of inbred parental lines (Werner et al. 2018). Furthermore, Reif *et al.* (2013) observed an increase in their model's predictive capacity when incorporating general combining ability (GCA) effects. Despite this, no studies of long-term breeding programs compare the superiority of models that incorporate specific combining ability (SCA) effects using the genomic best linear unbiased predictor (G-BLUP) model, including additive and dominance effects. Thus, our objectives were to: (i) assess the impact of marker density reduction on prediction accuracy and responses to selection in long-term breeding programs, (ii) assess whether the inclusion of dominance effects affects the main genetic parameters, and (iii) to compare genomic prediction models using combining ability, additive and additive-dominance effects over many cycles.

MATERIALS AND METHODS

To achieve our objectives, we worked with empirical data and simulation experiments.

Empirical data

Overview: we used genotyping chips to perform GS for grain yield in maize and evaluated the effects of different marker reduction strategies on GS accuracy.

Phenotypic data

Our study considered grain yield (GY, in Mg ha^{-1}) from two datasets of maize single-crosses carried out by the Helix Seeds Company (HEL) and the University of São Paulo (USP). The plots were manually harvested, and GY was corrected to 13% moisture. Each dataset was analyzed independently.

The HEL data were obtained from 452 single-crosses obtained from 111 inbred lines. The hybrids were evaluated in a randomized complete block design in five Brazilian sites during the 2015 season.

The USP data were composed of 903 maize single-crosses, obtained from a full diallel of 49 inbred lines. The single-crosses were evaluated in 2016 and 2017 at two sites (Piracicaba and Anhumas) in São Paulo State, Brazil. Field trials were performed in an augmented block design, each block consisting of 16 unique single-crosses and two commercial hybrids as checks. Single-crosses were evaluated at each site and year under two nitrogen fertilization levels (ideal and low N). These eight treatment combinations comprise what we herein call "environments". More details on the experimental design and cultivation practices for HEL and USP datasets can be found in Sousa *et al.* (2017) and Galli *et al.* (2020), respectively.

Genetic-statistical models to estimate BLUEs

Mixed model equations were used for the statistical analysis of single-cross phenotypes in each dataset. First, the joint analysis of each phenotype was performed to estimate genotype means across environments. Next, we fitted a mixed model to obtain the Best Linear Unbiased Estimator (BLUE) for each genotype's yield. We then estimated the adjusted means in each environment via the *breedR* package (Muñoz and Rodriguez, 2016) in R software (R Development Core Team, 2019). The analyses for the USP dataset were performed using this model:

$$\boldsymbol{y = Ql + Sb + Tc + Ug + Vi + \varepsilon}$$

where $\boldsymbol{y}$ is the vector of phenotypic values of single-crosses and checks; $\boldsymbol{l}$ is the vector of fixed effects of the environment (combination of site × year × N level); $\boldsymbol{b}$ is the vector of random effect of block nested within environment, where $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_b^2)$; $\boldsymbol{c}$ is the vector of fixed effects of checks; $\boldsymbol{g}$ is the vector of fixed effects of single-crosses; $\boldsymbol{i}$ is the vector of fixed effects of interaction checks × environments; $\boldsymbol{\varepsilon}$ is the vector of random residual effects, which are confounded in the final residual term, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{D_e})$. $\boldsymbol{Q}$, $\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are the incidence matrices for $\boldsymbol{l}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{g}$, and $\boldsymbol{i}$. We assumed an unstructured covariance matrix across environments for the residual term ($\boldsymbol{D_e}$).

For the phenotypic analysis of the HEL dataset, we used the following reduced model:

$$\boldsymbol{y} = \boldsymbol{Xr} + \boldsymbol{Tg} + \boldsymbol{Bs} + \boldsymbol{Hx} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{r}$ is the vector of block effect considered as random, where $\boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_r^2)$; $\boldsymbol{s}$ is the vector of random effects of the environment (sites), where $\boldsymbol{s} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_s^2)$; $\boldsymbol{x}$ is the vector of the random effects of single-crosses by environment interaction, where $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_x^2)$; and $\boldsymbol{\varepsilon}$ is the vector of error, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2)$. $\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{B},$ and $\boldsymbol{H}$ are incidence matrices for $\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{s}, \boldsymbol{x}$.

In both HEL and USP models, single-crosses' effects were considered random in order to obtain the variance components via the restricted maximum likelihood (REML/BLUP) procedure (Henderson 1975).

Genotypic data

All parental inbred lines (USP and HEL) were genotyped using the Affymetrix® Axiom Maize Genotyping array containing 616 K SNPs (Unterseer et al. 2014). As a standard quality control, markers with a low call rate (< 90%) and those found to be non-biallelic were removed. Missing data were imputed with the Beagle 5.0 algorithm (Browning et al. 2018). Next, loci showing heterozygosity in one or more individuals were removed. Single-cross genotypes were obtained *in silico* by genomic information from parental inbred lines. Finally, markers with minor allele frequency (MAF) < 0.05 were removed from the hybrid genomic matrix, resulting in 63,104 and 30,467 SNP markers for the USP and HEL datasets.

Marker reduction

Different SNP marker subsets were obtained via the following two approaches: (i) removing SNP markers based on the LD between them and (ii) applying the LA-GA-T genetic algorithm (Akdemir (2017) in the *STPGA R* package, which aims to select optimized training sets (OTS).

In the first approach, pairwise LD was calculated as the squared allele frequencies correlation (r^2). In order to remove redundant markers, we applied the following threshold r^2 values, 0.99, 0.90, 0.80, 070, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, and 0.01, using the *SNPRelate* package (Zheng et al., 2012). After performing quality control, we used the full SNP markers (M_Full) matrix as the benchmark scenario for downstream comparisons.

In the second approach, we applied a Singular Value Decomposition (SVD) using the number of eigenvalues that explained 98% of the variance, according to Pocrnic *et al.* (2016). To reduce computation time and represent only the LD among markers' physical components, we applied SVD + OTS within each chromosome. This procedure resulted in thirteen SNP matrices for use in genomic prediction models.

Genomic predictions and model comparisons

We used the thirteen SNP matrices to build genomic relationship matrices via the *snpReady* package (Granato et al. 2018). Then, we applied additive and additive-dominance G-BLUP models to perform genomic predictions of grain yield using the following equations, respectively:

$$\widehat{\boldsymbol{g}} = \mathbf{1}\mu + \boldsymbol{Z_a a} + \boldsymbol{\varepsilon}$$

$$\widehat{\boldsymbol{g}} = \mathbf{1}\mu + \boldsymbol{Z_a a} + \boldsymbol{Z_d d} + \boldsymbol{\varepsilon}$$

where $\widehat{\boldsymbol{g}}$ is the vector of adjusted environmental grain yield means of single-crosses described in the 'Phenotypic data' subsection; μ is the mean (intercept); $\boldsymbol{a}$ is the vector of additive genetic effects of the individuals, where $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{G_a}\sigma_a^2)$; $\boldsymbol{d}$ is the vector of dominance effects, where $\boldsymbol{d} \sim N(\mathbf{0}, \boldsymbol{G_d}\sigma_d^2)$; and $\boldsymbol{\varepsilon}$ is the vector of random residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_\varepsilon^2)$. $\mathbf{1}$ is the vector of ones; $\boldsymbol{Z_a}$ and $\boldsymbol{Z_d}$ are incidence matrices for $\boldsymbol{a}$ and $\boldsymbol{d}$. σ_a^2 is the genomic additive variance, σ_d^2 is the genomic dominance variance, and σ_ε^2 is the residual variance. $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ are the additive and dominance genomic relationship matrices, following the equations: $\boldsymbol{G_a} = \boldsymbol{W_A W_A'}/2\sum_{i=1}^{n} p_i\,(1-p_i)$ and $\boldsymbol{G_d} = \boldsymbol{W_D W_D'}/4\sum_{i=1}^{n}\big(p_i(1-p_i)\big)^2$, where p_i is the frequency of one allele of the locus i and $\boldsymbol{W}$ is the marker incidence matrix (VanRaden 2008). Genotypes in the $\boldsymbol{W_A}$ matrix were coded as 0 for homozygote A_1A_1, 1 for the heterozygote A_1A_2, and 2 for the homozygote A_2A_2. In the $\boldsymbol{W_D}$ matrix, genotypes were coded as 0 for both homozygotes and 1 for heterozygotes. The genomic prediction models were performed using the *sommer* package (Covarrubias-Pazaran 2016).

