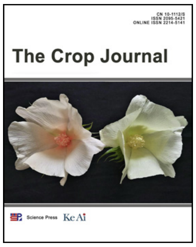
Crop Science Society of China

Available online at www.sciencedirect.com

ScienceDirect



Factors affecting genomic selection revealed by empirical evidence in maize

Xiaogang Liu[a], Hongwu Wang[a], Hui Wang[a], Zifeng Guo[a], Xiaojie Xu[a], Jiacheng Liu[a], Shanhong Wang[a], Wen-Xue Li[a], Cheng Zou[a], Boddupalli M. Prasanna[c], Michael S. Olsen[c], Changling Huang[a,*], Yunbi Xu[a,b,**]

[a]Institute of Crop Science, National Key Facility of Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[b]International Maize and Wheat Improvement Center (CIMMYT), El Batán 56130, Texcoco, Mexico
[c]International Maize and Wheat Improvement Center (CIMMYT), ICRAF campus, United Nations Avenue, Nairobi, Kenya

ARTICLE INFO

Article history:
Received 28 December 2017
Received in revised form 5 March 2018
Accepted 17 March 2018
Available online 22 April 2018

Keywords:
Marker density
Population size
Population relationship
Zea mays

ABSTRACT

Genomic selection (GS) as a promising molecular breeding strategy has been widely implemented and evaluated for plant breeding, because it has remarkable superiority in enhancing genetic gain, reducing breeding time and expenditure, and accelerating the breeding process. In this study the factors affecting prediction accuracy (r_{MG}) in GS were evaluated systematically, using six agronomic traits (plant height, ear height, ear length, ear diameter, grain yield per plant and hundred-kernel weight) evaluated in one natural and two biparental populations. The factors examined included marker density, population size, heritability, statistical model, population relationships and the ratio of population size between the training and testing sets, the last being revealed by resampling individuals in different proportions from a population. Prediction accuracy continuously increased as marker density and population size increased and was positively correlated with heritability; r_{MG} showed a slight gain when the training set increased to three times as large as the testing set. Low predictive performance between unrelated populations could be attributed to different allele frequencies, and predictive ability and prediction accuracy could be improved by including more related lines in the training population. Among the seven statistical models examined, including ridge regression best linear unbiased prediction (RR-BLUP), genomic BLUP (GBLUP), BayesA, BayesB, BayesC, Bayesian least absolute shrinkage and selection operator (Bayesian LASSO), and reproducing kernel Hilbert space (RKHS), the RKHS and additive-dominance model (Add + Dom model) showed credible ability for capturing non-additive effects, particularly for complex traits with low heritability. Empirical evidence generated in this study for GS-relevant factors will help plant breeders to develop GS-assisted breeding strategies for more efficient development of varieties.

* *Corresponding author.*
** *Correspondence to:* Y. Xu, Institute of Crop Science, National Key Facility of Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China.
E-mail addresses: huangchangling@caas.cn (C. Huang), y.xu@cgiar.org (Y. Xu).

Peer review under responsibility of Crop Science Society of China and Institute of Crop Science, CAAS.

https://doi.org/10.1016/j.cj.2018.03.005

1. Introduction

Genomic selection (GS), or genome-wide selection, has become increasingly important in the field of molecular plant breeding with development of high-throughput, cost-effective genotyping technology [1–3]. The GS strategy based on genome-wide polymorphic markers was initially proposed in 2001 with several publications of statistical models [4]. GS can be applied in improving preselection accuracies for complex agronomic traits based on genomic information. The hypothesis of GS primarily depends on the fact that chromosome segments contributing to phenotypic variation are in high linkage disequilibrium (LD) with a minimum of one marker locus within the genome [4]. The GS procedure, it usually utilizes both phenotypic and genotypic data from a training population (TP) to train a statistical model, which can be used to estimate genomic-estimated breeding values (GEBVs) for precise selection of each individual from a candidate (breeding) population that is genotyped without phenotyping [5]. The GS strategy obviously has advantages in comparison with previous molecular breeding technologies such as marker-assisted selection (MAS) and marker-assisted recurrent selection (MARS), which depend on identified significant markers, tagged genes or mapped quantitative trait loci (QTL). Moreover, MAS or MARS has some shortcomings because the search for significant marker-QTL associations has low power in capturing genes with minor effects [6,7]. GS removes the requirement to unearth QTL, and can directly estimate all marker effects in whole genome and capture genetic loci with minor effects for complex traits [8,9]. Beyond that, GS can substantially enhance the rate of annual genetic gain by accelerating breeding cycles and by reducing time and cost because selection of the candidate population depends only on genotypes of individuals without need for phenotypic records [10,11].

GS has long been practiced in livestock and animal breeding [12–14]. For plant breeding, simulation analysis and empirical evaluation of prediction accuracy of GS had been accomplished by using cross-validation method in the experimental populations of Arabidopsis [15], rice [16,17], wheat [18,19], barley [20,21], maize [3,22–24], and forest trees [25–27]. Prediction accuracy (r_{MG}) is regarded as a vital parameter to evaluate the performance of GS in breeding programs. It is usually defined as Pearson's correlation (r) between the true breeding value and the GEBVs of candidate individuals. The factors affecting the estimate accuracy of GEBV will have an influence on prediction ability of GS. These key factors are more or less interrelated in a comprehensive manner. They generally include model performances, relationship between training and breeding populations (BP), heritability of target trait, population size of both TP and BP, population structure, and marker density. r_{MG} varies with particular GS statistical models that depend on prior assumptions and treatment of marker effects [28–31]. Several statistical models have been applied in genomic prediction, including ridge regression best linear unbiased prediction (RR-BLUP) [32,33], genomic best linear unbiased prediction (GBLUP) [3,31,34,35], Bayesian models [4,28,30,36,37], and machine learning models [38–42]. Moreover, developing optimum models with consideration of genotype × environment interaction can significantly improve the predictive ability in multi-environment trials [43–46]. Designing the composition of the TP with reference to BP is an important factor for maintaining a high degree of prediction accuracy in GS breeding programs [47–49]. Agronomic traits with high heritability are regulated by major-effect genes and are rarely affected by environment, and thus they will be positively correlated with higher r_{MG} with good selection response [50,51]. High marker density can improve the proportion of genetic variation explained by molecular markers such as single nucleotide polymorphisms (SNPs), and thus result in high prediction accuracy, the latter being used as a selection criterion to assist plant breeders to select target traits with precision [51,52]. Population structure, as a specific factor affecting GEBV prediction, can give rise to biased estimates in GS [53–55]. Hence, taking all GS-relevant factors into account can be deemed a rational strategy that will be more helpful for plant breeders in making selections based on the superiority of breeding individuals rather than their phenotypic data alone.

