1 **Reduction of genotyping marker density for genomic selection is not an affordable**
2 **approach to long-term breeding in cross-pollinated crops**

3
4 Júlio César DoVale^{a*}, Humberto Fanelli Carvalho^b, Felipe Sabadin^b, Roberto Fritsche-
5 Neto^b

6
7 ^a Department of Crop Science, Federal University of Ceará, Fortaleza, Ceará, Brazil

8 ^b Department of Genetics, “Luiz de Queiroz” College of Agriculture, University of São Paulo,
9 Piracicaba, São Paulo, Brazil

10

11 *Corresponding author

12 Department of Crop Science, Federal University of Ceará, Fortaleza, Ceará, Brazil

13 Júlio César DoVale

14 email: juliodovale@ufc.br

15

16 **Running title:** Reduction of marker density is not affordable to long-term breeding

17

18 **Word count for the main text:** 6,471

19

20 ORCID:

21 Julio César DoVale: 0000-0002-3497-9793

22 Humberto Fanelli Carvalho: 0000-0003-0745-7583

23 Felipe Sabadin: 0000-0003-2937-1465

24 Roberto Fritsche-Neto: 0000-0003-4310-0047

25 ABSTRACT

26

27 The selection of informative markers has been studied massively as an alternative to reduce
 28 genotyping costs for the genomic selection (GS) application. Low-density marker panels are
 29 attractive for GS because they decrease computational time-consuming and multicollinearity
 30 beyond more individuals can be genotyped with the same cost. Nevertheless, these inferences
 31 are usually made empirically using "static" training sets and populations, which are adequate
 32 only to predict a breeding program's initial cycles but might not for long-term cycles.
 33 Moreover, to the best of our knowledge, none of these inferences considered the inclusion of
 34 dominance into the GS models, which is particularly important to predict cross-pollinated
 35 crops. Therefore, that reveals an important and unexplored topic for allogamous long-term
 36 breeding. To achieve this goal, we employed two approaches: the former used empirical
 37 maize datasets, and the latter simulations of long-term breeding cycles of phenotypic and
 38 genomic recurrent selection (intrapopulation and reciprocal). Then, we observed the reducing
 39 marker density effect on populations (mean, the best genotypes performance, accuracy,
 40 additive variance) over cycles and models (additive, additive-dominance, specific combining
 41 ability (SCA)). Our results indicate that the markers reduction based on different linkage
 42 disequilibrium (LD) levels is viable only within a cycle and brings a significant decrease in
 43 predictive ability over generations. Furthermore, in the long-term, regardless of the selection
 44 scheme adopted, the more makers, the better because they buffer LD losses caused by
 45 recombination over breeding cycles. Finally, regarding the accuracy, the additive-dominant
 46 models tend to outperform the additive ones and perform similar to the SCA.

47

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

50 INTRODUCTION

51

52 Genomic selection (GS), using high-density single-nucleotide polymorphisms
53 (SNPs) chips, has been widely adopted to improve prediction ability and response to selection
54 (Hayes *et al.*, 2009; Crossa *et al.*, 2010). However, these chips' genotyping cost is currently
55 prohibitive, especially for low profitability species (Hou *et al.*, 2020), and programs in poor
56 regions, such as those in Latin America, Africa, and Southwest Asia.

57 In this context, density reduction of markers has been evaluated massively as an
58 alternative to reduce genotyping costs for GS applications. This purpose can be achieved
59 based on different criteria, for instance, marker effects of effect (Sousa *et al.*, 2019), positions
60 (Zhang *et al.*, 2015), preselection based on haplotype block analysis (Ma *et al.*, 2016);
61 genome-wide association studies (GWAS) approach in which significant markers are
62 selected to fit a prediction model (Subedi *et al.*, 2013), and by linkage disequilibrium (LD),
63 where highly correlated markers are eliminated by redundancy. All strategies have
64 advantages and disadvantages. Among them, the LD criteria seem to be more suitable
65 because it reduces multicollinearity, which is an obstacle for highly dense SNP panel, not
66 delivers noise in the marker effects (Xu, 2013), and the marker reduction process is
67 independent of the phenotype, reducing the probability of overfitting.

68 Low-density marker panels are also attractive for breeding purposes because they
69 decrease computational time-consuming, and then more individuals can be genotyped with
70 the same cost (Gorjanc *et al.*, 2015). With larger datasets of phenotyped and genotyped
71 individuals, GS's predictive ability may be increasingly driven by LD rather than by linkage
72 information (Hickey *et al.*, 2014). Furthermore, several studies claim that it is possible to
73 substantially reduce the number of markers and maintain high predictive ability (Tayeh *et al.*,
74 2015; Ma *et al.*, 2016; Sousa *et al.*, 2019; Al-Tobasei *et al.*, 2020). Thus, the use of
75 genotyping techniques with highly discriminative markers could be a viable alternative to

76 current low-cost SNP array strategies, with the potential to increase the fraction of the
77 genome captured in a cost-efficient manner (Elshire *et al.*, 2011).

78 Nevertheless, in all these studies, it was possible to obtain an informative marker
79 subset either with empirical or simulated data. The results are usually good because they
80 consider the LD within a generation with a fixed scenario, i.e., using fixed TS, in the early
81 stages of a breeding program, to infer across multiple breeding cycles. Although, during
82 breeding cycles (with selection and recombination), LD between markers and QTLs may
83 change due to crossing-over occurrences, especially for more complex traits (Xu, 2013). As
84 a result, there are changes in allelic frequencies and, consequently, allele substitution effects,
85 which may even cause allele fixation (Walsh and Lynch, 2018).

86 Although previous studies like the ones mentioned above evaluated the effect of
87 markers' density on genomic prediction models, there are no reports about their influence on
88 a recurrent selection (RS) scheme using models that incorporate additive and dominance
89 effects on cross-pollinated crops. The vast majority of published papers so far have
90 considered only additive models for genomic predictions, which may not represent the
91 genetic complexity of these species. Even though only additive effects are transmitted over
92 generations, the *per se* performance of the population or their single-crosses depends on the
93 dominance deviations (Falconer and Mackay, 1996), as this is the basis of heterosis and
94 crucial in the expression of the agronomic traits of these species (Bernardo, 2010).
95 Additionally, Galli *et al.* (2020) observed negative covariance between single-crosses genetic
96 values predicted from marker effects estimated from parental lines and single-crosses. Hence,
97 this reinforces the inability to predict single-crosses performance based on parental
98 information with additive effects.

99 Simulation is a suitable tool to verify the effect of marker density reduction in a
100 long-term breeding program using GS. We can control several factors to infer what happens
101 in given genetic parameters over breeding cycles in a fast, cheap and consistent way (Dai *et al.*, 2020). It is essential for breeding programs of cross-pollinated crops that usually adopt
102

103 RS-based methods for obtaining cultivars, either by intrapopulation recurrent selection (IRS)
 104 or by reciprocal (RRS), because genetic progress is achieved very slowly with enough
 105 selection and recombination over cycles. Additionally, the flexibility of the models and
 106 methods used in GS allows integrating allelic interaction effects, as the combining abilities
 107 (general and specific) of parents inbred lines (Werner *et al.*, 2018).