In order to compare marker reduction scenarios, we used model accuracy, effective population size (N_e), and several metrics of the genetic relationship among individuals. We obtained estimates of N_e based on the $\boldsymbol{G_a}$ matrix. Correlation among $\boldsymbol{G_a}$ matrices across scenarios were assessed with Mantel tests (*vegan* R package, (Oksanen et al. 2019)), as was the correlation among $\boldsymbol{G_d}$ ones. We used $\boldsymbol{G_a}$ and $\boldsymbol{G_d}$ calculated from the M_Full SNP marker matrix to evaluate clustering of single-crosses based on Euclidean distance. The *superheat R* package created genetic distance

heatmaps (Barter and Yu 2018) using M_Full, M_0.5, M_0.01, and M_OTS. Subsequently, the singles-crosses order was used to sort heatmaps M_0.5, M_0.01, and M_OTS.

The predictive ability of the genomic prediction models was estimated as the correlation between the predicted and the observed genotypic values via a cross-validated, alpha-based design (CV-α, Yassue *et al.*, 2020), which is an extension of the methodology presented by Shao (1993). This model validation procedure consists of assigning observations to train or validation folds (groups) in each replication using alpha-lattice sorting premises and evaluating the model's prediction in each assignment iteration performed. CV-α was developed in a way that allows for the creation of validation scenarios with two, three, or four replicates per group, regardless of the number of treatments. The number of groups is determined by the percentage of training and validation sets. Our assignments of observations to groups were based on the α(0,1) lattice design in order to reduce the concurrences of any two treatments in the same group across the replicates (Patterson and Williams 1976). Here we used a validation scenario with four replicates and five groups (i.e., 5-fold validation), and model predictive abilities were averaged across the different assignments of replicates to those groups.

Simulation experiments

Overview: We simulated long-term breeding programs that apply intrapopulation (IRS) and reciprocal (RRS) recurrent selection schemes-based, using genomic (GRS) and phenotypic recurrent selection (PRS). Different marker densities and model types were applied.

Characteristics of the founder population, trait genetic architecture, and genotyping chips

We simulated a population of maize single-crosses from inbred parental lines to perform phenotypic and genomic predictions for a quantitative trait. For this, we used the *AlphaSimR* package (Gaynor et al. 2020).The scripts can be found in the supplementary files.

A historical population founded by 1,000 unique individuals was simulated stochastically (i.e., without artificial selection), with genomic information consisting of ten chromosomes pairs each using a Markovian Coalescent Simulator (MaCS) (Chen et al. 2009). Thirty-one thousand biallelic segregating sites were considered uniformly distributed across chromosomes, and 1,000 segregating loci were randomly sampled as quantitative trait nucleotides (QTN). The founders were diploidized but not inbred. From the historical population, base populations were obtained to perform recurrent selection (RS) schemes on, as described below.

The target of the simulation was a quantitative trait, such as yield. The genetic parameters obtained by Hallauer *et al.* (2010) were used. Each QTN received randomly allocated additive and dominance effects. Genetic values for each genotype were obtained by summing all additive and dominance effects for all QTN. Additive effects (a) were sampled of a gamma distribution with scale and shape parameters equal to 1 and randomly assigned for each QTN. Similarly, dominance effects (d) for each QTN were computed by multiplying the absolute value of its additive effect (a_i) by locus-specific dominance degree (δ_i). Dominance degrees were sampled of a Gaussian distribution with $\delta_i \sim N(\mu_\delta, \sigma_\delta^2)$, where μ_δ is the average dominance degree equal to 0.50 and σ_δ^2 is the dominance variance equal to 1. Dominance effects were assigned for each QTN according to the equation below:

$$d_i = \begin{cases} 0, & if\ QTN\ is\ homozygous \\ \delta_i \times |a_i|, & if\ QTN\ is\ heterozygous \end{cases}$$

Phenotypic values were obtained by adding a random error sampled of a Gaussian distribution with mean equal to 0 and variance (σ_e^2) equal to 1, which was defined by broad-sense ($H^2 = 0.50$) and narrow ($h^2 = 0.33$) heritabilities.

In order to represent the genotyping chips, we used three marker densities sampled from the segregating loci as SNP: the first was 30K SNPs, which on average is roughly equivalent to 15 markers per cM; the second was 1,5K SNPs equivalent to 1 marker per cM; and the third, 0,75K SNPs equivalent to one marker for every 2 cM.

Breeding schemes: base population and burn-in phase

We used IRS and RRS, two main breeding schemes RS-based, to evaluate the effect of marker reduction on predictions over twenty cycles with steps conducted independently. To create the base population for the IRS scheme, we selected 15 parents from the historical population (Fig. 1) and built a full diallel to generate the base population (Cycle 0, C_0). For the RRS scheme, the base population was made similarly, but in this case, 100 parents were used for the diallel. Parents of the best single-crosses were selected to build two heterotic groups (HG). Regardless of the breeding scheme, stage, or cycle, the progeny of 50 individuals was considered (equivalent to one trial with 2 plots of 25 plants).

Fig. 1 Schematic representation of the recurrent selection breeding schemes applied in the simulations. After obtaining the historical population, base populations were selected from it randomly, and these served as the basis for the IRS (orange) and RRS (blue) schemes. Thus, information in black applies to both schemes.

Once the base populations were created, a burn-in stage followed. For the burn-in stage (not represented in detail in Figure 1), we used three traditional recurrent selection cycles, as described below:

IRS – 200 half-sib progenies were obtained from the base population; then, the top 30 progenies were selected, and random recombination of 1,500 self-fertilized individuals was simulated (S_1) from those 30 best parents; using the resulting population, the process was repeated two more times, finally resulting in the C_0 population.

RRS – 200 full-sib offspring were obtained through a crossing-plan with 200 unique parents from each HG. Then, the best single-crosses were selected, and we proceeded with the recombination (S_1) of the parents within HG. At the end of each cycle, the fixation index (Fst) was calculated to assess the genetic distance between the HGs.

Breeding cycles: genomic and phenotypic selection

After the three cycles of the burn-in phase, 3,000 individuals were randomly selected as each scheme's C_0 (Fig. 1). From those 3,000 individuals, 2,000 were assigned to the training set (TS) for genomic predictions. First, marker effects were estimated using the three different SNP-chip densities described earlier. For that, genotyping and phenotyping information was used. Then, these 3,000 individuals were also used as population C_0 to begin the subsequent stage, consisting of twenty cycles of RS (GRS or PRS).

In each cycle, for PRS simulations (in both schemes – IRS or RRS), breeding cycles were run as described in the burn-in phase. For GRS simulations, the genotype of the 200 parents allowed us to skip the generation of progeny, select directly based on genomic breeding values, and simulate planting only the selected individuals for recombination, closing one breeding cycle in just one growing season. Furthermore, based on the parents' genotype, we obtained *in silico* all 40,000 possible single-crosses genotypes (200 x 200), increasing the population size and the intensity of selection.

Genomic predictions, methods, and model comparisons

As in the equations mentioned in the 'Genomic prediction' subsection for empirical data, we used additive and additive-dominance G-BLUP models to perform genomic predictions using three SNP-chip densities in both breeding schemes. Additionally, for the RRS scheme, we modeled using combining ability effects. Thus, the genomic estimated value (GEV) of single-crosses was obtained using estimates of the general combining ability (GCA) and specific combining ability (SCA) of the parents, according to the following expression:

$$\boldsymbol{GEV}_{ij} = \boldsymbol{GCA}_i + \boldsymbol{GCA}_j + \boldsymbol{SCA}_{i\times j}$$

where $\boldsymbol{GEV}_{\mathrm{ij}}$ is the genomic estimated value of single-cross ij, $\boldsymbol{GCA}_{\mathrm{i}}$ is the general combining ability of parent i, $\boldsymbol{GCA}_{\mathrm{j}}$ is the general combining ability of parent j, and $\boldsymbol{SCA}_{i\times j}$ is the specific combining ability from the cross between i and

j parentals. Kronecker products were used to model the interaction between single-crosses' parental genomes to capture SCA effects (Acosta-Pech et al. 2017; Basnet et al. 2019).

We evaluated two methods at different levels for both IRS and RRS schemes. In the first, based on the GRS, we considered the three markers densities already mentioned (GRS_30K, GRS_1.5K, and GRS_0.75K). In the second, based on PRS, used as benchmark strategies, we consider two selection ways: RS_Traditional – traditional recurrent selection based on selection and recombination per cycle of the best phenotypically identified individuals, and RS_Drift – recurrent selection also conducted with phenotypic data, but from randomly identified individuals per cycle. In this sense, we call for RS_Traditional the phenotypic model while for RS_Drift the random model. The methods within each breeding scheme were simulated over 20 breeding cycles and replicated 20 times.