In this research article, we focus on better understanding GS-relevant factors to seek a few available measures to instruct plant breeders in reasonably designing GS breeding programs for enhancing genetic gain per unit time while reducing cost. The major challenge in GS is how to integrate empirical results into plant breeding practices to improve crop yield and create more economic values. The datasets used in the study contain genotypic data from the 55 K SNP array and phenotypic data for six complex traits generated for one natural and two biparental populations. The objectives of the case study were to (1) evaluate the prediction accuracy of different traits in three maize populations; (2) assess the effects of six GS-relevant factors on prediction accuracy, including population size, marker density, heritability, model performance, relationship between TP and BP, and the relative population size between the training and testing sets; and (3) utilize the results to make recommendations for plant breeders in implementing GS in commercial breeding programs.

2. Materials and methods

2.1. Plant materials

The experiment started with one natural population (NAT) and three biparental populations, including recombinant inbred line (RIL), doubled haploid (DH) and $F_{2:3}$ populations. The natural population comprised 435 maize elite inbred lines derived from temperate and tropical regions around the world, so that it has rich genetic diversity. The three biparental populations were derived from a widely grown single-cross maize hybrid, Zhongdan 909, with elite Chinese inbred lines Zheng 58 and HD568 as parents, and included 212 RILs, 79 DH lines, and 304 $F_{2:3}$ families. As the RILs were generated with many generations of selfing, and DH population had only a limited number of individuals, we combined the RILs and DH lines as a single population, and named it the RIL&DH population.

2.2. Field trial and phenotyping

The natural population was evaluated in Henan province in 2014 and 2015. The other lines were grown at the same location in 2015 and 2016. Two-row plots were grown and harvested as grain trials. The experiment in each year was performed in a randomized incomplete block design with two replications. Phenotyping was done for six agronomic traits related to yield: plant height (PH, cm), ear height (EH, cm), ear length (EL, cm), ear diameter (ED, cm), grain yield per plant (GYP, kg) and hundred-kernel weight (HKW, g). Phenotypic values of GYP and HKW were adjusted to 14% grain moisture.

2.3. Phenotypic data analysis and heritability estimation

SAS 9.2 (SAS Institute Inc., Cary, NC) was used to evaluate multi-year trials with best linear unbiased prediction (BLUP) using the SAS PROC MIXED procedure with all factors as random effects; the model followed was

$$y_{ilj} = \mu + g_i + e_l + ge_{il} + b_j + \varepsilon_{ilj},$$

where y_{ilj} is the phenotypic value, μ is overall mean, g_i is the effect of ith genotype, e_l is the effect of the lth environment, ge_{il} is the effect of genotype × environment interaction, b_j is the effect of jth replication, and ε_{ilj} is the model residuals. Broad-sense heritability was estimated as the ratio of genetic variance to total phenotypic variance:

$h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_{ge}^2/e + \sigma_\varepsilon^2/re)$,where σ_g^2, σ_{ge}^2, and σ_ε^2 are genotypic, genotype × environment interaction and random error variance components, respectively, and e and r are the numbers of environments and replicates. Analyses of correlation and variance (ANOVA) were performed by SAS CORR PROC and SAS GLM PROC procedures.

2.4. Genotypic data analysis

The newly developed maize 55 K SNP Array [56] was used to genotype all inbred lines in the natural and RIL&DH populations and F_2 plants that were used for development of $F_{2:3}$ population. Markers with minor allele frequencies (MAF) less than 5% and missing values greater than 10% were removed. Finally there were 37,803 and 8271 SNP markers in the genotypic data set for the natural population and biparental populations, respectively. A total of 7861 SNP markers shared among different populations were used as genotypic data to evaluate the effect of genetic relationship on prediction accuracy.

2.5. GS-relevant factors and their settings

Prediction accuracy was investigated using different ranges of genetic parameters. The effect of marker density was tested using 12 to 15 levels (N_m) (each level containing 50 to 8271 or 37,803 markers depending on populations) (i.e., 50, 100, 200, 300, 400, 500, 750, 1000, 3000, 5000, 7000, 10,000, 20,000, 30,000, and all SNPs). To understand the effect of population size, 5 to 7 training population sizes (N) (each with 100 to 400 lines) (i.e., 100, 150, 200, 250, 300, 350, 400, and all lines in the RIL&DH population) and 11 different ratios of the population size allocated between training and testing sets (1/5 to 10/1) (i.e., 1/5, 1/4, 1/3, 1/2, 1/1, 2/1, 3/1, 4/1, 5/1, 6/1, and 10/1) were examined. To evaluate the effect of ratio of population size between the training and testing sets, we randomly divided the experimental population into two portions with one as training set to estimate the GEBVs for the other (testing set). Population structure was assessed for the natural population by the fastSTRUCTURE algorithm [57] with K values set as 1 to 10. The effect of population structure was studied by classifying the natural population into temperate, tropical and other germplasm groups based on clustering analysis. The effect of genetic relationship on prediction accuracy was then evaluated by pooling the natural and RIL&DH training populations to predict $F_{2:3}$ performance, and by pooling temperate and tropical inbred lines as a training population in different proportions to predict tropical line performance. The natural population was used as a training population to evaluate GEBVs for RIL&DH and $F_{2:3}$ populations, whereas the biparental populations were used to predict each other, and temperate lines were used as a training population to predict tropical lines. F-statistics (F_{ST}, a descriptive index of genetic differentiation and distance) within and among natural and biparental populations were evaluated by VCFtools [58].