108 Reif *et al.* (2013) observed an increase in the single-crosses performance's predictive
 109 capacity incorporating the model of the effects of the general combining ability (GCA).
 110 Despite this, no studies compare the superiority of models that incorporate specific
 111 combining ability (SCA) effects concerning genomic best linear unbiased predictor (G-
 112 BLUP) model and include additive and dominance effects, mainly if we consider long-term
 113 breeding cycles. Thus, our objectives were: (i) to assess the impact of reduction marker
 114 density on the prediction accuracy in long-term breeding programs, (ii) to verify how the
 115 inclusion of dominance effects has implications on the main genetic parameters, and (iii) to
 116 compare genomic prediction models using combining ability, additive and additive-
 117 dominance effects over many cycles.

118 MATERIAL AND METHODS

119

120 Empirical datasets

121

122 *Phenotypic data*

123

124 This study considered two maize single-crosses datasets, obtained from Helix Seeds
 125 Company (HEL) and the University of São Paulo (USP), phenotyped for the grain yield (GY,
 126 in Mg ha⁻¹). Plots were manually harvested, and GY was corrected to 13% moisture. Each
 127 dataset was analyzed independently. The HEL dataset consists of 452 maize single-crosses
 128 obtained by crossing 111 inbred lines. The hybrids were evaluated in a randomized complete
 129 block design in five Brazilian sites during the 2015 season. The USP dataset consists of 903
 130 maize single-crosses obtained from a diallel mating design between 49 inbred lines. The
 131 single-crosses were evaluated in 2016 and 2017 at two sites (Piracicaba and Anhumas) in
 132 São Paulo State, Brazil. Field trials were performed in an augmented block design, each block
 133 consisting of 16 unique single-crosses and two commercial hybrids as checks. Single-crosses
 134 were evaluated at each site and year under two nitrogen fertilization levels (ideal and low N),
 135 composing eight environments. More details about the experimental design, cultivation
 136 practices for HEL and USP datasets can be accessed in Sousa *et al.* (2017) and Galli *et al.*
 137 (2020), respectively.

138

139 *Genetic-statistical models to estimate BLUEs*

140

141 Mixed model equations were used for the statistical analysis of the single-crosses at
 142 each data set. The joint analysis of each phenotype was performed to estimate genotype
 143 means across environments. Then, we fitted the following mixed model to obtain the best
 144 linear unbiased estimator (BLUE) for each genotype and then estimate the adjusted means

across environments via the *breedR* package (Muñoz and Rodriguez, 2016) in R software (R Development Core Team, 2019). The analysis for the USP dataset were performed by this model:

$$\mathbf{y} = \mathbf{Q}\mathbf{l} + \mathbf{S}\mathbf{b} + \mathbf{T}\mathbf{c} + \mathbf{U}\mathbf{g} + \mathbf{V}\mathbf{i} + \boldsymbol{\varepsilon}$$

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

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

$$\mathbf{y} = \mathbf{X}\mathbf{r} + \mathbf{T}\mathbf{g} + \mathbf{B}\mathbf{s} + \mathbf{H}\mathbf{x} + \boldsymbol{\varepsilon}$$

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

Subsequently, for both models, the single-crosses effect was considered random to obtain the variance components, performed by restricted maximum likelihood (REML/BLUP) procedure (Henderson, 1975).

Genotypic data

170 All parental inbred lines (USP and HEL datasets) were genotyped using the
171 Affymetrix® Axiom Maize Genotyping array containing 616 K SNPs (Unterseer *et al.*, 2014).
172 As for standard quality controls (QC), markers with low call rate (< 90%) and non-biallelic
173 were removed, and the remaining missing data was imputed by Beagle 5.0 algorithm
174 (Browning *et al.* 2018). Heterozygous *loci* on at least one individual were removed, and
175 single-crosses genotypes were obtained *in silico* by genomic information from parental
176 inbred lines. Finally, markers with minor allele frequency (MAF) < 0.05 were removed from
177 the hybrid genomic matrix resulting in 63,104 and 30,467 SNP markers for USP and HEL
178 datasets.

179

180 *Markers reduction on datasets*

181

182 Different SNP marker sets were obtained via two approaches: (i) removing SNP
183 markers based on the LD between them, and (ii) using the genetic algorithm LA-GA-T
184 implemented in the *STPGA* R package that aims to optimize training sets (OTS) developed
185 by Akdemir (2017).

186 In the first, pairwise LD was calculated as squared allele frequencies correlation
187 (r^2). Then, in order to remove redundant markers, we applied the following threshold values,
188 0.99, 0.90, 0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10, and 0.01 using the *SNPRelate* package
189 (Zheng *et al.*, 2012). As the benchmark scenario, we used the full matrix of SNP markers
190 (M_Full), after quality control, for downstream comparisons. In the second approach, we
191 applied a singular value decomposition (SVD) using the number of eigenvalues that
192 explained 98% of the variance, according to Pocrnici *et al.* (2016). To reduce computational
193 time and represent only the LD among markers' physical components, we process SVD +
194 OTS within each chromosome. Lastly, we had thirteen SNP matrices to perform genomic
195 prediction models.

196

197 Genomic prediction

198

199 We used the thirteen SNP matrices to build genomic relationship matrices (GRM).

200 We applied additive and additive-dominance G-BLUP models, respectively, to perform the
201 genomic prediction, following the equations:

$$202 \hat{\mathbf{g}} = \mathbf{1}\mu + \mathbf{Z}_a\mathbf{a} + \boldsymbol{\varepsilon}$$

$$203 \hat{\mathbf{g}} = \mathbf{1}\mu + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_d\mathbf{d} + \boldsymbol{\varepsilon}$$

204 where $\hat{\mathbf{g}}$ is the vector of adjusted means of single-crosses from the phenotypic analysis; μ is

205 the mean (intercept); $\mathbf{a}$ is the vector of additive genetic effects of the individuals, where $\mathbf{a} \sim$

206 $N(\mathbf{0}, \mathbf{G}_a\sigma_a^2)$; $\mathbf{d}$ is the vector of dominance effects, where $\mathbf{d} \sim N(\mathbf{0}, \mathbf{G}_d\sigma_d^2)$; and $\boldsymbol{\varepsilon}$ is the vector

207 of random residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$. $\mathbf{1}$ is the vector of ones; $\mathbf{Z}_a$ and $\mathbf{Z}_d$ are incidence matrices

208 for $\mathbf{a}$ and $\mathbf{d}$. σ_a^2 is the genomic additive variance, σ_d^2 is the genomic dominance variance, and

209 σ_ε^2 is the residual variance. $\mathbf{G}_a$ and $\mathbf{G}_d$ are the additive and dominance genomic relationship

210 matrices, following the equations: $\mathbf{G}_a = \mathbf{W}_A\mathbf{W}_A'/2\sum_{i=1}^n p_i(1-p_i)$ and $\mathbf{G}_d =$

211 $\mathbf{W}_D\mathbf{W}_D'/4\sum_{i=1}^n (p_i(1-p_i))^2$, where p_i is the frequency of one allele of the locus i and $\mathbf{W}$ is

212 the incidence matrix of markers (VanRaden, 2008). The $\mathbf{W}_A$ matrix was coded as 0 for

213 homozygote A_1A_1 , 1 for the heterozygote A_1A_2 , and 2 for the homozygote A_2A_2 . For $\mathbf{W}_D$,

214 genotypes were coded as 0 for both homozygotes and 1 for heterozygote.