The average genetic values (population mean), the performance of the best-crosses, Fst between heterotic groups, additive genetic variance, and prediction accuracy were calculated in each breeding cycle. The prediction accuracy for methods GRS-based was calculated as a Pearson correlation between true genetic values and GEV estimates. For methods PRS-based, the prediction accuracy was computed as the square root of heritability (h^2). Furthermore, the response to selection was calculated by modified breeder's equation with other components that remained fixed between the studied methods:

$$RS = \frac{PA}{t}$$

where RS is the relative response to selection to the previous cycle; PA is prediction accuracy; and t is the breeding cycle time. We consider that a PRS (RS_Traditional or RS_Drift) cycle takes 1.75 years on average to complete, while 0.35 years is needed for GRS.

RESULTS

Marker density reduction effect - empirical data

The number of markers in the SNP sub-sets decreased substantially with the decreasing LD level cut-offs applied to the empirical data (Table 1). In general, in both datasets, the reduction was steadier leading up to the subset with 10% LD between markers (M_0.1) than after it. The matrix with all markers (M_Full) had approximately 11.5 and 3.3 times more markers (USP and HEL datasets, respectively) than the one with only 1% LD between markers (M_0.01). However, the subset obtained applying the algorithm to optimize training sets (M_OTS) had drastically fewer markers than M_0.01 (3.9 and 5.4 times less, in USP and HEL, respectively).

Table 1 Number of SNP-markers per chromosome as a function of the marker density reduction performed using different levels of linkage disequilibrium and the OTS algorithm. USP and HEL datasets.

Despite considerable marker reduction using LD level cut-offs, effective population sizes (Ne) did not change in either dataset when considering the different SNP subsets (reduction scenarios) for estimating this parameter (Table 2). In addition, the significant and high correlation coefficients among additive matrices ($\boldsymbol{G_a}$) (> 0.98 for both datasets) as well as among the dominance ($\boldsymbol{G_d}$) matrices (> 0.73 for USP and > 0.99 for HEL) across marker densities (Tables S1 and S2), indicates that the matrices with low-density markers are as representative as those with high density. Furthermore, comparing the genetic distance heatmaps across marker reduction scenarios, including the most extreme scenarios, shows that genetic relationships were not altered by marker density (Figs. S1 and S2).

Table 2 Effective population size (Ne) obtained using additive effect matrices (*A*) with marker density reduction performed using different linkage disequilibrium levels and the OTS algorithm. USP and HEL datasets.

Similarly, results showed that the predictive ability remained unchanged throughout the various marker reduction scenarios, including the most extreme ones (Fig. 2). Therefore, with approximately 1,500 markers (1,408 for USP and 1,734 for HEL), it was possible to predict maize single-cross performance with the same accuracy as that achieved with all markers after the quality control process. Additionally, greater predictive abilities were observed when

we adopted the additive-dominance model, regardless of marker density. On average, the compound model accuracy surpassed the additive one by 9.8% and 19% for the USP and HEL datasets, respectively.

Fig. 2 Predictive ability across marker density reduction scenarios based on different linkage disequilibrium levels and an OTS algorithm, using additive and additive-dominant models. **A** USP dataset **B** HEL dataset.

Marker density reduction effect - long-term IRS breeding simulations

In simulations of long-term IRS breeding, RS_Traditional surpassed that of GRS for the population mean (genetic mean), regardless of marker density (Fig. 3A). RS_Traditional' superiority started to become evident in the third cycle. In relation to GRS, the density of 30K SNP had superior average performance compared to the other two densities, reaching almost double at the end of the twenty breeding cycles. We also observed that the additive model tended to be superior to the additive-dominance one for this parameter. The IRS scheme conducted by the drift method had the worst performance and, unlike the others, it declined over the cycles.

Fig. 3 Simulation of 20 breeding cycles via PRS (Traditional and Drift) and GRS (with three marker densities) using the IRS breeding scheme with additive and additive-dominant G-BLUP models. **A** Population mean **B** Additive variance **C** Accuracy

The additive variance available for GRS increased up till the second breeding cycle and then consumed vertiginously, cycle after cycle, but at a lower rate than that observed for RS_Traditional (Fig. 3B). Thus, the results of this genetic parameter were exactly opposite to those observed in the previous one (population mean). On the other hand, the trend of the curves for the accuracy parameter was very similar to those obtained for the population mean (Fig. 3C). The difference is that the GRS with the highest marker density started better than the RS_Traditional. In general, after the second cycle, GRS showed abrupt reductions in reliability. As for RS_Traditional, there was a loss in accuracy over the cycles, but it was less steep. In contrast, RS_Drift led to accuracies tending to zero throughout the entire breeding process. Regardless of the genetic model used, GRS accuracy increased with the number of markers used. Furthermore, the additive model has slightly outperformed the additive-dominant one in accuracy for IRS long-term breeding.

Marker density reduction effect - long-term RRS breeding simulations

The simulations of breeding programs with the RRS scheme showed different results regarding population mean with IRS (Fig. 4A). Here GRS outperformed PRS, regardless of marker density and model adopted. The differences started to become bigger after the fourth breeding cycle. The curves referring to the performance of the best crosses behaved similarly to the population mean parameter (Fig. 4B). The best crosses obtained via GRS also outperformed those obtained via PRS. In general, the yield achieved by applying GRS exceeded 18 Mg ha^{-1}, on average 64% and 350% higher than that obtained by using RS_Traditional and RS_Drift, respectively. In addition, GRS allowed for the maintenance of greater genetic divergence (as measured by fixation index, Fst) between heterotic groups (Fig. 4C) than did PRS, especially from the eighth cycle.

Fig. 4 Simulation of 20 breeding cycles via PRS (Traditional and Drift) and GRS (with three marker densities) using the RRS breeding scheme with additive, additive-dominant, CGA, and SCA G-BLUP models. **A** Population mean **B** The best crosses **C** Fst between HGs **D** Additive variance in HG1 **E** Additive variance in HG2 **F** Accuracy

Unlike what was observed in the IRS scheme, additive variance (for both heterotic groups) tended to be consumed more quickly in the breeding program conducted using GRS than in the program applying PRS, regardless of marker density or model (Fig. 4D and 4E). Similar to the IRS scheme, the PRS led by drift consumed less genetic variability, and RS_Traditional showed greater accuracy considering the whole simulation, from the first to the last breeding cycle. (Fig. 4F). Once again, the accuracy tended to be higher with a higher marker density. Overall, the models did not significantly affect parameters analyzed in the RRS scheme. However, there was a slight tendency for the additive-dominant and SGA models' accuracy to be higher than the additive and GCA models.

Genetic gains in both breeding schemes

Our results show that genetic gains are directly proportional to marker density regardless of the scheme used (Fig. 5) concerning the selection response to selection after seven years of breeding. In the IRS breeding scheme, the genetic gain obtained with the GRS_30K SNP reached more than 100% than that obtained with GRS_1,5K SNP and approximately 250% with the GRS_0,75K SNP (Fig. 5A). Regarding RS (traditional or via drift), the genetic gain of GRS_30K reached almost 600% than that obtained by RS_Traditional. In this scheme, the additive model outperformed the additive-dominant one when the GRS was conducted with any mark density.

Fig. 5 Responses to selection after seven years of breeding via PRS (Traditional and Drift) and GRS (with three marker densities) using different G-BLUP models. **An** IRS **B** RRS breeding schemes

Genetic gains obtained in the RRS reached levels higher than those IRS schemes (Fig. 5B). However, the differences between the GRS methods were smaller in the RRS scheme. GRS_30K achieved 48% superiority over GRS_1.5K and 96% over GRS_0.75K. At the end of the seven years of breeding for GRS_30K, the dominant additive model surpassed by almost 22, 56, and 8% of the additive, GCA, and SCA models. Concerning PRS, the genetic gain of GRS_30K reached nearly 1,150% than that obtained by RS_Traditional. In this scheme, the additive-dominant model surpassed all mark density and was followed by the SCA model. RS_Drift generated genetic gains close to zero in both breeding schemes.

DISCUSSION

High-density genotyping chips can span the whole genome with markers, making it possible to capture the total genetic variability of a trait throughout the genome and allow predictions of individuals' breeding values without collecting their phenotypic data (Meuwissen et al. 2001).

Implementing low-density marker chips in breeding programs applying GS is a desirable, cost-saving strategy, given a reduction in genotyping costs. It is especially important in species with low profitability. Several studies have documented the benefits of using marker subsets in GS analyses (Gorjanc et al. 2015; Tayeh et al. 2015; Ma et al. 2016; Li et al. 2018; e Sousa et al. 2019). Thus, the finding that few markers were sufficient to achieve a predictive ability similar to that obtained with high marker density, without changing the genetic relationship estimates among individuals (Tables 1 and 2; Figs. S1 and S2) spells significant advantages for low-density chips. Their use could lead to technological improvements and diminishing costs (Heaton et al. 2005; Chessa et al. 2007), as well as reduced computational demand.