2.6. Statistical models

Seven statistical models were examined in each experiment, including ridge regression best linear unbiased prediction (RR-BLUP), genomic BLUP (GBLUP), BayesA, BayesB, BayesC, Bayesian least absolute shrinkage and selection operator (Bayesian LASSO, BayesL), and reproducing kernel Hilbert space (RKHS).

2.6.1. RR-BLUP model

The RR-BLUP model was fitted using the R package *rrBLUP* version 4.4 [33]; the mixed model is described as:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z}\boldsymbol{u} + \varepsilon$$

where $\mathbf{y}$ is the vector (n × 1) of observations, $\mathbf{1}_n$ is the vector (n × 1) of ones and μ is the fixed effects, ε is the vector (n × 1) of independently random residuals with assumed distribution N $(0, \boldsymbol{I}\,\sigma_\varepsilon^2)$, $\mathbf{Z}$ is the design matrix (n × m) for random effects, and $\boldsymbol{u}$ is the vector of random effects with $\boldsymbol{u} \sim N(0, \boldsymbol{K}\,\sigma_u^2)$, $\boldsymbol{K}$ being an identity matrix in this case [33]. In addition, n is the number of individuals, and m is the number of markers.

2.6.2. GBLUP model

The standard GBLUP model focusing on additive genetic effects is described as:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z}\boldsymbol{u} + \varepsilon$$

where $\mathbf{Z}$ is the design matrix for random effects $\boldsymbol{u}$ with $\boldsymbol{u} \sim N(0, \mathbf{G}\,\sigma_u^2)$, and $\mathbf{G}$ is the genomic relationship matrix calculated using marker information [35]. GBLUP was fitted using the R package *BGLR* version 1.0.5 [59].

2.6.3. Extended GBLUP model for additive and dominance effects (Add + Dom model)

The extended GBLUP model including additive and dominance effects can be described as:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z}_a \boldsymbol{u}_a + \mathbf{Z}_d \boldsymbol{u}_d + \varepsilon$$

where $\boldsymbol{u_a}$ is the vector of random effects for additive genetic effects with $\boldsymbol{u_a} \sim N\ (0, \mathbf{G}_a\ {\sigma_{u_a}}^2)$, $\boldsymbol{u_d}$ is the vector of random effects for dominance effects with $\boldsymbol{u_d} \sim N\ (0,\ \mathbf{G}_d\ \ {\sigma_{u_d}}^2)$, $\mathbf{G}_a = \mathbf{W}_a\ \mathbf{W}_d'/\sum 2p_m q_m$, $\mathbf{W}_a$ is an $n \times m$ matrix with $\mathbf{W}_{anm} = \mathbf{Z}_{anm} - 2p_m$, $\mathbf{G}_d = \mathbf{W}_d\mathbf{W}_d'/\sum 2p_m q_m\ (1–2p_m q_m)$, $\mathbf{W}_d$ is an $n \times m$ matrix with $\mathbf{W}_{dnm} = \mathbf{Z}_{dnm} - 2p_m q_m$, $\mathbf{Z}_a$ is a design matrix for $\boldsymbol{u_a}$ composed of 0, 1, and 2 for genotype A_1A_1, A_1A_2, and A_2A_2, respectively, $\mathbf{Z}_d$ is a design matrix for $\boldsymbol{u_d}$ which consists of 0 and 1 for homozygotes and heterozygotes, respectively, p_m is the frequency of allele 2 at locus m, q_m is the frequency of allele 1 at locus m [60]. The Add + Dom model for additive and dominance effects was fitted using the R package *sommer* version 3.0 [61].

2.6.4. *Bayesian models*

The Bayesian models have different prior distributions [30,36,37], and the general model can be represented as:

$$\mathbf{y} = \mathbf{1}_n\ \mu + \mathbf{Z}\boldsymbol{u} + \varepsilon$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{Z}$ is the design matrix for random effects, and $\boldsymbol{u}$ is the vector of random effects. The hyperparameter settings of each Bayesian model were based on default choices in the R package *BGLR* described in built-in rules of the introduction [59]. These Bayesian models were fitted using R package *BGLR* version 1.0.5 [59], and the Gibbs sampler was run for 15,000 iterations with the first 5000 samples discarded as burn in.

2.6.5. *Reproducing kernel Hilbert space (RKHS) model*

RKHS, as a semiparametric approach for GS [38,39,62,63], could have the ability to capture nonadditive and epistatic effects [64], and the RKHS model was implemented using R package *BGLR* version 1.0.5 [59]. The multi-kernel method was used and the model is depicted as:

$$\mathbf{y} = \mathbf{1}_n\ \mu + \boldsymbol{u} + \varepsilon;$$

where $\boldsymbol{u}$ is a vector of random effects, and the assumed distribution of $\boldsymbol{u}$ is normal distribution $N\ (0, \overline{\boldsymbol{K}}\,\overline{\sigma}_u^2)$, where $\overline{\sigma}_u^2 = \sum_{l=1}^{L} \sigma_{u_l}^2$, $\overline{\boldsymbol{K}} = \sum_{l=1}^{L} \boldsymbol{K}l\ \sigma_{u_l}^2 \overline{\sigma}_u^{-2}$, and $\overline{\boldsymbol{K}}$ is the weighted average of multiple reproducing kernels (hence named kernel averaging), $\boldsymbol{K}_l$ is the reproducing kernel that was evaluated at the lth value of the bandwidth parameter [39,59]. In this model, bandwidth parameters were set to values 1/5*M*, 1/*M*, and 5/*M* based on the reference of the *BGLR* package [59], *M* is the median squared Euclidean distance between lines.

2.7. *Model comparisons*

Prediction accuracies obtained from cross-validation were used to perform a hierarchical clustering analysis to evaluate similarities between models. First, the prediction accuracies of all traits and populations for each model were standardized to zero mean and unit variance. A matrix of Euclidean distances between the statistical models was then generated using the standardized values and was averaged and used to perform a hierarchical clustering based on Ward's criterion. The cluster dendrogram was subsequently constructed using R version 3.3.3 [65].