215 The *snprReady* package (Granato *et al.*, 2018) was used to build GRM and obtain

216 estimates of effective population size (N_e) based on $\mathbf{G}_a$ matrix. $\mathbf{G}_a$ and $\mathbf{G}_d$ matrices obtained

217 from SNP datasets were used to calculate the correlation between them via Mantel correlation

218 using the *vegan* R package (Oksanen *et al.*, 2019). The *superheat* R package built the

219 heatmaps (Barter and Yu, 2018) using M_Full, M_0.5, M_0.01, and M_OTS. For

220 comparison, we used $\mathbf{G}_a$ and $\mathbf{G}_d$ from M_Full dataset to cluster single-crosses using the

221 Euclidean distance. Afterward, the singles-crosses order was used to sort heatmaps M_0.5,

222 M_0.01, and M_OTS. The genomic prediction models were performed using the *sommer*
223 package (Covarrubias-Pazarán, 2016).

224 Predictive ability was estimated as the correlation between the predicted and
225 observed genotypic values via cross-validation alpha-based design (CV- α , Yassue *et al.*,
226 2020), which is an extension of the methodology presented by Shao (1993) and consists of
227 assigning treatments to folds in each replication using alpha-lattice sorting premises. The
228 CV- α was intended to create scenarios with two, three, or four replicates, regardless of the
229 number of treatments. Each replicate is split into folds, and the number of folds is determined
230 by the percentage of training and validation sets. Each fold across replicates is based on the
231 $\alpha(0,1)$ lattice design aiming to reduce the concurrences of any two treatments in the same
232 fold (block) across the replicates (Patterson and Williams, 1976). We used here four
233 replicates with five-folds, and predictive abilities were averaged.

234

235 **Simulated experiments**

236

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

238

239 We simulated a population of maize single-crosses from parental inbred lines to
240 perform phenotypic and genomic predictions for a quantitative trait. For this, we used the
241 *AlphaSimR* package (Gaynor *et al.*, 2020), and the scripts can be found in the supplementary
242 files. A historical population founded of 1,000 individuals was simulated stochastically with
243 ten chromosomes containing 31,000 segregating loci. The individuals were diploidized but
244 not inbred.

245 Trait genetic architecture was simulated considering: (i) only additive effects and
246 (ii) additive + dominance effects. In both, we used randomly sampling 1,000 QTLs (100 per
247 chromosome) with effects drawn from a gamma distribution with a scale parameter of 1.0.
248 We considered an average degree of dominance of 0.5 in each QTL for a trait with additive-

dominance effects, according to the ratio between additive effects and dominance in Hallauer *et al.* (2010). We consider heritabilities in the broad and narrow sense equal to 0.5 and 0.33, respectively. From this, it was possible to estimate the error variance component. The phenotypic values were obtained by sum effects of the QTLs plus the random error from the segregating loci with mean 0 and variance 1.

In order to represent the genotyping chips, we used three marker densities sampled from segregating loci: 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 means one marker to each 2 cM.

Breeding schemes based on recurrent selection

We used two main breeding schemes to evaluate the effect of markers reduction on predictions over 15 cycles: (i) intrapopulation recurrent selection (IRS) and (ii) reciprocal recurrent selection (RRS). To compose the base-population IRS, we selected 15 inbred lines (Fig. 1). Subsequently, a full diallel was composed between them to generate the initial cycle (C_0). For RRS, the base population was built similarly, but in this case, parents of the best single-crosses were selected to build the two heterotic groups (HG), being ten unique parents in each. All progenies were composed of 50 individuals (equivalent to one trial with 2 plots of 25 plants), regardless of the breeding scheme, stage, or cycle.

The historical population and trait parameters were equal in both schemes. For the burn-in stage, we used two traditional recurrent selection schemes:

IRS – 200 half-sib progenies were obtained; then, the top 30 progenies were selected, and there was the random recombination of 1,500 self-fertilized individuals (S_1), corresponding the 30 best parents;

RRS – 200 full-sib progenies were obtained - via North Carolina mating design II (NCII). Then, the best hybrids were selected, and we proceed with the recombination of the 10 best

276 parents for each HG. At the end of each cycle, the fixation index (Fst) was estimated to
277 observe the genetic distance between HG.

278 After the three cycles of the burn-in phase, 3,000 individuals were randomly selected
279 in each scheme (Fig. 1). Among them, 2,000 were sampled to compose the training set (TS)
280 for genomic predictions, where the marker effects were estimated using the three different
281 SNP-chip densities. For that, genotyping and phenotyping of these individuals were
282 simulated. These 3,000 individuals were also used as the base population (C_0) for the next 15
283 recurrent selection cycles (genomic and phenotypic). In each cycle, for both phenotypic
284 breeding schemes, obtaining progenies, selection, and recombination were made based on
285 200 progenies, similarly to the burn-in phase. Conversely, for the genomic schemes, the
286 genotype of the 200 parents allows us to skip the progeny phase, select directly based on the
287 genomic breeding values, and plant only the select individuals to recombine, closing one
288 breeding cycle in just one growing season. Furthermore, based on the parents' genotype, we
289 can obtain *in silico* all 40,000 possible single-crosses genotypes (200 x 200), increasing the
290 population size and the intensity of selection.

291

292 *Genomics predictions and comparison models*

293

294 According to the equations mentioned above, we used additive and additive-
295 dominance G-BLUP models to perform the genomic prediction on three SNP-chip densities
296 in both breeding schemes. Additionally, we model for the RRS scheme with combining
297 abilities effects. In general, the genomic estimated value (GEV) of single-crosses was
298 obtained using the estimates of GCA and SCA according to the following expression:

$$299 \quad \mathbf{GEV}_{ij} = \mathbf{GCA}_i + \mathbf{GCA}_j + \mathbf{SCA}_{i \times j}$$

300 where $\mathbf{GEV}_{ij}$ is the genomic estimated value of single-cross ij , $\mathbf{GCA}_i$ is the general combining
301 ability of parent i , $\mathbf{GCA}_j$ is the general combining ability of parent j , and $\mathbf{SCA}_{i \times j}$ is the
302 specific combining ability from the cross between i and j parentals. Kronecker products were

303 used to model the interaction between single-crosses' parental genomes to capture SCA
304 effects (Acosta-Pech *et al.*, 2017; Basnet *et al.*, 2019). We repeated the simulations 10 times
305 and averaged the estimates above. Accuracy for phenotypic selection was estimated by the
306 square root of heritability and for GS by the correlation between genetic values and estimated
307 breeding values.

308 RESULTS

309

310 *Markers reduction on empirical datasets*

311

312 The number of markers decreased substantially with reducing LD levels on
313 empirical datasets (Table 1). In general, the reduction was smoother in both datasets up to
314 the matrix with 10% LD between markers (M_0.1). Hence, from the matrix with all markers
315 (M_Full) to the one with only 1% LD between markers (M_0.01), we observed a reduction
316 of approximately 11.5 and 3.3 times for the USP and HEL datasets. However, from M_0.01
317 to matrix markers obtained with the algorithm for the optimization training set (M_OTS),
318 drastic reductions were observed for both datasets, ranging from 3.9 to 5.4 times for USP and
319 HEL.