Nevertheless, over breeding cycles, recombination between markers and QTLs will cause LD decreases, while selection and drift will potentially act to generate new LD or to strengthen the LD between closely linked loci (Hill and Robertson, 1968; Lorenz *et al.*, 2011). Thus, there is a need to study the long-term effects of reducing marker density, particularly given that long-term recurrent selection reduces the effectiveness of GS. Simulations of different reduction scenarios and controlled factors may do this quickly and at a lower cost.

Consequences of marker reduction for GRS in long-term breeding schemes

Simulated breeding program conducted with the IRS scheme

When a breeding program is conducted using recurrent selection on a single population (IRS), the aim is to release open-pollinated varieties into the field. Hence, the population means, genetic variance, and accuracy are essentials parameters for evaluating IRS scheme performance. However, at the end of the breeding cycles, our results showed that GRS applied in an IRS scheme is inferior to phenotypic selection. Probably because RS_Traditional is driven based on owner phenotypes. The GRS, on the other hand, is conducted based on the effect of the markers to predict the phenotype of supposedly superior individuals. As several recombinations are necessary over the breeding cycles, the effect of the markers must lose efficiency in predicting phenotype, which is reflected in these parameters.

GRS accuracy with higher markers density was superior in the first cycle but decreased abruptly from the second breeding cycle. That suggests that the model training set (TS) must be re-calibrated (updated) during the recurrent selection process to maintain predictive ability (Neyhart et al. 2017). Indeed, simulated experiments with eucalyptus

(Jannink 2010), barley (Denis and Bouvet 2013), and empirical data from advanced-cycle rye (Auinger et al. 2016) revealed that the accuracy of GS was improved by recalibration of the TS. However, each recurring selection cycle's recombination events change the QTL-marker LD pattern, diminishing its usefulness for GS. Furthermore, changes in allele frequencies due to selection over breeding cycles impact the estimation of the effects of allelic substitutions and, consequently, affects genomic prediction accuracy.

The above situation can be aggravated by very low initial marker density, compromising the TS's recalibration. However, our results show that loss of accuracy over breeding cycles was lower when higher marker densities were used (Fig. 3C). Furthermore, Müller *et al.* (2017) used simulations to show the persistence of predictive accuracy for a longer timeframe and, consequently, better selection outcomes when higher marker density was used. Thus, high marker density seems to buffer the effect of the loss of LD between markers and QTLs over the recombination cycles more effectively than low marker density, even without TS recalibration. Although, however, there is an expectation that genotyping costs for low-density chips are lower than those for high-density chips. In practice, the actual differences in terms of genotyping cost per sample among these chips are negligible. In light of this, and given our results, we stipulate that the economic gain between high- and low-density chips do not compensate for the prediction accuracy loss in the case of long-term breeding programs.

Simulated breeding program conducted with the RRS scheme

When a breeding program is conducted using an RRS scheme, two populations are improved simultaneously due to reciprocal crosses. At first, the frequencies of desirable alleles in each population are maximized, improving performance via additive effects. Subsequently, the populations are crossed to capitalize on the non-additive effects of heterozygosity (Hallauer *et al*. 2010). Thus, RRS is more complex than IRS. However, when conducted with GS support, it can be carried out with efforts similar to those requiring programs applying an IRS scheme. Population performance was better with RRS than IRS (Figs. 3A and 4A), likely due to more significant heterosis in the former than in the open-pollinated individuals of the latter.

The performance of single-crosses selected by GRS was higher than those selected using phenotypic selection (Figs. 4A and 4B). This may be explained by the fact that GRS allows *in-silico* generation of all single-cross combinations (40,000 in the present study), a process that increases the chances of identifying the best crosses but is unfeasible with PRS (traditional or via drift). Furthermore, genomic information makes it possible to distinguish between alleles belonging to different heterotic groups and hence allows for the maintenance of divergence over breeding cycles (Fig. 4C). Given the selection of individuals is carried out on single-crosses, which have high heterotic effects, GRS is better equipped to capture complementarity, heterozygosity, and genetic distance elements that may contribute to an individual's

performance. That is important for a breeding program that aims to release single-crosses, since those are the main components of heterosis (Moll et al. 1965; Prasad and Singh 1986).

Long-term RS may reduce genetic variance to the point where genetic gains are limited (Jannink, 2010). Indeed, genetic variance showed a substantial decline over breeding cycles in both selection schemes and methods. However, in the RRS scheme, GRS resulted in more substantial losses of genetic diversity than PRS (Figs. 4D and 4E). Muleta *et al.* (2019) observed similar results, with higher losses in the simulation of an oligogenic trait with high heritability and lower losses in a polygenic trait with low heritability. The authors attributed this finding to higher selection intensity in GS, given more cycles per year (two and three cycles of GS per year were simulated). Essentially, selection intensity (standardized for cycles per year) will be higher in GRS than in PRS because it has a larger sample size, as previously mentioned.

The maintenance of prediction accuracy across selection cycles is critical for long-term genetic gain (Müller *et al*., 2017). Interestingly, in our study, RS_Traditional accuracy surpassed that of GRS throughout cycles. Mathematically, accuracy is a function of additive variance; a reduction in variance causes a decrease in prediction accuracy. Our finding may be explained by the breakdown of LD between markers and QTLs due to recombination (Jannink, 2010). Moreover, over cycles to selection and drift of alleles are not tagged by markers in TS, which cannot be targeted by GRS (Rutkoski *et al.*, 2015), culminating in low accuracy. Another suggested explanation for this result is that selection is based on genetic divergence, in which the alleles present in individuals from both heterotic groups change in direction and sense (upper and lower tail of the genetic variation curve), affecting the reliability of the GS.

As seen with the IRS scheme, we noted that higher marker density results in higher accuracy over breeding cycles with RRS. In addition, a higher number of markers should better capture the LD between QTLs and markers since some markers may dissociate from the QTLs due to recombination during the cycles, while others may not, keeping some level of LD.

Comparison between prediction models in response to selection

We consider in this study that a PRS cycle takes 1.75 years on average to complete while 0.35 years is needed for GRS. Regarding time in a study of this nature is important because each method has a different length. Therefore, weighting for the time factor makes comparisons fairer (realistic). Thus, over seven years of breeding, it would be possible to carry out the twenty cycles by GRS, while PRS only four.

Predictions obtained using GRS tended to be more accurate than those obtained using PRS, as the former considers the real genetic relatedness between genotypes rather than the average expectation. Studies have been conducted with empirical data (Al-Tobasei et al. 2020) and with simulations (Hou et al. 2020) for animal and plant

breeding to understand markers' effects on the accuracy of genomic predictions. However, these only consider additive genetic effects to identify marker subsets and, consequently, to make the predictions.

Several other studies have already shown that information about dominance deviations incorporated via kernels is essential to increase the efficiency of GS prediction models (Azevedo et al. 2015; Dias et al. 2018; Matias et al. 2018; Dai et al. 2020), especially when the species is cross-pollinated and heterosis is explored in the final product (Santos *et al.*, 2016). Our empirical data findings (Fig. 2) corroborate others that show the importance of incorporating additive effects and dominance deviations in the models to increase predictive ability (Technow et al. 2012; Alves et al. 2019). In the presence of dominance, the additive-dominant model is expected to be more accurate, as the additive model falsely assumes that residuals are independent and identically distributed (Duenk *et al*., 2017). Furthermore, additive-dominant models allow for finer-grain assessment of genomic contributions to performance than models that only consider additive effects. However, when we consider only the additive kernel in the models, additive effects, dominance, and residuals are not readily distinguishable from one another (Alves *et al*., 2019).

We found in long-term breeding simulations that the additive model was superior primarily in terms of accuracy to the additive-dominant model for the IRS scheme. This result is most evident in response to selection (Fig. 5) and makes sense, as the products generated by this scheme (open-pollinated varieties and inbred lines) effectively exploit additive effects (Hallauer et al., 2010). Contrasting the RRS scheme, the additive-dominant model was superior to those accompanied by the SGA model. Analogously to the previous explanation, this scheme's genetic material (single-crosses) effectively exploits additive and non-additive genetic effects. Thus, a model that better captures and separates these effects should generate the best responses to selection. Furthermore, Duenk *et al.* (2017) showed that in the presence of dominance, the mean squared error of the average effect of allelic substitution with the additive-dominant model was always smaller, especially when the heritability was low. Therefore, for cross-pollinated species in which single-crosses are commercially exploited, it can be expected that genetic models that include dominance will be more robust and have higher accuracy in estimating the average effect of allelic substitution purely additive models.

GS is expected to be not consistently well-suited for RRS since, in the modeling process, only one vector of additive effects and dominance deviations is considered for the entire population and their respective heterotic groups. Thus, we use two more approaches to simulate breeding cycles with RRS; G-BLUP modeled only with the GCA effects and another with SCA effects. Furthermore, based on quantitative genetics theory, with the application RS_Traditional, individuals from both heterotic groups can be expected to be selected for breeding based on the same allele's effects, which might lead to a reduction in the genetic distance among HG over the cycles, vanishing the possibility of the materialization of heterosis (Hallauer et al. 2010).