2.8. *Cross-validation*

To assess the effect for each factor, such as marker density, population size, and statistical models, we used a 10-fold cross-validation scheme that randomly partitions each population dataset into training and testing sets and calculated the correlation between true breeding values and GEBVs within the testing group. The cross-validation scheme was implemented with 100 replications, and the average of correlation coefficients was defined as the genome-wide prediction accuracy (r_{MG}). The GBLUP model was used to evaluate the effects of GS-relevant factors.

3. Results

3.1. *Effects of marker density, population size and heritability*

The estimates of r_{MG} continuously increased as marker density and population size increased until they reached plateaux. For example, in the natural population, only a non-significant increase was obtained on r_{MG} for yield-related traits, when the number of SNPs increased from 7000 to 30,000, and the r_{MG} estimate based on 37,803 markers was similar to that from 7000 high-quality markers selected randomly (Fig. 1a, c). However, only a slight gain in r_{MG} was obtained when the number of markers increased from 1000 to 8000 in the RIL&DH population, showing that 1000 markers were enough to achieve a reasonable accuracy in biparental populations (Fig. 1b, d). As the population size increased, r_{MG} increased, and a slight gain in r_{MG} was obtained when the population size reached 250 (Fig. 2). Generally, targeted traits with high h^2 showed high prediction accuracies. The correlation coefficients between r_{MG} and heritability among traits were 0.83 ($P = 0.04$), 0.89 ($P = 0.02$) and 0.95 ($P < 0.01$) in natural, RIL&DH and $F_{2:3}$ populations, respectively (Table 1). Further improvement on prediction accuracy could be achieved with increased h^2.

3.2. *Balanced population sizes between training and testing sets in GS*

The effect of population size on r_{MG} in genomic selection has been evaluated based on population size per se rather than the ratio of population sizes between training and testing sets. In this study, we examined various ratios of population size between training and testing sets. The prediction accuracy of genome-wide prediction increased as more samples were allocated to the training set. In the natural population, a slight gain in r_{MG} was observed when the ratio of population size increased from 3 to 10 times (Fig. 3). A similar result was observed in other types of populations, where a relatively higher estimate of r_{MG} was obtained when the training set was three times as large as the testing set, indicating that the training set should be three times larger than the testing set when the total population size is smaller than 400. A larger population size is required in further research for verifying whether the threefold ratio of population size of training: testing sets is enough to achieve the best performance of GS.

3.3. *Effects of genetic relationship between training and breeding populations*

To evaluate the effects of genetic relationship on prediction accuracy, the natural population was used as a training

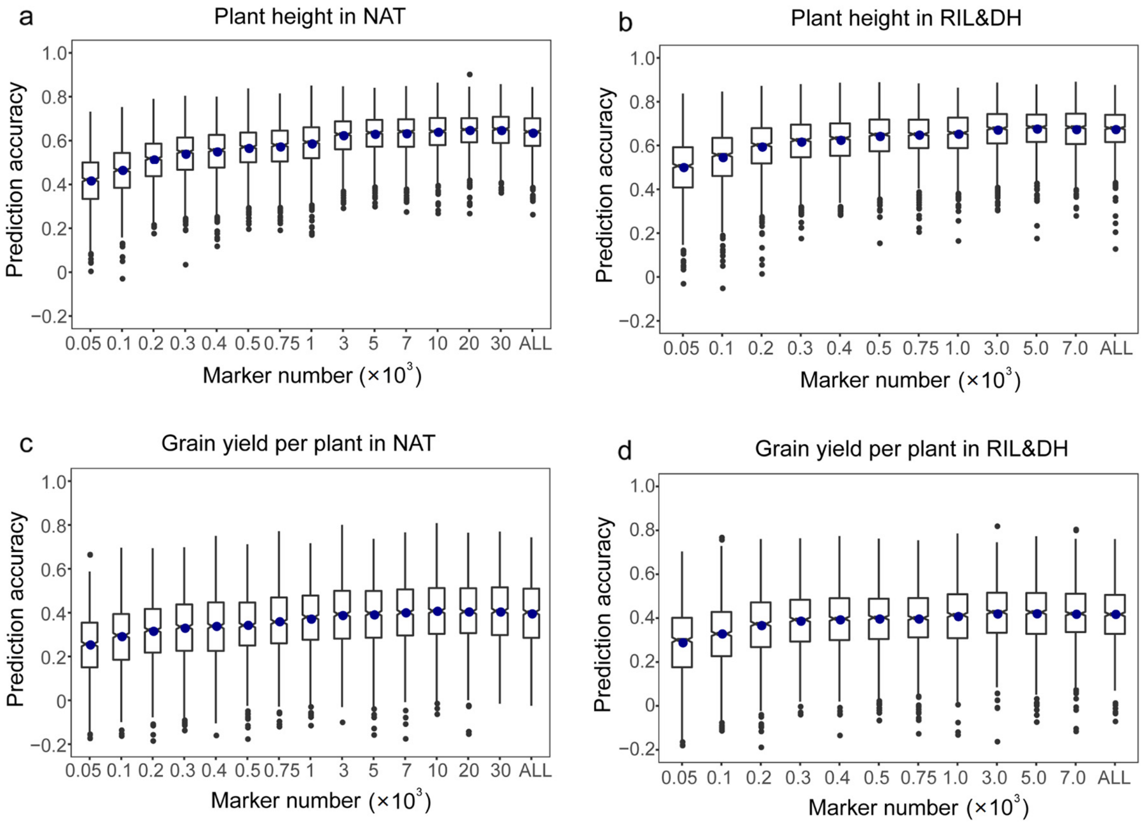


Fig. 1 – Effects of marker density on prediction accuracy determined by natural and RIL&DH populations. (a) and (c) Effect of marker density: natural population (N = 435) with 50 to 37,803 (ALL) markers; (b) and (d) RIL&DH population (N = 291) with 50 to 8271 (ALL) markers; NAT, natural population; RIL&DH, a population containing RILs and DH lines; the GBLUP model with a 10-fold cross-validation scheme was implemented.