320 Although considerable markers reduction over LD levels, the effective population
321 size (N_e) has not been changed for both single-crosses datasets when considering the
322 different matrices (scenarios) for estimating this parameter (Table 2). In addition, the
323 significant and high correlation coefficients between additive (G_a) (> 0.98 for USP and HEL
324 datasets) and dominance (G_d) matrices (> 0.73 for USP and > 0.99 for HEL dataset) obtained
325 from different markers densities (Tables S1 and S2), reiterate the matrices with low-density
326 markers are as representative as those with high-density. Besides that, heatmaps obtained
327 with additive and dominance matrices among the most extreme scenarios prove that the
328 genetic relationship has not been altered with the marker density reduction (Figs. S1 and S2).

329 Similar to the previous results for the USP and HEL datasets, we noted that the
330 predictive ability has remained mostly unchanged (minor changes) throughout the various
331 markers reduction scenarios, including the most extremes ones (Fig. 2). Therefore, with
332 approximately 1,500 markers (1,408 for USP and 1734 for HEL datasets), it is possible to
333 predict maize single-crosses' performance with the same reliability as all marks after the
334 quality control process. Additionally, greater predictive abilities were observed when we

335 adopted the additive-dominance model regardless of marker density. On average, this model
336 surpassed the compound only by additive effects by 9.8% and 19% for the USP and HEL
337 datasets, respectively.

338

339 *Marker density reduction effect on long-term IRS breeding*

340

341 The simulated data showed that GS overcame phenotypic selection for the
342 population mean during all breeding cycles in an IRS scheme, regardless of marker density
343 (Fig. 3A). GS's superiority began to become evident in the third cycle. At the end of the 15
344 cycles, the GS allowed an average of 14.5 Mg ha⁻¹ to be reached, while the phenotypic
345 selection a maximum of 88% of this. We also observed that for this parameter, the additive
346 model tended to be superior to the additive-dominance, regardless of the selection method
347 (phenotypic or genomic). However, in the phenotypic selection, the difference between
348 models tended to be more significant over the cycles.

349 The additive variance available for GS increased until the second breeding cycle and
350 was consumed similarly to the phenotypic selection (Fig. 3B). Selection methods and genetic
351 models did not show a significant difference of additive variance tendencies over breeding
352 cycles. On the other hand, GS using the additive-dominance model generated higher accuracy
353 estimates (Fig. 3C). In the initial cycle, GS with additive-dominance model reached almost
354 twice the accuracy (> 0.7) of GS with the additive model (maximum 0.43) and also from
355 phenotypic selection regardless of the model (at most 0.40). In general, after the second cycle,
356 GS showed abrupt reductions in reliability. As for the phenotypic selection, there was a loss
357 in accuracy over the cycles, but with a smaller variation. However, higher accuracy was
358 obtained when the breeding program was conducted with GS, and more genomic information
359 was available, regardless of the genetic model. It is worth noting that after the reduction in
360 accuracy of the second cycle, GS using the SNP-chip of 30K promoted less variation and a

superior accuracy over other scenarios. Furthermore, the additive-dominance model outperformed the additive one in terms of GS accuracy for IRS long-term breeding.

Marker density reduction effect on long-term RRS breeding

When the RRS scheme conducts the breeding program, similar outcomes were observed for the population mean (Fig. 4A). Also, the differences between this scheme's methods were more significant, with performance above 15 Mg ha⁻¹ for GS and around 9 Mg ha⁻¹ for phenotypic selection at the end of the 15 cycles. The best crosses obtained via GS also outperform those obtained via phenotypic selection, regardless of SNP-chip densities (Fig. 4B). In general, single-crosses with the adoption of GS exceeding 17 Mg ha⁻¹, on average 70% higher than those obtained by phenotypic selection. GS allowed maintaining greater genetic divergence by fixation index (Fst) between heterotic groups (Fig. 4C), especially from the eighth cycle.

There was a similar behavior for the population mean and best crosses regardless of the models' GS method. However, for the Fst parameter, the additive model showed a slightly higher tendency than the additive-dominance model. The phenotypic selection follows the fashion showing the additive model's superiority for the population mean and best crosses parameters. On the other hand, the GCA model tended to be inferior in all of them. For population mean, the best-crosses, and Fst, the additive-dominance and SCA models presented practically the same performance. This same behavior was observed when we considered the phenotypic selection for the best crosses and population mean parameters, as seen in the IRS scheme.

Unlike what was observed in the IRS scheme, the additive variance (for both heterotic groups) tended to be consumed more quickly when GS conducted the breeding program, regardless of the marker density and model (Fig. 4D and 4E). The phenotypic selection also showed greater accuracy from the beginning to the end of breeding cycles.

388 (Fig. 4F). Overall, the accuracy tended to be superior in the breeding process with a higher
389 density of markers. The modeling did not interfere decisively in the performance of this
390 parameter. However, there was a slight tendency to superiority the additive-dominant model
391 for accuracy, similar to the IRS scheme.

392 DISCUSSION

393

394 The high-density genotyping chips permit whole coverage of the genome with
395 markers, making it possible to capture the total genetic variability of the trait across the
396 genome and predict the breeding value without collecting phenotypic data of the individuals
397 (Meuwissen *et al.*, 2001).

398 Low-density marker chips are desirable to implement GS in breeding programs to
399 reduce genotyping costs, especially for species with low profitability. Several studies have
400 documented the benefits of using marker subsets in GS analyses (Gorjanc *et al.*, 2015; Tayeh
401 *et al.*, 2015; Ma *et al.*, 2016; Li *et al.*, 2018; Sousa *et al.*, 2019). Thus, the fact that few
402 markers are sufficient to generate predictive ability similar to those obtained with high
403 markers density, without changing the genetic relationship (Tables 1 and 2; Figs. S1 and S2)
404 brings excellent advantages for the low-density chips, leading to technology improvements
405 and cost (Heaton *et al.*, 2005; Chessa *et al.*, 2007), as well as computational demand.

406 Nevertheless, long-term recurrent selection difficulties GS effectiveness. Over
407 cycles, recombination between markers and QTL will cause LD decreases, while selection
408 and drift will potentially act to generate new LD or tighten the LD between closely linked
409 *loci* (Hill and Robertson, 1968; Lorenz *et al.*, 2011). Thus, there is a demand to study the
410 effects of reducing the markers density over breeding cycles. Therefore, simulations of
411 different scenarios and controlled variants may do this quickly and less costly.

412

413 *Consequences of marker reduction for GS to long-term breeding schemes*

414

415 *Breeding program conducted by IRS scheme*

416

417 When the breeding program is conducted by recurrent selection with a single
418 population (IRS), generally, the purpose is to release open-pollinated varieties. Thus, the

419 population means is an essential parameter to evaluate the performance of the IRS scheme.
 420 Our results showed that GS applied in an IRS scheme overcomes phenotypic selection,
 421 regardless of SNP-chip density (Fig. 3A). This was expected because GS is more informative
 422 than phenotypic selection. With the availability of genomic and phenotypic data, it becomes
 423 easier to identify superior genotypes with precision. Despite this, the additive variance was
 424 consumed by both strategies (phenotypic and genomic selection) similarly (Fig. 3B).