Breeders that exploit heterosis to generate products (especially single-crosses) can theoretically be expected to benefit from models that can predict non-additive genetic effects, as this would allow more accurate predictions of the

performance of single-crosses. In essence, the two models that best capture dominance deviations are the additive-dominance model and the SCA one. These models presented similar estimates for the parameters studied, such as the population mean, the best crosses, and Fst. Werner et al. (2018) used different models to verify the increase in predictive ability in oilseed rape with the GS approach. Interestingly, that study found that incorporating SCA effects combined with GCA into the RR-BLUP method (similar to G-BLUP) was the approach that caused the most fluctuation in predictions over 200 validation cycles. Except for this, all methods (including those based on Bayesian statistics) revealed comparatively robust predictive ability and high accuracy.

We found different trends among the results generated with the models employed after seven years of breeding. The responses to selection were different, mainly for GRS_30K. However, the selection process is standardized in the RRS scheme, i.e., we select the best single-crosses first and then recombine their parents. Even under these circumstances, it is important to be aware of how to model. The smallest difference between the additive-dominant and SCA models must be related to the fact that the dominance deviations are captured by SCA effects for heterosis (Falconer and Mackay, 1996).

Additionally, it is well-known that non-additive effects are partially included in the parents breeding value, GCA, or in allele substitution effects (Falconer and Mackay, 1996; Bernardo, 2010; Werner et al., 2018). Despite this, it is noteworthy that even if, in some situations, the additive-dominance model achieves lower performance than the additive one, the additive-dominance model generates more reliable estimates (Alves et al. 2019). Reif et al. (2007) demonstrated that the dominance variance decreases the respective additive variance but increases according to the populations' divergence. Thus, dominance effects are increasingly absorbed by the population mean or become inseparable from additive effects over breeding cycles (Technow et al. 2014).

Final considerations

Disentangling the different factors that may contribute to heterogeneity and predictive ability failure in empirical datasets is challenging (Dai et al. 2020). However, several models and tools are available to minimize bias. Using simulations, we can control several uncontrolled factors in the field and focus on answering specific questions. Our results clearly showed that reducing the marker density is efficient only in a restricted timeframe (within a breeding cycle or a static group of parents). Hence, low-density marker sets are ineffective for GRS (either with IRS or RRS schemes), usually conducted for long-term breeding programs, and that does not aim to reach a plateau as inbreeding schemes to increase homozygosity. Overall, the higher the marker density, the higher the capacity of the marker set to retain relevant genetic information while being subject to the LD losses between markers and QTLs that result from crossing-over events during recombination cycles. This brings a direct gain (linear and ascending) for the response to selection. Therefore, this

study complements the results of others published recently. Although GS allows a higher genetic gain, especially when weighted by time, Muleta et al. (2019) also observed more significant accuracy losses in GRS than in PRS throughout the breeding cycles. In this sense, there is a need for more studies on how often and in what way to update the TS to maximize prediction accuracy (Rincent et al. 2012), retaining in satisfactory levels the LD between markers and QTLs (Neyhart et al., 2017), and the relationship between TS and target populations Finally, we recommend that genotyping should be carried out with high marker density, so that, in long-term breeding programs, the marker set's ability to capture LD over cycles is maintained, and does not represent a bottleneck. This may avoid the need to update the TS with each breeding cycle. The additive model is sufficient to predict breeding populations' performance in the IRS breeding scheme and the additive-dominance one when employed RRS scheme. Furthermore, regarding non-additive genetic effects, we found a slight difference between the additive-dominance and SCA G-BLUP models.

REFERENCES

Acosta-Pech R, Crossa J, de los Campos G, et al (2017) Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theor Appl Genet 130:1431–1440. https://doi.org/10.1007/s00122-017-2898-0

Akdemir D (2017) STPGA: Selection of training populations with a genetic algorithm. bioRxiv. https://doi.org/10.1101/111989

Al-Tobasei R, Ali A, Garcia A, et al (2020) Genomic Predictions for Muscle Yield and Fillet Firmness in Rainbow Trout using Reduced-Density SNP Panels. https://doi.org/10.21203/rs.3.rs-36925/v1

Alves FC, Granato ÍSC, Galli G, et al (2019) Bayesian analysis and prediction of hybrid performance. Plant Methods 15:1–18. https://doi.org/10.1186/s13007-019-0388-x

Auinger HJ, Schönleben M, Lehermeier C, et al (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). Theor Appl Genet 129:2043–2053. https://doi.org/10.1007/s00122-016-2756-5

Azevedo CF, de Resende MDV, e Silva FF, et al (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genet 16:1–13. https://doi.org/10.1186/s12863-015-0264-2

Bandeira e Sousa M, Cuevas J, Couto EG de O, et al (2017) Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. G3 Genes, Genomes, Genet 7:1995–2014. https://doi.org/10.1534/g3.117.042341

Barter RL, Yu B (2018) Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. J Comput Graph Stat 27:910–922. https://doi.org/10.1080/10618600.2018.1473780

Basnet BR, Crossa J, Dreisigacker S, et al (2019) Hybrid Wheat Prediction Using Genomic, Pedigree, and Environmental Covariables Interaction Models. Plant Genome 12:180051. https://doi.org/10.3835/plantgenome2018.07.0051

Bernardo R (2010) Breeding for quantitative traits in plants Stemma Press. Stemma Press, Woodbury

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Res 19:136–142. https://doi.org/10.1101/gr.083634.108

Chessa S, Chiatti F, Ceriotti G, et al (2007) Development of a single nucleotide polymorphism genotyping microarray platform for the identification of bovine milk protein genetic polymorphisms. J Dairy Sci 90:451–464. https://doi.org/10.3168/jds.S0022-0302(07)72647-4

Covarrubias-Pazaran G (2016) Genome-Assisted prediction of quantitative traits using the r package sommer. PLoS One 11:1–15. https://doi.org/10.1371/journal.pone.0156744

637 Crossa J, De Los Campos G, Pérez P, et al (2010) Prediction of genetic values of quantitative traits in plant breeding
638 using pedigree and molecular markers. Genetics 186:713–724. https://doi.org/10.1534/genetics.110.118521
639 Dai Z, Long N, Huang W (2020) Influence of genetic interactions on polygenic prediction. G3 Genes, Genomes, Genet
640 10:109–115. https://doi.org/10.1534/g3.119.400812
641 Denis M, Bouvet JM (2013) Efficiency of genomic selection with models including dominance effect in the context of
642 Eucalyptus breeding. Tree Genet Genomes 9:37–51. https://doi.org/10.1007/s11295-012-0528-1
643 Dias KODG, Gezan SA, Guimarães CT, et al (2018) Improving accuracies of genomic predictions for drought
644 tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity
645 (Edinb) 121:24–37. https://doi.org/10.1038/s41437-018-0053-6
646 Dos Santos JPR, De Castro Vasconcellos RC, Pires LPM, et al (2016) Inclusion of dominance effects in the
647 multivariate GBLUP model. PLoS One 11:1–21. https://doi.org/10.1371/journal.pone.0152045
648 Duenk P, Calus MPL, Wientjes YCJ, Bijma P (2017) Benefits of dominance over additive models for the estimation of
649 average effects in the presence of dominance. G3 Genes, Genomes, Genet 7:3405–3414.
650 https://doi.org/10.1534/g3.117.300113
651 e Sousa MB, Galli G, Lyra DH, et al (2019) Increasing accuracy and reducing costs of genomic prediction by marker
652 selection. Euphytica 215:. https://doi.org/10.1007/s10681-019-2339-z
653 Elshire RJ, Glaubitz JC, Sun Q, et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high
654 diversity species. PLoS One 6:1–10. https://doi.org/10.1371/journal.pone.0019379
655 Falconer D, Mackay T (1996) Introduction to quantitative genetics, Longman, 4th edn. Longmans Green, Harlow,
656 Essex, UK.
657 Galli G, Alves FC, Morosini JS, Fritsche-Neto R (2020) On the usefulness of parental lines GWAS for predicting low
658 heritability traits in tropical maize hybrids. PLoS One 15:1–15. https://doi.org/10.1371/journal.pone.0228724
659 Gaynor RC, Gorjanc G, Hickey JM (2020) AlphaSimR: An R-package for Breeding Program Simulations 2 3. bioRxiv
660 2020.08.10.245167
661 Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015) Potential of genotyping-by-sequencing for genomic
662 selection in livestock populations. Genet Sel Evol 47:. https://doi.org/10.1186/s12711-015-0102-z
663 Granato ISC, Galli G, de Oliveira Couto EG, et al (2018) snpReady: a tool to assist breeders in genomic analysis. Mol
664 Breed 38:. https://doi.org/10.1007/s11032-018-0844-8
665 Hallauer A, Carena M, Filho JM (2010) Quantitative genetics in maize breeding
666 Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle:
667 Progress and challenges. J Dairy Sci 92:433–443. https://doi.org/10.3168/jds.2008-1646
668 Heaton MP, Keen JE, Clawson ML, et al (2005) Use of bovine single nucleotide polymorphism markers to verify