population to predict the estimated breeding value using the RIL&DH and $F_{2:3}$ populations as breeding populations. The RIL&DH population was also used to predict the $F_{2:3}$ population. When the natural population was used to predict biparental populations, the prediction accuracy was either negative or very low for almost all traits. Outperformance results were obtained when the two biparental populations were used to predict each other. Taking plant height as an example, using the natural population to predict the $F_{2:3}$ and RIL&DH populations resulted in negative r_{MG} (−0.15 and −0.14, respectively), whereas r_{MG} estimates were up to 0.68 and 0.61 when the two biparental populations were used to predict each other (Table 2). To further examine genetic relationship, we constructed mixed natural and biparental populations in different modes. In the first mode, the mixed populations comprised 300 inbred lines that were randomly selected from the natural and RIL&DH populations. In the second mode, the mixed populations were composed of all inbred lines from the natural population but different proportions of inbred lines from RIL&DH population. For all traits, prediction accuracies increased with more related individuals included in the training population. For plant height, r_{MG} increased from 0.53 to 0.66 in the first mixed mode and from 0.36 to 0.65 in the second (Table 2). A similar analysis was implemented in the natural population that was classified into temperate, tropical and mixed group by the fastSTRUCTURE algorithm. r_{MG} was −0.04 for plant height when the temperate group was used as the training population to predict the tropical group, but it increased from 0.30 to 0.38 when more tropical inbred lines were added to the training population (Table 2). Similar tendencies were observed for other traits.

As a descriptive statistics index of genetic differentiation, F_{ST} can be used to show the genetic distance and relationship between populations, and also indicate the variance in allele frequencies of genomic regions between populations. F_{ST} values between natural and biparental populations, and between RIL&DH and $F_{2:3}$, were 0.1002 and 0.1013, respectively, which were significantly large in comparison with that between the two biparental populations, RIL&DH and $F_{2:3}$ (0.0006, Table 3). The F_{ST} value between temperate and tropical groups was relatively high (0.0982, Table 3).

3.4. *Comparative results from different statistical models*

Compared with other models, RKHS had slightly higher prediction accuracies for most traits in the $F_{2:3}$ population, which contains heterozygous individuals. As for plant height, the accuracies obtained by using RR-BLUP, GBLUP, BayesA, BayesB, BayesC, and BayesL models were similar for all the natural, RIL&DH and $F_{2:3}$ populations, yet a higher r_{MG} was observed by using the RKHS model with the $F_{2:3}$ population (Fig. 4a, b). Prediction accuracy was significantly higher for