425 The GS accuracy decreased abruptly from the second breeding cycle (Fig. 3C). That
 426 suggests that TS must be re-calibrated (updated) during recurrent selection to maintain
 427 prediction ability (Neyhart *et al.*, 2017). Indeed, studies simulating recurrent selection in
 428 eucalyptus (Jannink, 2010), barley (Denis and Bouvet, 2013), and with empirical data from
 429 advanced-cycle rye (Auinger *et al.*, 2016) revealed that the accuracy of GS was improved by
 430 recalibration of the TS. Therefore, each recurring selection cycle's recombination events
 431 should shift in the QTL-marker LD pattern, making its capture unfeasible. Further, the change
 432 in allele frequencies due to the selection over the cycle's impact on estimating the effects of
 433 allelic substitution and, consequently, on the genomic prediction.

434 Furthermore, this situation can be aggravated when the markers density is initially
 435 very low, and then it will compromise the recalibration of the TS. Our results showed a lower
 436 loss of accuracy over the breeding cycle when higher marker densities were used (Fig. 3C).
 437 Müller *et al.* (2017) illustrated by simulations that more remarkable persistence of predictive
 438 accuracy and, consequently, selection gains when higher markers density were used. Thus,
 439 most markers seem to buffer better the effect of the loss LD between markers and QTLs over
 440 the recombination cycles, even without TS recalibration. Although there is an expectation
 441 with reducing genotyping costs with low-density chips, in practice, budgets in companies
 442 specialized in genotyping show that the price differences between chips of high and low
 443 markers density are negligible (personal communication). Therefore, our results showed that
 444 genotyping with low-density chips for a breeding program conducted to long-term breeding
 445 cycles is not compensatory.

446

447 *Breeding program conducted by RRS scheme*

448

449 In RRS, two populations are improved simultaneously, not to increase their
450 performance by themselves but the cross between them. In principle, the frequencies
451 desirable alleles in each population are maximized and improved of additive effects.
452 Subsequently, both populations are crossed to capitalize on heterozygous individuals' non-
453 additive effects (Hallauer *et al.* 2010). Thus, the RRS is more complex than the IRS.
454 However, when conducted with the GS support, it is possible to run similarly to the IRS
455 scheme. The population performance was better (Fig. 3A and 4A) probably because single-
456 crosses are generated (with more significant heterosis than the open-pollinated varieties).

457 The performance of single-crosse selected by GS was higher than phenotypic
458 selection (Figs. 4A and 4B). It may be explained due to GS permits to generate *in-silico* all
459 single-crosses combinations (in the present study 40,000), which is unable with traditional
460 phenotypic selection, which increases the chances of identifying the best crosses.
461 Furthermore, the genomic information allows discriminating alleles belonging to each
462 heterotic group and keeping them more divergent over the breeding cycles (Fig. 4C). As the
463 selection is made based on single-crosses, which have a high heterosis effect, the GS better
464 captures the complementarity, heterozygosity, and genetic distance components. That is
465 important for a breeding program that aims to release single-crosses since those are the main
466 components of heterosis (Moll *et al.*, 1965; Prasad and Singh, 1986).

467 Long-term recurrent selection may reduce genetic variance to the point where
468 genetic gains are limited (Jannink, 2010). Indeed, genetic variance showed a substantial
469 decline over breeding cycles in both selection schemes and methods. However, in the RRS
470 scheme, GS resulted in a higher genetic diversity loss than phenotypic selection (Figs. 4D
471 and 4E). Muleta *et al.* (2019) observed similar results with higher losses in the simulation of
472 an oligogenic trait with high heritability and lower for a polygenic trait with low heritability.

473 These authors attributed this fact to higher selection intensity in GS due to more selection
474 cycles per year (simulated two and three cycles of GS per year). In our study, we simulate
475 only one cycle per year for both GS and phenotypic selection. Even so, it was clear that the
476 additive variance is consumed more quickly by GS. Essentially, selection intensity
477 (standardized) will be higher in GS compared to phenotypic selection because it has a larger
478 sample, as previously mentioned.

479 Maintenance of prediction accuracy across selection cycles is critical for long-term
480 genetic gain (Müller *et al.*, 2017). Interestingly, the phenotypic selection also overcame the
481 GS for accuracy throughout cycles (Fig. 4F). Mathematically, the accuracy is a function of
482 the additive variance. Therefore, a reduction of that will promote a decrease in prediction
483 accuracy. This finding may be explained by the breakdown of LD between markers and QTL
484 due to recombination (Jannink, 2010). Moreover, over cycles to selection and drift of alleles
485 are not tagged by markers in TS, which cannot be targeted by GS (Rutkoski *et al.*, 2015),
486 culminating in low accuracy. Another suggested explanation for this result is that the
487 selection is based on genetic divergence, in which the alleles present in individuals captured
488 from both heterotic groups change in direction and sense (upper and lower tail), affecting the
489 reliability of the GS.

490 Similarly to the IRS scheme, we noted that the higher marker density results in a
491 higher accuracy over the breeding cycles. A larger marker quantity should better capture the
492 LD between QTLs and markers since some of them can disassociate from the QTLs due to
493 recombination during the cycles, whereas other markers can remain associated with the
494 QTLs, maintaining some level of LD.

495

496 *Comparison between prediction models*

497

498 The predictions by GS tended to be more accurate than phenotypic selection because
499 it considers the real genetic relatedness between genotypes rather than the average

500 expectation. Studies have been conducted with empirical data (Al-Tobasei *et al.*, 2020) and
501 with simulations (Hou *et al.*, 2020) whether for animal or plant breeding, to understand the
502 markers' effect on the accuracy of genomic predictions. However, they only consider the
503 additive genetic effects to identify marker subsets and, consequently, make the predictions.

504 Nevertheless, several other studies have already shown that information about the
505 dominance deviations incorporated via kernels are essential to increase the efficiency of the
506 prediction models GS (Azevedo *et al.*, 2015; Dias *et al.*, 2018; Matias *et al.*, 2018; Dai *et al.*,
507 2020), especially when the species is cross-pollination and heterosis is explored in the final
508 product (Santos *et al.*, 2016). Our reports for empirical datasets (Fig. 2) corroborate these
509 others that showed the importance of incorporating additive effects and deviations of
510 dominance in the model to increase predictive ability (Technow *et al.*, 2012; Alves *et al.*,
511 2019). In the presence of dominance, the additive-dominant model is expected to be more
512 accurate because the additive model falsely assumes that residuals are independent and
513 identically distributed (Duenk *et al.*, 2017). Furthermore, in addition, the additive-dominant
514 models allow separating the effects better than the models that consider only the additive
515 effects. When we consider only the additive kernel in the models, there is a "confusion"
516 among additive effects, dominance, and residuals (Alves *et al.*, 2019).

517 However, in the long-term breeding, we found that the additive-dominance model
518 was consistently superior to the IRS scheme's accuracy parameter. On the other hand, the
519 additive model tended to be slightly superior to the others for population means (for both IRS
520 and RRS), the best crosses, and Fst between heterotic groups. Duenk *et al.* (2017) showed
521 that the additive model contributed to the error with sampling deviations of genotype
522 frequencies from Hardy-Weinberg Equilibrium (HWE). Conversely, the additive-dominant
523 model is only sampling deviations of allele frequencies. Consequently, according to these
524 authors, in the presence of dominance, the root mean squared error of the average effect of
525 allelic substitution with the additive-dominant model was always smaller, especially when
526 the heritability is low. Therefore, for cross-pollinated species, in which single-cross are

527 commercially exploited, it is expected genetic models that include dominance more robust
528 and generated higher accuracies of the average effect of allelic substitution than purely
529 additive models when dominance is present. However, our results show that this benefit did
530 not have advantages in estimating the parameters studied.