669 sample tracking in beef processing. J Am Vet Med Assoc 226:1311–1314.
670 https://doi.org/10.2460/javma.2005.226.1311
671 Henderson CR (1975) Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31:423.
672 https://doi.org/10.2307/2529430
673 Hickey JM, Dreisigacker S, Crossa J, et al (2014) Evaluation of genomic selection training population designs and
674 genotyping strategies in plant breeding programs using simulation. Crop Sci 54:1476–1488.
675 https://doi.org/10.2135/cropsci2013.03.0195
676 Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231.
677 https://doi.org/10.1080/03071848408522227
678 Hou L, Liang W, Xu G, et al (2020) Accuracy of genomic prediction using mixed low-density marker panels. Anim
679 Prod Sci 60:999–1007. https://doi.org/10.1071/AN18503
680 Jannink JL (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:1–11. https://doi.org/10.1186/1297-
681 9686-42-35
682 Li B, Zhang N, Wang YG, et al (2018) Genomic prediction of breeding values using a subset of SNPs identified by
683 three machine learning methods. Front Genet 9:1–20. https://doi.org/10.3389/fgene.2018.00237
684 Lorenz AJ, Chao S, Asoro FG, et al (2011) Genomic Selection in Plant Breeding. Knowledge and Prospects., 1st edn.
685 Elsevier Inc.
686 Ma Y, Reif JC, Jiang Y, et al (2016) Potential of marker selection to increase prediction accuracy of genomic selection
687 in soybean (Glycine max L.). Mol Breed 36:1–10. https://doi.org/10.1007/s11032-016-0504-9
688 Matias FI, Barrios SCL, Bearari LM, et al (2018) Contribution of additive and dominance effects on agronomical and
689 nutritional traits, and multivariate selection on Urochloa spp. hybrids. Crop Sci 58:2444–2458.
690 https://doi.org/10.2135/cropsci2018.04.0261
691 Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genomewide dense marker
692 maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819
693 Moll R, Lonnquist J, Vélez Fortuno J, Johnson E (1965) The relationship of heterosis and genetic divergence in maize.
694 Genetics 52:139–144
695 Muleta KT, Pressoir G, Morris GP (2019) Optimizing genomic selection for a sorghum breeding program in Haiti: A
696 simulation study. G3 Genes, Genomes, Genet 9:391–401. https://doi.org/10.1534/g3.118.200932
697 Müller D, Schopp P, Melchinger AE (2017) Persistency of prediction accuracy and genetic gain in synthetic
698 populations under recurrent genomic selection. G3 Genes, Genomes, Genet 7:801–811.
699 https://doi.org/10.1534/g3.116.036582
700 Muñoz F, Rodriguez LS (2016) breedR : An open statistical package to analyse genetic data

Neyhart JL, Tiede T, Lorenz AJ, Smith KP (2017) Evaluating methods of updating training data in long-term genomewide selection. G3 Genes, Genomes, Genet 7:1499–1510. https://doi.org/10.1534/g3.117.040550

Oksanen J, Blanchet FG, Friendly M, et al (2019) Package 'vegan' Title Community Ecology Package. Community Ecol Packag 2:1–297

Patterson HD, Williams ER (1976) A new class of resolvable incomplete block designs. Biometrika 63:83–92. https://doi.org/10.1093/biomet/63.1.83

Pocrnic I, Lourenco DAL, Masuda Y, Misztal I (2016) Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genet Sel Evol 48:1–9. https://doi.org/10.1186/s12711-016-0261-6

Prasad SK, Singh TP (1986) Heterosis in relation to genetic divergence in maize (Zea mays L.). Euphytica 35:919–924. https://doi.org/10.1007/BF00028600

Reif JC, Gumpert FM, Fischer S, Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176:1931–1934. https://doi.org/10.1534/genetics.107.074146

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. https://doi.org/10.1534/genetics.112.141473

Rutkoski J, Singh RP, Huerta-Espino J, et al (2015) Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust Resistance in Wheat. Plant Genome 8:1–10. https://doi.org/10.3835/plantgenome2014.09.0046

Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494. https://doi.org/10.1016/j.jspi.2003.10.004

Subedi S, Feng Z, Deardon R, Schenkel FS (2013) SNP selection for predicting a quantitative trait. J Appl Stat 40:600–613. https://doi.org/10.1080/02664763.2012.750282

Tayeh N, Klein A, Le Paslier MC, et al (2015) Genomic prediction in pea: Effect of marker density and training population size and composition on prediction accuracy. Front Plant Sci 6:1–11. https://doi.org/10.3389/fpls.2015.00941

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. https://doi.org/10.1007/s00122-012-1905-8

Technow F, Schrag TA, Schipprack W, et al (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355. https://doi.org/10.1534/genetics.114.165860

Unterseer S, Bauer E, Haberer G, et al (2014) A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:1–15. https://doi.org/10.1186/1471-2164-15-823

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

Walsh B, Lynch M (2018) Evolution and Selection of Quantitative Traits

Werner CR, Qian L, Voss-Fels KP, et al (2018) Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. Theor Appl Genet 131:299–317. https://doi.org/10.1007/s00122-017-3002-5

Xu S (2013) Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195:1103–1115. https://doi.org/10.1534/genetics.113.155309

Yassue RM, Sabadin JFG, Galli G, et al (2020) CV-α: designing validations sets to increase the precision and enable multiple comparison tests in genomic prediction

Zhang Z, Erbe M, He J, et al (2015) Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 Genes, Genomes, Genet 5:615–627. https://doi.org/10.1534/g3.114.016261

Figures

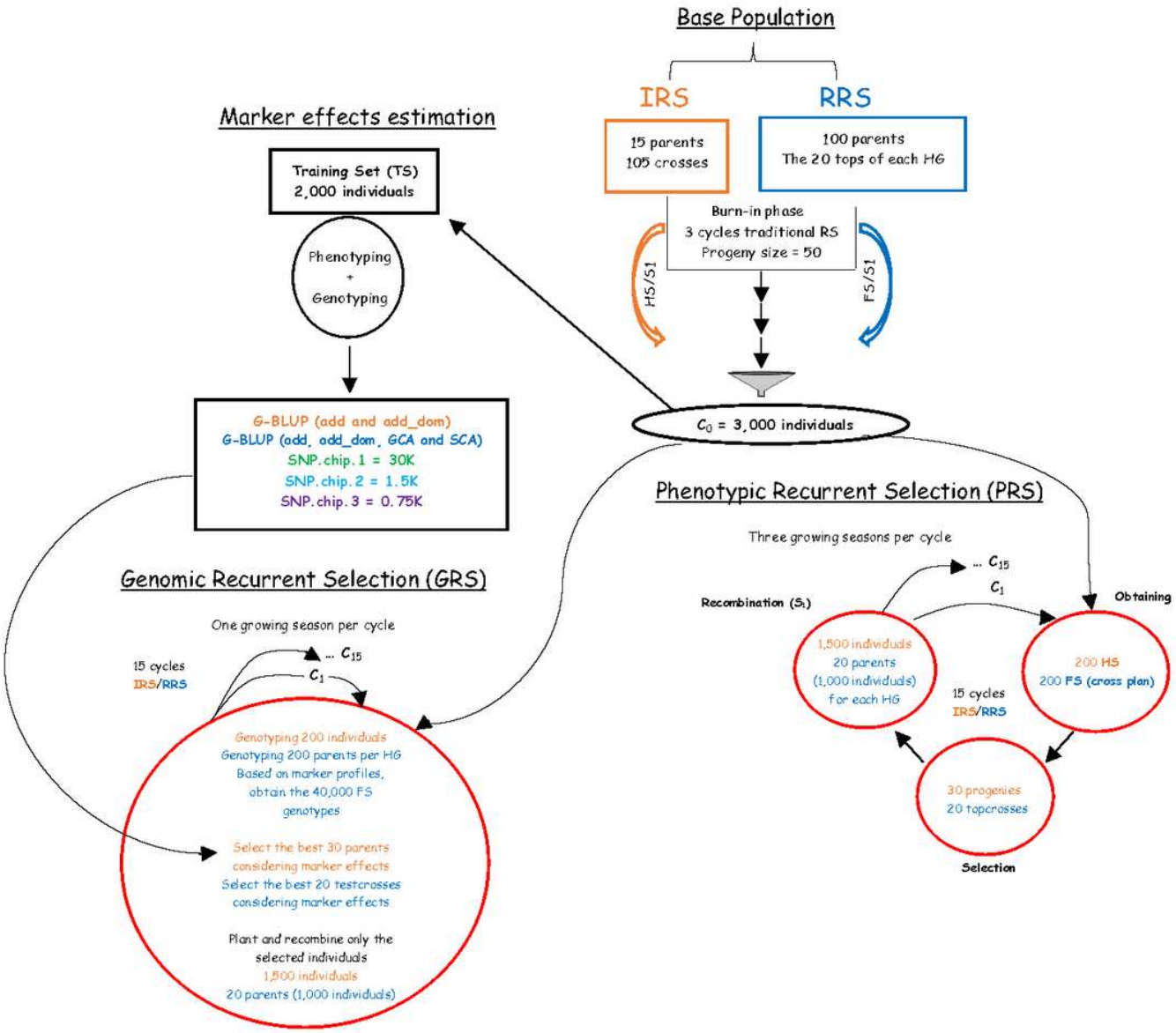

Figure 1

Schematic representation of the recurrent selection breeding schemes applied in the simulations. After obtaining the historical population, base populations were selected from it randomly, and these served as the basis for the IRS (orange) and RRS (blue) schemes. Thus, information in black applies to both schemes.