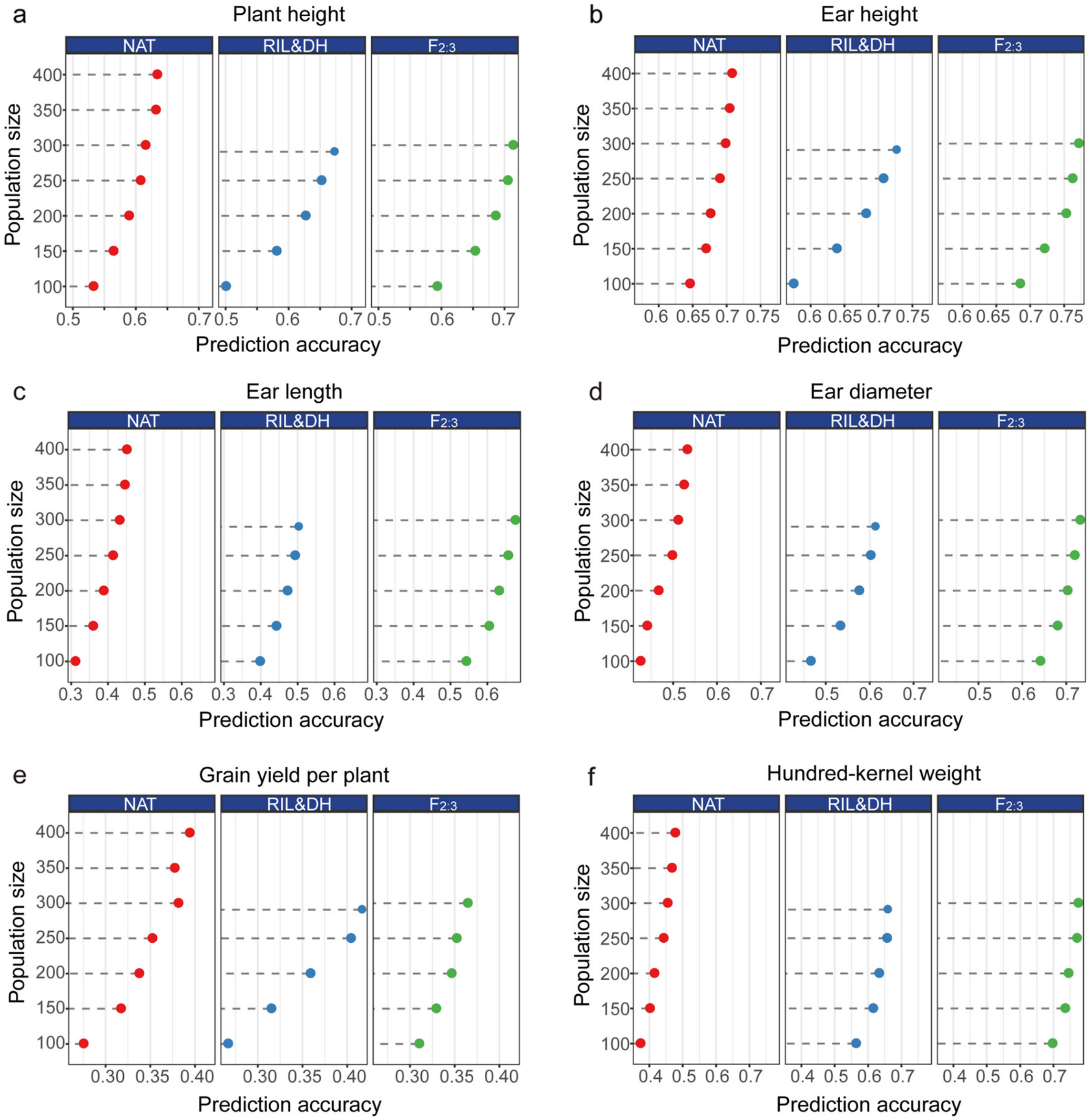


Fig. 2 – Effects of population size on prediction accuracy as determined by three tested populations. (a) to (f) Effect of population size: natural population (N_m = 37,803) with 100 to 400 lines; RIL&DH population (N_m = 8271) with 100 to 291 lines; $F_{2:3}$ population (N_m = 8271) with 100 to 300 families; NAT, natural population; RIL&DH, a population containing RILs and DH lines; $F_{2:3}$, F_3 families derived from F_2 plants; the GBLUP model with a 10-fold cross-validation scheme was implemented.

GYP with the RKHS model and prediction with the $F_{2:3}$ population was markedly better than those with other populations (Fig. 4b). Moreover, prediction accuracy for GYP in the $F_{2:3}$ population was 0.54 with the additive and dominance model, which is significantly better than 0.36 using the additive model alone ($P < 0.01$), although no similar estimates between these two models were obtained for other traits (Fig. 4c). A dendrogram was constructed with hierarchical clustering using prediction accuracies from different statistical models, traits and populations, indicating that the RKHS model was distinctly different from others and was branched out. The BayesA and BayesB models were clustered into one group, while RR-BLUP, GBLUP, BayesC, and BayesL models were clustered together (Fig. 4d).

4. Discussion

Marker density and population size are important factors that affect prediction accuracy. Many reports indicate that the accuracy of genome-wide prediction increases as marker density and population size increase separately [50–52,66]. However, the number of markers required for r_{MG} to plateau varies from population to population [2,67]. This phenomenon can be attributed to the complexity of genetic structure between groups within a population and different levels of diversity among populations. More specifically, natural populations, including the one used in this study, usually have significant genetic structure and high LD levels between

Table 1 – Heritability (h^2) and prediction accuracy (r_{MG}) estimated for agronomic traits, and their Pearson's correlation in natural (NAT), RIL&DH, and $F_{2:3}$ populations.

Traits [a]	NAT		RIL&DH		$F_{2:3}$	
	h^2	r_{MG} [d]	h^2	r_{MG}	h^2	r_{MG}
PH	0.95	0.63	0.92	0.66	0.83	0.71
EH	0.94	0.71	0.92	0.72	0.86	0.77
EL	0.88	0.45	0.82	0.49	0.86	0.66
ED	0.87	0.53	0.88	0.60	0.89	0.72
GYP	0.68	0.40	0.70	0.41	0.65	0.36
HKW	0.79	0.48	0.84	0.65	0.87	0.77
Pearson's r [b]	0.83		0.89		0.95	
P-value [c]	0.04		0.02		< 0.01	

[a] PH, plant height (cm); EH, ear height (cm); EL, ear length (cm); ED, ear diameter (cm); GYP, grain yield per plant (kg); HKW, hundred-kernel weight (g).

[b] Pearson's correlation coefficient between h^2 and r_{MG}.

[c] Numbers denote the P-values for significant test of correlation.

[d] GBLUP model was implemented; NAT ($N = 435$, $N_m = 37{,}803$); RIL&DH ($N = 291$, $N_m = 8271$); $F_{2:3}$ ($N = 304$; $N_m = 8271$).

adjacent markers [68], and high-density markers will be needed to ensure that at least one marker can be in LD with trait-associated loci to achieve precise prediction [4]. However, biparental populations have a clear genetic structure, and finite chromosome recombination events will be introduced in the course of development [69]. Thereby, a moderate marker density will be enough to ensure that each gene-related locus can be in linkage with at least one marker, and it will most likely achieve a better prediction [15]. The results from this study provide some instructive guidance for molecular breeding programs with maize. For example, the gains in r_{MG} began to plateau once the number of markers increased to 7000 and 1000 in the natural and biparental populations, respectively. Additional markers become largely redundant with no further improvement of predictive ability when the r_{MG} has reached a plateau. Hence, to determine how many markers should be used in GS-assisted commercial breeding, schemes based on the above-mentioned results can be regarded as a reference to reduce breeding costs. With regard to population size in GS, marker effects could be more precisely estimated with increased population size [70,71]. However, more attention should be given to use of historical data collected from multi-environment trials conducted for several years. With such information GS performance is more likely to improve and should accelerate breeding outcomes with significantly enhanced genetic gain.

The ratio of population size between training and testing sets was rarely taken into consideration in previous GS studies that primarily focused on the effect of population size on prediction accuracy. In this study, we revealed that r_{MG} generally increased as the ratio increased (with more samples included in the training set). A slight gain in r_{MG} was still achieved even when the training set became three times as large as testing set, compared to a previous study where a slight gain of r_{MG} was observed when the training set was more than half of total population [2]. The reasons for this difference are probably due to the population size and targeted traits. However, both previous and current results could be used to provide some guidance to develop GS-assisted breeding projects. Currently, identifying an optimized training population is one of the most important challenges for plant breeders when they implement GS prior to multi-environment trials. Large breeding companies, such as the multinationals, which have adequate financial support, could develop massive training populations to increase r_{MG}. Small companies could implement GS-assisted breeding by having the training population three times larger than the breeding population, as revealed by testing different ratios of population size in training and testing sets in this study. However, limited population sizes ranging from 300 to 400 were used in this study, and larger population sizes should be considered in future GS studies and GS-assisted breeding. This study also revealed that the increase in r_{MG} with current ratios largely depended on the population types and structures. For example, the gains in r_{MG} for plant height were 0.11, 0.15, and 0.15 for natural, RIL&DH and $F_{2:3}$ populations, respectively, when the ratio increased from 0.2 to 3.0 times (data not shown). The natural population showed a relatively small increase of r_{MG} compared to the other two populations. This might be due to different levels of complexity in genetic structure among populations and can result in biased and inaccurate estimation of marker effects.