531 It is expected that GS does not consistently adjust to RRS, since in the modeling
532 process, only a vector of additive effects and deviations dominance is considered for the
533 entire population and their respective heterotic groups. Thus, we use two more approaches in
534 the simulation of breeding cycles with RRS; G-BLUP modeled only with the GCA effects
535 and the other with SCA effects. Based on quantitative genetics theory, it is expected that via
536 traditional GS, individuals from both heterotic groups would be selected by the same allele
537 effects, which might lead to a reduction in the genetic distance among HG over the cycles,
538 vanishing the heterosis expression (Hallauer *et al.*, 2010).

539 In theory, it is hoped that breeders who exploit heterosis to generate products
540 (especially single-crosses) can benefit from models that can predict non-additive genetic
541 effects. Thus, it is possible to predict the performance of single-crosses with greater accuracy.
542 In essence, the two models that best capture dominance deviations are the additive-
543 dominance and SCA. These models presented similar estimates for the parameters studied,
544 such as, for example, population means, best crosses, and F_{st} . Interestingly, Werner *et al.*
545 (2018) used different models to verify the increase in predictive ability in oilseed rape with
546 the GS approach. This study found that incorporating SCA effects combined with GCA into
547 the RR-BLUP method (similar to G-BLUP) was the approach that caused the most
548 fluctuation over 200 validation cycles. Except for this, all methods (including those based on
549 Bayesian statistics) revealed comparatively robust predictive ability and high accuracy.
550 However, we did not notice any trends of considerable differences between results generated
551 by four models over the 15 cycles, although it is known that the dominance deviations
552 captured by the SCA effect are paramount for heterosis (Falconer and Mackay, 1996).

553 We believe that the similarity of results is because we select the best single-crosses
 554 first, then recombine their parents. This procedure is standard in the RRS scheme and does
 555 not change with both modeling approaches. Regarding the single-cross genomic prediction,
 556 it is well-known that the non-additive effects are partially included in the parents breeding
 557 value, GCA, or in allele substitution effects (Falconer and Mackay, 1996; Bernardo, 2010;
 558 Werner *et al.*, 2018). Even though it is noteworthy that even in some circumstances
 559 generating lower estimates than the additive, the additive-dominance model generates more
 560 reliable estimates (Alves *et al.*, 2019). Reif *et al.* (2007) demonstrated that the dominance
 561 variance decreases concerning additive variance but increases according to the populations'
 562 divergence. The dominance effects are increasingly absorbed by the population mean or
 563 become inseparable from additive effects over breeding cycles (Technow *et al.*, 2014).

564

565 *Final considerations*

566

567 It is challenging to disentangle the different factors that may contribute to
 568 heterogeneity and predictive ability failure in real data sets (Dai *et al.*, 2020). However,
 569 several genetic-statistical models and tools are available to minimize bias. Using simulations,
 570 we can control several uncontrolled factors in the field and focus on answer specific
 571 questions. Our results clearly showed that reducing the markers is punctually efficient (within
 572 a breeding cycle or a static group of parents). Hence, the low-density marker sets are not
 573 affordable for recurrent genomic selection (either IRS or RRS), usually conducted for the
 574 long-term. Overall, the more markers, the higher the genotyping chip's buffering capacity to
 575 support LD losses (between markers and QTL) by crossing-over during recombination
 576 cycles. Therefore, this study complements the results of others published recently. Although
 577 GS allows a higher genetic gain, especially when weighted by time, Muleta *et al.* (2019) also
 578 observed higher GS accuracy losses about phenotypic selection throughout the breeding
 579 cycles. In this sense, it is evident the need for more studies about how often and how to make

580 updates to the TS to maximize the reliability and prediction accuracy (Rincent *et al.*, 2012),
581 keeping the LD between markers and QTLs (Neyhart *et al.*, 2017), and the relationship
582 between TS and target populations satisfactory levels. Finally, we recommend that
583 genotyping should be done with a high marker density so that this in a long-term breeding
584 program maintains the ability to capture LD over cycles and does not become a bottleneck.
585 The additive model is sufficient to predict breeding populations' performance, but the
586 additive-dominance generates slightly better estimates of accuracy throughout the breeding
587 cycles. Furthermore, regarding non-additive genetic effects, there are no differences between
588 the additive-dominance and SCA G-BLUP models.

589 **Acknowledgments:** We thank Luiz de Queiroz College of Agriculture (University of São
590 Paulo), Federal University of Ceará and Conselho Nacional de Desenvolvimento Científico
591 e Tecnológico (CNPq) – Process 104371/2019-6, for the financial support.

592

593 **Conflicts of interest:** The authors declare no conflict of interest.

594

595 **Data archiving:** <https://data.mendeley.com/datasets/4nccgtcpgn/1>

596 REFERENCES

597

598 Acosta-Pech R, Crossa J, de los Campos G, Teyssèdre S, Claustres B, Pérez-Elizalde S, *et*
599 *al.* (2017). Genomic models with genotype $\times$ environment interaction for predicting
600 hybrid performance: an application in maize hybrids. *Theor Appl Genet* **130**: 1431–
601 1440.

602 Akdemir D (2017). STPGA: Selection of training populations with a genetic algorithm. :
603 111989.

604 Al-Tobasei R, Ali A, Garcia A, Lourenco D, Leeds T, Salem M (2020). Genomic
605 Predictions for Muscle Yield and Fillet Firmness in Rainbow Trout using Reduced-
606 Density SNP Panels.

607 Alves FC, Granato ÍSC, Galli G, Lyra DH, Fritsche-Neto R, De Los Campos G (2019).
608 Bayesian analysis and prediction of hybrid performance. *Plant Methods* **15**: 1–18.

609 Auinger HJ, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, *et al.*
610 (2016). Model training across multiple breeding cycles significantly improves
611 genomic prediction accuracy in rye (*Secale cereale* L.). *Theor Appl Genet* **129**: 2043–
612 2053.

613 Azevedo CF, de Resende MDV, e Silva FF, Viana JMS, Valente MSF, Resende MFR, *et*
614 *al.* (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC*
615 *Genet* **16**: 1–13.

616 Barter RL, Yu B (2018). Superheat: An R Package for Creating Beautiful and Extendable
617 Heatmaps for Visualizing Complex Data. *J Comput Graph Stat* **27**: 910–922.

618 Basnet BR, Crossa J, Dreisigacker S, Pérez-Rodríguez P, Manes Y, Singh RP, *et al.* (2019).
619 Hybrid Wheat Prediction Using Genomic, Pedigree, and Environmental Covariables
620 Interaction Models. *Plant Genome* **12**: 1–13.

621 Bernardo R (2010). *Breeding for quantitative traits in plants Stemma Press*. Stemma Press:
622 Woodbury.

623 Chessa S, Chiatti F, Ceriotti G, Caroli A, Consolandi C, Pagnacco G, *et al.* (2007).
624 Development of a single nucleotide polymorphism genotyping microarray platform
625 for the identification of bovine milk protein genetic polymorphisms. *J Dairy Sci* **90**:
626 451–464.