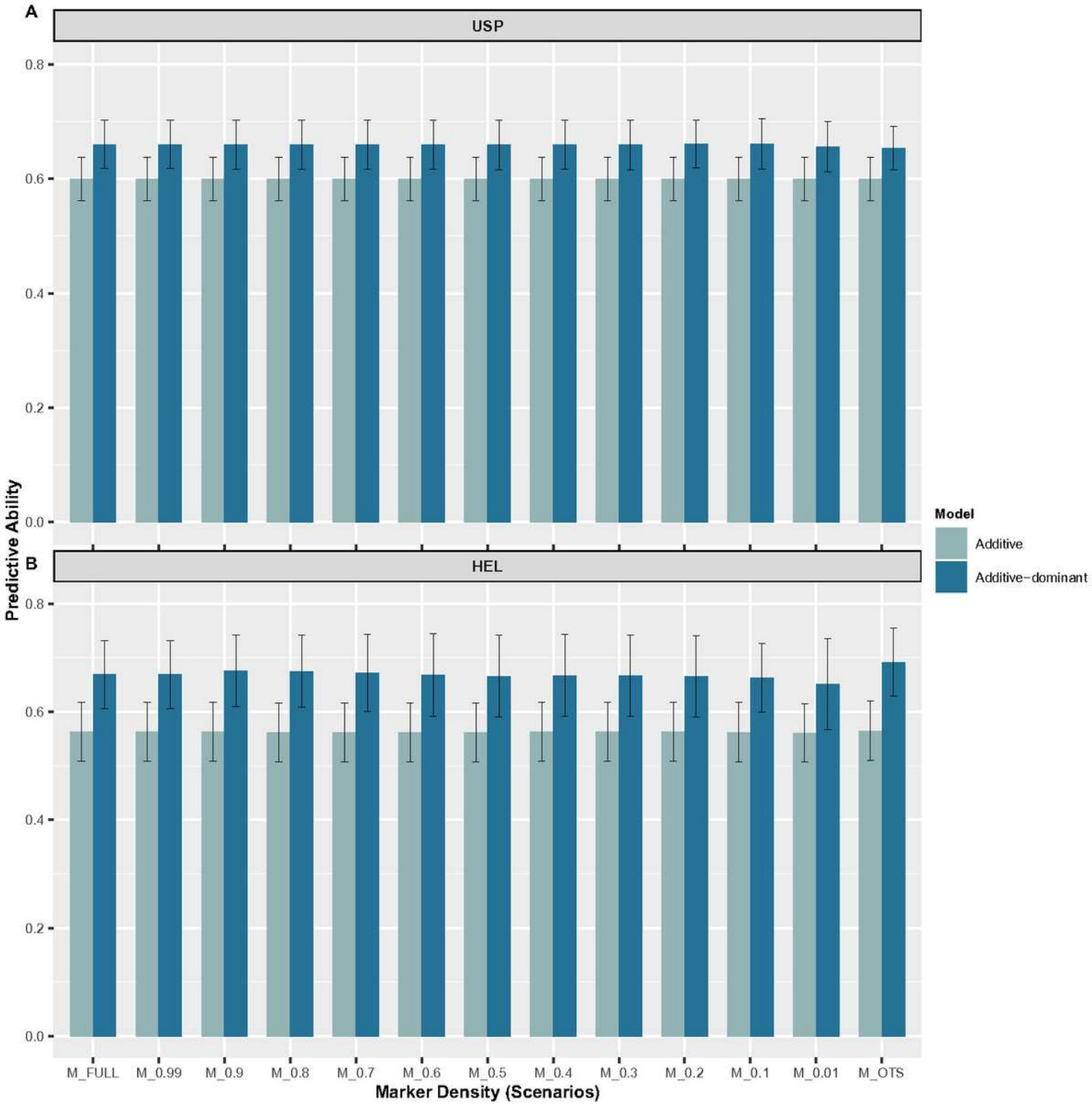

Figure 2

Predictive ability across marker density reduction scenarios based on different linkage disequilibrium levels and 363 an OTS algorithm, using additive and additive-dominant models. A USP dataset B HEL dataset.

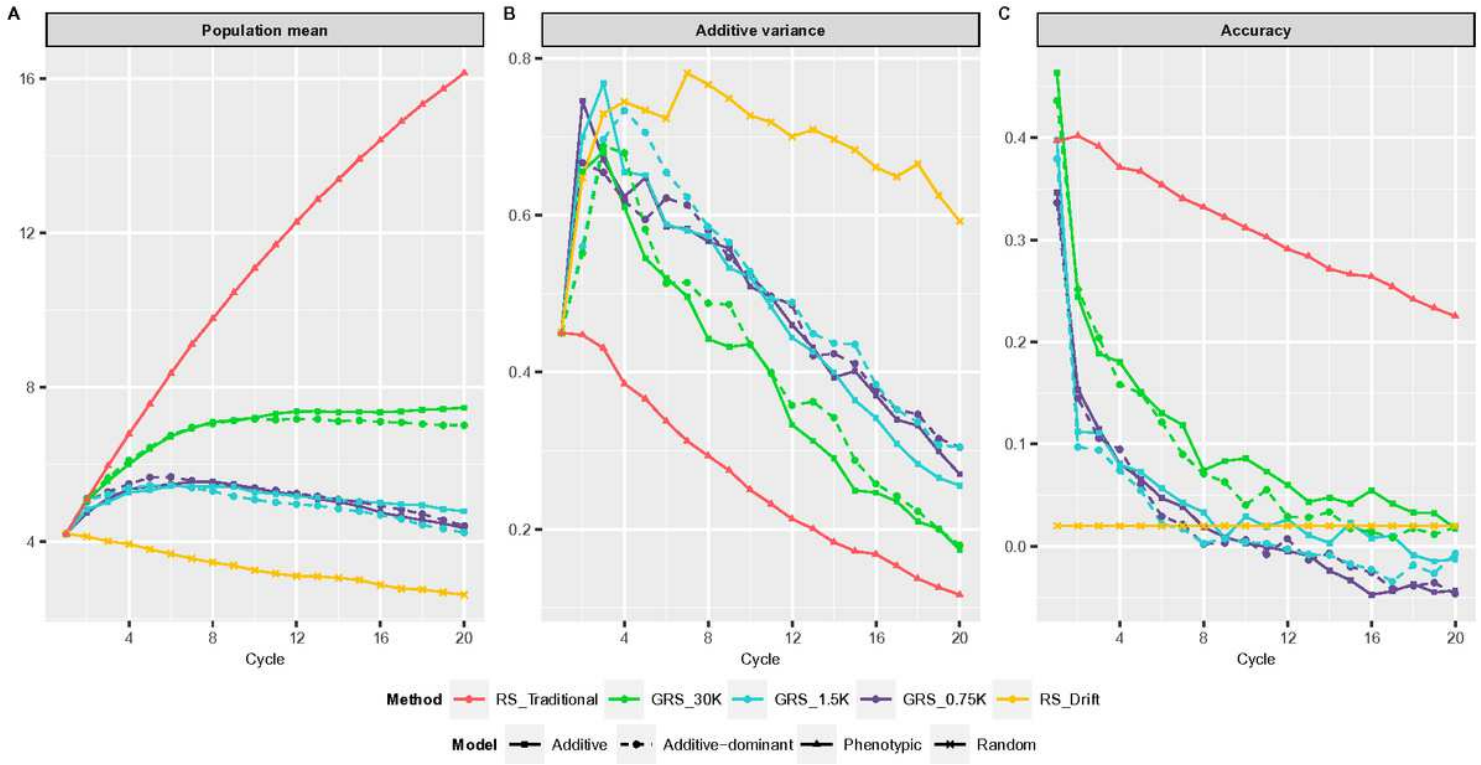

Figure 3

Simulation of 20 breeding cycles via PRS (Traditional and Drift) and GRS (with three marker densities) using the IRS breeding scheme with additive and additive-dominant G-BLUP models. A Population mean B Additive variance C Accuracy