The effect of genetic relationship on prediction accuracy, as one of the most important factors, has been extensively evaluated in previous empirical studies, concluding that the relationship between training and breeding populations has a significant influence on prediction ability [52,72]. This study revealed that the prediction accuracy could be improved significantly with increasing genetic relationship, from distant populations to biparental populations. The r_{MG} was extremely low when the natural population was used as a training population to predict biparental populations, indicating that such prediction has no significant potential in practical breeding. However, r_{MG} increased with more RIL&DH lines added into the natural population for training. With genetic relationships changing from less-related to closely-related in this study prediction accuracies were improved from a negative low level to positively higher levels. In fact, candidate inbred lines in commercial seed companies would continually accumulate in the process of selection and breeding, and these will be added to improve the training set based on the genetic background in reference and inference groups. For example, historical data for candidate inbred lines derived from the Iowa Stiff Stalk Synthetic population, a famous heterotic maize group, can be used to estimate breeding values, and thus alternative inbred lines could be selected and included in the parental panel in the next-round selection. However, one of the most important steps in applying genetic relationship data in GS could be analysis of the initial populations to understand their population structure and genetic background, and thus more precise prediction accuracy and virtually accelerated breeding progress can be achieved. A few previous studies of genetic relationship were performed by including related individuals in the training population [73] or using half-sib populations to investigate GS prediction accuracy [74], without taking allele frequencies between training and breeding populations into account.

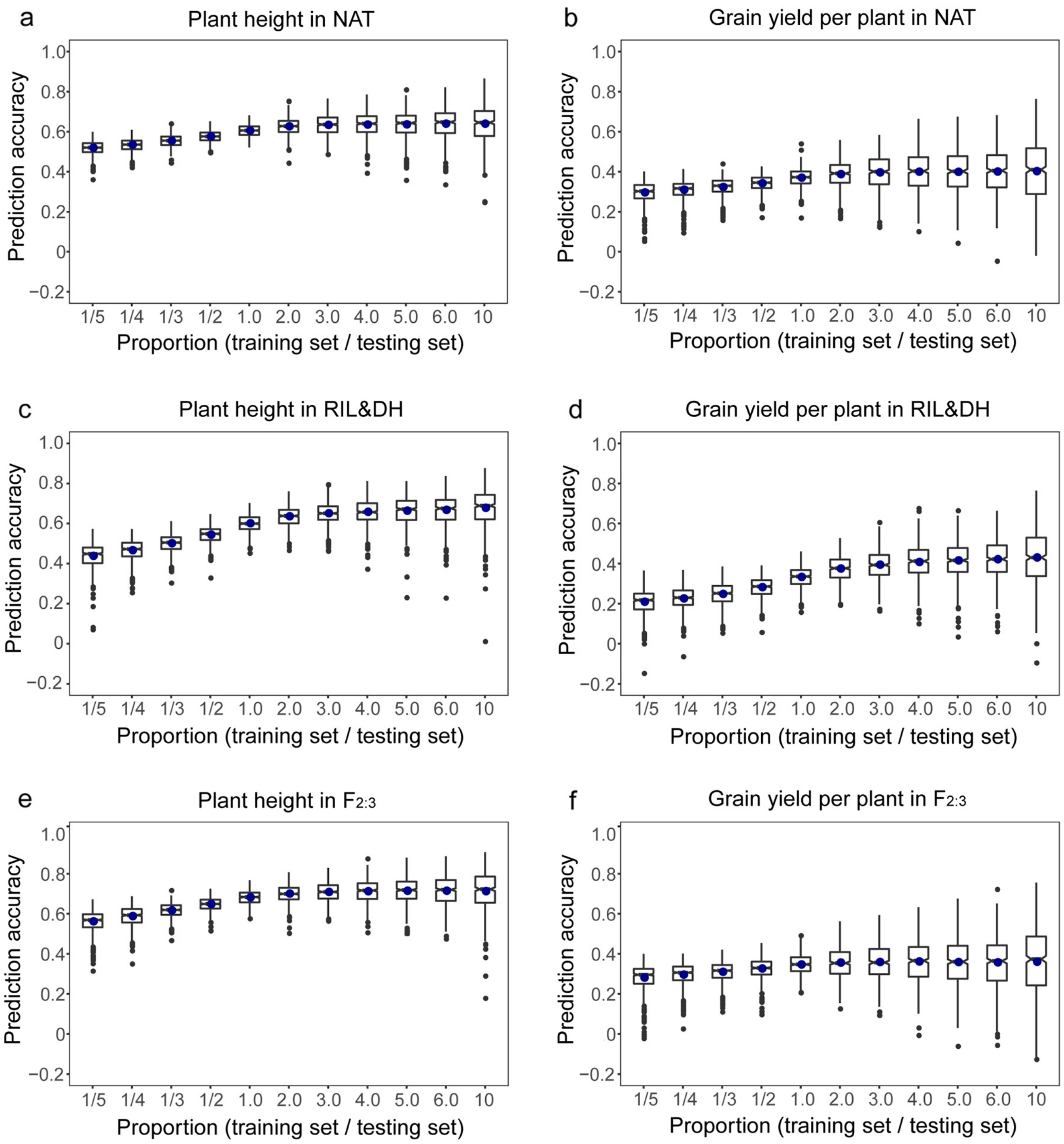


Fig. 3 – Effects of ratio of population size between training and testing sets on prediction accuracy. (a) and (b) natural population (*N* = 435) with 37,803 markers. (c) and (d) RIL&DH population (*N* = 291) with 8271 markers; (e) and (f) $F_{2:3}$ population (*N* = 304) with 8271 markers; NAT, natural population; RIL&DH, a population containing RILs and DH lines; $F_{2:3}$, F_3 families derived from F_2 plants; GBLUP model was implemented.