627 Covarrubias-Pazaran G (2016). Genome-Assisted prediction of quantitative traits using the
628 r package sommer. *PLoS One* **11**: 1–15.

629 Crossa J, De Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, *et al.* (2010).
630 Prediction of genetic values of quantitative traits in plant breeding using pedigree and
631 molecular markers. *Genetics* **186**: 713–724.

632 Dai Z, Long N, Huang W (2020). Influence of genetic interactions on polygenic prediction.
633 *G3 Genes, Genomes, Genet* **10**: 109–115.

634 Denis M, Bouvet JM (2013). Efficiency of genomic selection with models including
635 dominance effect in the context of Eucalyptus breeding. *Tree Genet Genomes* **9**: 37–
636 51.

637 Dias KODG, Gezan SA, Guimarães CT, Nazarian A, Da Costa E Silva L, Parentoni SN, *et*
638 *al.* (2018). Improving accuracies of genomic predictions for drought tolerance in
639 maize by joint modeling of additive and dominance effects in multi-environment
640 trials. *Heredity (Edinb)* **121**: 24–37.

641 Duenk P, Calus MPL, Wientjes YCJ, Bijma P (2017). Benefits of dominance over additive
642 models for the estimation of average effects in the presence of dominance. *G3 Genes,*
643 *Genomes, Genet* **7**: 3405–3414.

644 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, *et al.* (2011). A
645 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.
646 *PLoS One* **6**: 1–10.

647 Falconer D, Mackay T (1996). *Introduction to quantitative genetics*, Longman, 4th edn.
648 Longmans Green: Harlow, Essex, UK.

649 Galli G, Alves FC, Morosini JS, Fritsche-Neto R (2020). On the usefulness of parental lines

650 GWAS for predicting low heritability traits in tropical maize hybrids. *PLoS One* **15**:
651 1–15.

652 Gaynor RC, Gorjanc G, Hickey JM (2020). AlphaSimR: An R-package for Breeding
653 Program Simulations 2 3. *bioRxiv*: 2020.08.10.245167.

654 Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015). Potential of genotyping-by-
655 sequencing for genomic selection in livestock populations. *Genet Sel Evol* **47**.

656 Granato ISC, Galli G, de Oliveira Couto EG, e Souza MB, Mendonça LF, Fritsche-Neto R
657 (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol Breed* **38**.

658 Hallauer A, Carena M, Filho JM (2010). Quantitative genetics in maize breeding.

659 Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009). Invited review: Genomic
660 selection in dairy cattle: Progress and challenges. *J Dairy Sci* **92**: 433–443.

661 Heaton MP, Keen JE, Clawson ML, Harhay GP, Bauer N, Shultz C, *et al.* (2005). Use of
662 bovine single nucleotide polymorphism markers to verify sample tracking in beef
663 processing. *J Am Vet Med Assoc* **226**: 1311–1314.

664 Henderson CR (1975). Best Linear Unbiased Estimation and Prediction under a Selection
665 Model. *Biometrics* **31**: 423.

666 Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, *et al.* (2014).
667 Evaluation of genomic selection training population designs and genotyping
668 strategies in plant breeding programs using simulation. *Crop Sci* **54**: 1476–1488.

669 Hill W, Robertson A (1968). Linkage disequilibrium in finite populations. *Theor Appl*
670 *Genet* **38**: 226–231.

671 Hou L, Liang W, Xu G, Huang B, Zhang X, Hu CY, *et al.* (2020). Accuracy of genomic
672 prediction using mixed low-density marker panels. *Anim Prod Sci* **60**: 999–1007.

673 Jannink JL (2010). Dynamics of long-term genomic selection. *Genet Sel Evol* **42**: 1–11.

674 Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y (2018). Genomic prediction of
675 breeding values using a subset of SNPs identified by three machine learning methods.
676 *Front Genet* **9**: 1–20.

677 Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, *et al.* (2011). *Genomic*
678 *Selection in Plant Breeding. Knowledge and Prospects.*, 1st edn. Elsevier Inc.

679 Ma Y, Reif JC, Jiang Y, Wen Z, Wang D, Liu Z, *et al.* (2016). Potential of marker selection
680 to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.).
681 *Mol Breed* **36**: 1–10.

682 Matias FI, Barrios SCL, Bearari LM, Meireles KGX, Mateus RG, Do Amaral PNC, *et al.*
683 (2018). Contribution of additive and dominance effects on agronomical and
684 nutritional traits, and multivariate selection on *Urochloa* spp. hybrids. *Crop Sci* **58**:
685 2444–2458.

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

688 Moll R, Lonnquist J, Vélez Fortuno J, Johnson E (1965). The relationship of heterosis and
689 genetic divergence in maize. *Genetics* **52**: 139–144.

690 Muleta KT, Pressoir G, Morris GP (2019). Optimizing genomic selection for a sorghum
691 breeding program in Haiti: A simulation study. *G3 Genes, Genomes, Genet* **9**: 391–
692 401.

693 Müller D, Schopp P, Melchinger AE (2017). Persistency of prediction accuracy and genetic
694 gain in synthetic populations under recurrent genomic selection. *G3 Genes, Genomes,*
695 *Genet* **7**: 801–811.

696 Muñoz F, Rodriguez LS (2016). breedR : An open statistical package to analyse genetic
697 data.

698 Neyhart JL, Tiede T, Lorenz AJ, Smith KP (2017). Evaluating methods of updating training
699 data in long-term genomewide selection. *G3 Genes, Genomes, Genet* **7**: 1499–1510.

700 Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, *et al.* (2019).
701 Package ‘vegan’ Title Community Ecology Package. *Community Ecol Packag* **2**: 1–
702 297.

703 Patterson HD, Williams ER (1976). A new class of resolvable incomplete block designs.

704 *Biometrika* **63**: 83–92.

705 Pocrnic I, Lourenco DAL, Masuda Y, Misztal I (2016). Dimensionality of genomic
706 information and performance of the Algorithm for Proven and Young for different
707 livestock species. *Genet Sel Evol* **48**: 1–9.

708 Prasad SK, Singh TP (1986). Heterosis in relation to genetic divergence in maize (*Zea*
709 mays L.). *Euphytica* **35**: 919–924.

710 Reif JC, Gumpert FM, Fischer S, Melchinger AE (2007). Impact of interpopulation
711 divergence on additive and dominance variance in hybrid populations. *Genetics* **176**:
712 1931–1934.

713 Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, *et al.* (2012). Maximizing
714 the reliability of genomic selection by optimizing the calibration set of reference
715 individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea*
716 mays L.). *Genetics* **192**: 715–728.

717 Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, *et al.* (2015).
718 Efficient Use of Historical Data for Genomic Selection: A Case Study of Stem Rust
719 Resistance in Wheat. *Plant Genome* **8**: 1–10.

720 Santos JPR, De Castro Vasconcellos RC, Pires LPM, Balestre M, Von Pinho RG (2016).
721 Inclusion of dominance effects in the multivariate GBLUP model. *PLoS One* **11**: 1–
722 21.