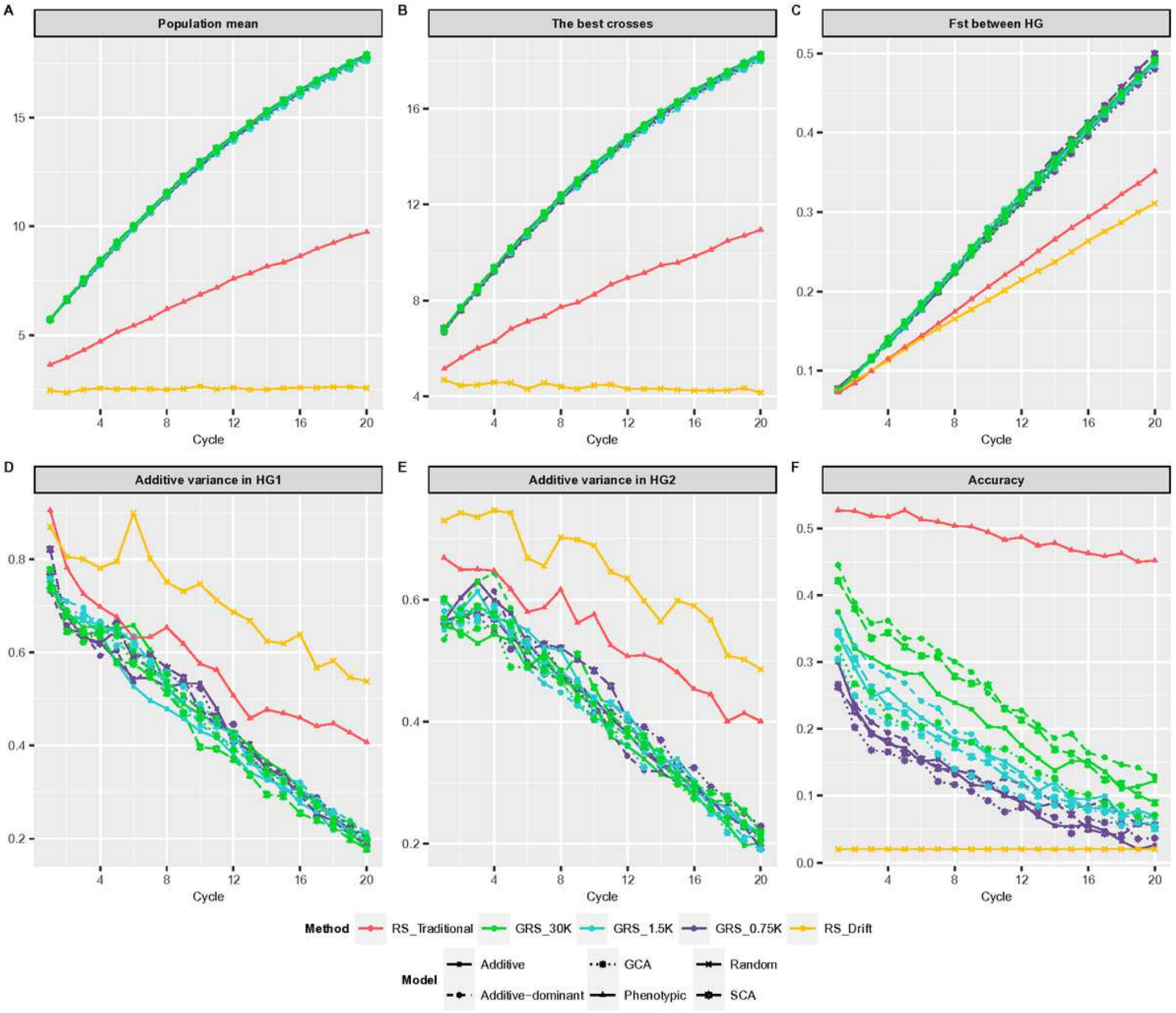

Figure 4

Simulation of 20 breeding cycles via PRS (Traditional and Drift) and GRS (with three marker densities) using the RRS breeding scheme with additive, additive-dominant, CGA, and SCA G-BLUP models. A Population mean B The best crosses C Fst between HGs D Additive variance in HG1 E Additive variance in HG2 F Accuracy

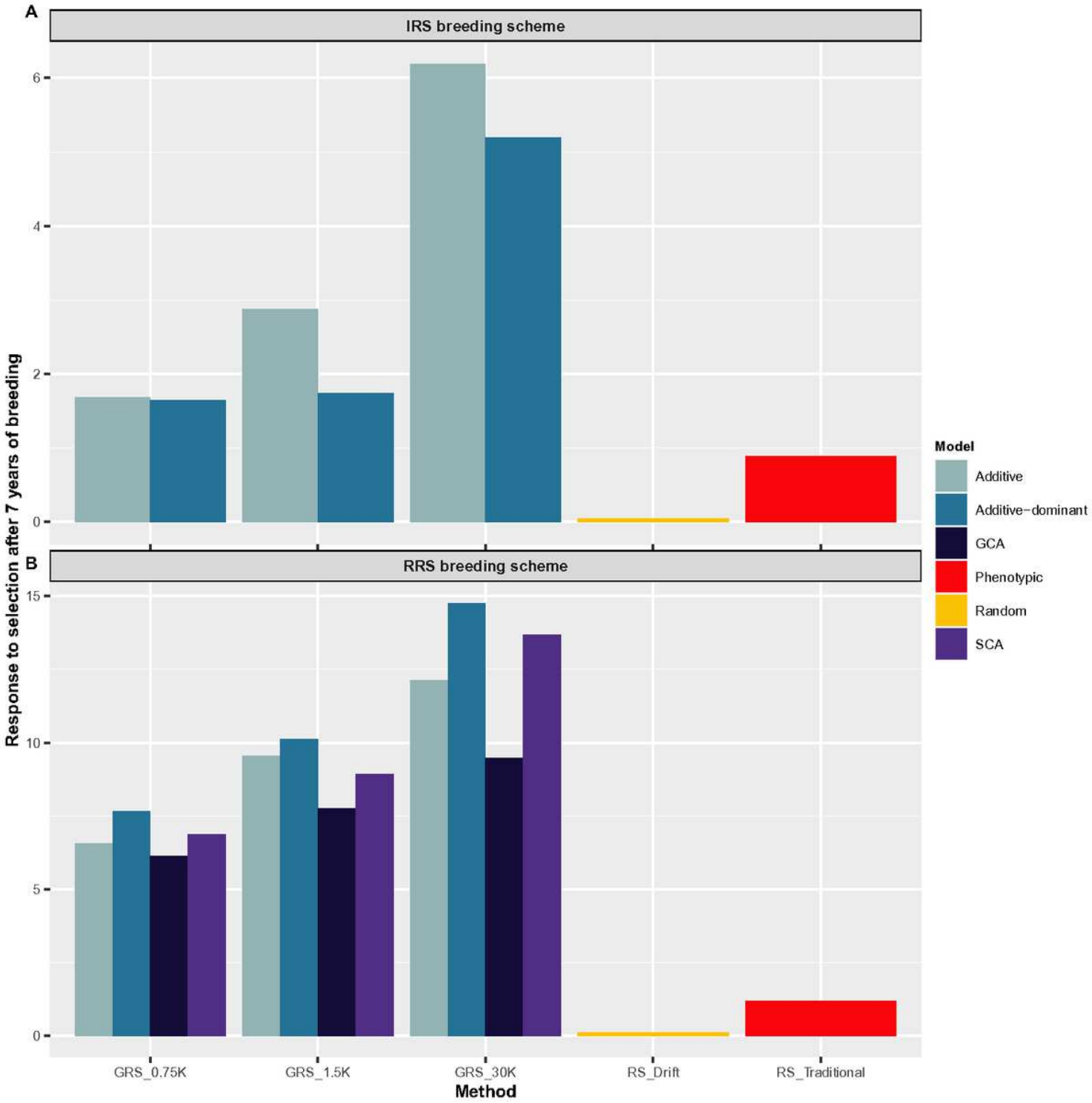

Figure 5

Responses to selection after seven years of breeding via PRS (Traditional and Drift) and GRS (with three marker densities) using different G-BLUP models. An IRS B RRS breeding schemes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Table2.xlsx
- SUPPLEMENTARYMATERIAL.pdf