723 Shao J (1993). Linear model selection by cross-validation. *J Am Stat Assoc* **88**: 486–494.

724 Sousa MB, Cuevas J, Couto EG de O, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, *et al.*
725 (2017). Genomic-enabled prediction in maize using kernel models with genotype ×
726 environment interaction. *G3 Genes, Genomes, Genet* **7**: 1995–2014.

727 Sousa MB, Galli G, Lyra DH, Granato ÍSC, Matias FI, Alves FC, *et al.* (2019). Increasing
728 accuracy and reducing costs of genomic prediction by marker selection. *Euphytica*
729 **215**.

730 Subedi S, Feng Z, Deardon R, Schenkel FS (2013). SNP selection for predicting a

731 quantitative trait. *J Appl Stat* **40**: 600–613.

732 Tayeh N, Klein A, Le Paslier MC, Jacquin F, Houtin H, Rond C, *et al.* (2015). Genomic
733 prediction in pea: Effect of marker density and training population size and
734 composition on prediction accuracy. *Front Plant Sci* **6**: 1–11.

735 Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012). Genomic prediction of
736 hybrid performance in maize with models incorporating dominance and population
737 specific marker effects. *Theor Appl Genet* **125**: 1181–1194.

738 Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014).
739 Genome properties and prospects of genomic prediction of hybrid performance in a
740 breeding program of maize. *Genetics* **197**: 1343–1355.

741 Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, *et al.* (2014). A
742 powerful tool for genome analysis in maize: Development and evaluation of the high
743 density 600 k SNP genotyping array. *BMC Genomics* **15**: 1–15.

744 VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**:
745 4414–4423.

746 Walsh B, Lynch M (2018). *Evolution and Selection of Quantitative Traits*. Oxford
747 University Press: New York.

748 Werner CR, Qian L, Voss-Fels KP, Abbadi A, Leckband G, Frisch M, *et al.* (2018).
749 Genome-wide regression models considering general and specific combining ability
750 predict hybrid performance in oilseed rape with similar accuracy regardless of trait
751 architecture. *Theor Appl Genet* **131**: 299–317.

752 Xu S (2013). Genetic mapping and genomic selection using recombination breakpoint data.
753 *Genetics* **195**: 1103–1115.

754 Yassue RM, Sabadin JFG, Galli G, Alves FC, Fritsche-Neto R (2020). *CV- α : designing
755 validation sets to increase the precision and enable multiple comparison tests in
756 genomic prediction*.

757 Zhang Z, Erbe M, He J, Ober U, Gao N, Zhang H, *et al.* (2015). Accuracy of whole-

758 genome prediction using a genetic architecture-enhanced variance-covariance matrix.

759 *G3 Genes, Genomes, Genet* **5**: 615–627.

760

761 **Fig. 1 Schematic representation of the breeding schemes based on recurrent selection**
 762 **applied in simulated data.** After obtaining the historical population, were developed
 763 populations that served as the basis for the IRS (exclusive steps in orange color) and RRS
 764 (blue color) schemes. Black information was common to both schemes.

765
 766 **Table 1 Number SNP-markers per chromosome as a function of the marker density**
 767 **reduction by different levels of linkage disequilibrium and OTS algorithm.** USP and
 768 HEL datasets.

769
 770 **Table 2 Effective population size (N_e) obtained by additive matrices (A) with markers**
 771 **density reduction in different linkage disequilibrium levels and the OTS algorithm.** USP
 772 and HEL datasets.

773
 774 **Fig. 2 Predictive ability with several scenarios of markers density reduction by different**
 775 **levels of linkage disequilibrium and algorithm for OTS using additive and additive-**
 776 **dominant models. A** USP dataset **B** HEL dataset.

777
 778 **Fig. 3 Simulation of 15 breeding cycles via phenotypic and genomic selection (with three**
 779 **marker densities) by the recurrent intrapopulation selection (IRS) scheme with additive**
 780 **and additive-dominant models. A** Population mean **B** Additive variance **C** Accuracy

781
 782 **Fig. 4 Simulation of 15 breeding cycles via phenotypic and genomic selection (with three**
 783 **marker densities) by the reciprocal recurrent selection (RRS) scheme with additive,**
 784 **additive-dominant, CGA, and SCA G-BLUP models. A** Population mean **B** Best crosses
 785 **C** F_{st} between HG **D** Additive variance HG1 **E** Additive variance HG2 **F** Accuracy

786

USP dataset													
Chrm/MD	M_FULL	M_0.99	M_0.9	M_0.8	M_0.7	M_0.6	M_0.5	M_0.4	M_0.3	M_0.2	M_0.1	M_0.01	M_OTS
1	10 689	10 548	9442	8359	7504	6701	5859	5049	4096	3146	2152	916	233
2	8872	8789	7876	7015	6296	5595	4910	4175	3370	2574	1675	792	186
3	5468	5420	4869	4344	3897	3522	3069	2614	2078	1569	1057	451	134
4	5466	5394	4821	4308	3873	3457	3049	2604	2119	1661	1159	545	121
5	8091	8007	7224	6416	5794	5145	4473	3802	3055	2292	1515	678	173
6	5250	5197	4693	4201	3783	3405	2970	2554	2045	1533	1036	430	113
7	6097	6045	5436	4853	4397	3963	3493	2956	2408	1783	1197	545	136
8	6520	6440	5794	5160	4650	4150	3640	3093	2530	1950	1247	563	151
9	3421	3379	3030	2717	2464	2205	1955	1665	1330	1048	663	303	83
10	3230	3190	2858	2584	2335	2096	1843	1583	1266	979	652	276	78
Total	63 104	62 409	56 043	49 957	44 993	40 239	35 261	30 095	24 297	18 535	12 353	5499	1408
HEL dataset													
1	3977	3976	3762	3558	3405	3247	3059	2881	2673	2442	2089	1325	222
2	3460	3460	3283	3097	2948	2790	2647	2487	2286	2047	1736	1053	188
3	3329	3329	3170	2989	2830	2685	2552	2377	2241	2008	1684	1045	185
4	3057	3057	2918	2771	2651	2516	2389	2246	2097	1880	1612	1005	177
5	3737	3734	3531	3335	3179	3019	2872	2703	2464	2199	1832	1102	212
6	2943	2942	2766	2614	2464	2332	2193	2017	1854	1614	1327	803	172
7	3039	3037	2856	2695	2551	2418	2286	2147	1977	1765	1428	860	173
8	2615	2613	2452	2328	2230	2125	2010	1872	1718	1523	1299	812	147
9	2242	2241	2133	2018	1935	1844	1746	1648	1524	1379	1132	705	138
10	2068	2067	1976	1875	1790	1709	1628	1515	1388	1234	1021	623	120
Total	30 467	30 456	28 847	27 280	25 983	24 685	23 382	21 893	20 222	18 091	15 160	9333	1734

MD	Ne	
	USP	HEL
M_FULL	456.21	230.55
M_0.99	456.19	230.55
M_0.9	456.26	230.48
M_0.8	456.37	230.43
M_0.7	456.32	230.38
M_0.6	456.33	230.26
M_0.5	456.30	230.26
M_0.4	456.55	230.14
M_0.3	456.59	230.12
M_0.2	456.73	230.10
M_0.1	456.85	230.16
M_0.01	457.56	231.33
M_OTS	455.88	229.84

A**USP** **Marker Density (Scenarios)****Model**

add

add_dom

B**HEL**