Prediction accuracy in GS largely depends on accurate estimation of marker effects by statistical models. The estimation of marker effects is affected by allele frequencies at each locus in the whole genome and vary from population to population. The F_{ST} value between natural and biparental populations, which provides important insights into genetic distance, was larger than that between the biparental, RIL&DH and $F_{2:3}$ populations. In this study, low r_{MG} values were obtained using the natural population to predict the biparental populations and using the temperate group as training set to predict tropical lines. Such low r_{MG} estimates can be largely attributed to the remarkably different allele frequencies in most genomic regions that probably give rise to biased estimation of marker effects. With biased marker effects, unreasonable breeding values for candidate individuals can be generated by matrix multiplication, and therefore cannot be applied in GS-assisted breeding programs. Hence, it could be advisable to add some correlative inbred lines into the training population in order to improve prediction accuracy as proposed [48,75].

An optimal GS model should have maximum prediction ability. The RR-BLUP and GBLUP models, which meet the assumption that random effects had equal variance, had a similar predictive ability in comparison with the other Bayesian methods with different prior assumptions [28,30,35]. Our results indicate that prediction accuracy was not significantly influenced by Bayesian models, although they have different assumptions of priori distributions for marker effects. Several previous studies reported similar results for different traits and populations [31,36,50]. However, the RKHS model had a slight superiority over other models especially when the population was comprised of heterozygous individuals, indicating that the RKHS model is capable of evaluating marker effects for heterozygous genotypes as revealed in previous studies [62,76]. More specifically, the RKHS model as a semiparametric and nonlinear dimensionality reduction approach can efficiently capture non-additive effects and improve the prediction accuracy for heterozygous populations [62,63]. In this study, the higher prediction accuracy was also observed for GYP when the additive-dominance model was used to evaluate breeding values. The additive-dominance model efficiently captures non-additive effects when traits have low heritability. The RKHS model showed a similar prediction accuracy to the additive-dominance model in cross-validation of GYP, indicating that both models have equal predictive ability and better performance in evaluating non-additive effects and in estimating breeding values for low heritability traits. Thus, implementing optimum models into heterozygous populations, which is more likely to improve breeding efficiency, is important when GS is used in practical breeding.

Table 2 – Evaluation of the effects of genetic relationship on prediction accuracy.

Types	PH[d]	EH	EL	ED	GYP	HKW
NAT - F$_{2:3}$[a]	−0.15	−0.19	−0.02	−0.29	−0.09	−0.10
NAT - RIL&DH	−0.15	−0.22	0.06	−0.09	−0.20	−0.18
RIL&DH - F$_{2:3}$	0.68	0.76	0.54	0.60	0.28	0.71
F$_{2:3}$ - RIL&DH	0.61	0.70	0.47	0.52	0.23	0.63
50NAT + 250RIL&DH - F$_{2:3}$[b]	0.66	0.73	0.53	0.59	0.27	0.70
100NAT + 200RIL&DH - F$_{2:3}$	0.63	0.71	0.51	0.57	0.26	0.68
150NAT + 150RIL&DH - F$_{2:3}$	0.59	0.67	0.49	0.54	0.23	0.65
200NAT + 100RIL&DH - F$_{2:3}$	0.53	0.61	0.45	0.48	0.19	0.61
250NAT + 50RIL&DH - F$_{2:3}$	0.35	0.41	0.37	0.37	0.15	0.46
435NAT + 50RIL&DH - F$_{2:3}$	0.36	0.39	0.34	0.34	0.14	0.43
435NAT + 100RIL&DH - F$_{2:3}$	0.52	0.58	0.43	0.48	0.20	0.59
435NAT + 150RIL&DH - F$_{2:3}$	0.59	0.66	0.48	0.53	0.24	0.64
435NAT + 200RIL&DH - F$_{2:3}$	0.63	0.70	0.51	0.57	0.27	0.67
435NAT + 250RIL&DH - F$_{2:3}$	0.65	0.72	0.52	0.59	0.27	0.69
TEM - TRO[c]	−0.04	0.05	0.01	0.05	−0.05	−0.10
50TEM + 100TRO - TRO	0.38	0.37	0.29	0.29	0.24	0.39
100TEM + 50TRO - TRO	0.32	0.30	0.18	0.23	0.16	0.29
155TEM + 50TRO - TRO	0.31	0.30	0.18	0.23	0.15	0.26
155TEM + 100TRO - TRO	0.37	0.36	0.25	0.29	0.23	0.36

[a] Training-breeding population set: training population is given at the left of the hyphen “-”, the breeding population is included at the right. The natural population (NAT) was used as the training population to estimate GEBVs for the RIL&DH population. RIL&DH, a population containing RILs and DH lines; F$_{2:3}$, F$_3$ families derived from F$_2$ plants; GBLUP model was implemented.
[b] A mixed population that contains 50 lines from the natural population and 250 lines from the RIL&DH population were used to evaluate the GEBVs for the F$_{2:3}$ population. The number is defined as the number of lines selected randomly from the targeted population.
[c] TEM, temperate group; TRO, tropical group.
[d] PH, plant height (cm); EH, ear height (cm); EL, ear length (cm); ED, ear diameter (cm); GYP, grain yield per plant (kg); HKW, hundred-kernel weight (g).

Table 3 – F_{ST} values estimated between pairwise comparisons among three tested populations and among two inbred line groups.

Population	NAT	RIL&DH	F$_{2:3}$	TEM	TRO
NAT[a]	0	0.1002	0.1013		
RIL&DH	0.1002	0	0.0006		
F$_{2:3}$	0.1013	0.0006	0		
TEM				0	0.0982
TRO				0.0982	0

[a] NAT, natural population; RIL&DH, a population containing RILs and DH lines; F$_{2:3}$, F$_3$ families derived from F$_2$ plants; TEM, temperate group; TRO, tropical group.

5. Conclusions

Plant breeding as a scientific and aesthetic procedure should be implemented by considering not only the grain yield of hybrids created from elite inbred lines but also the genetic gain achieved in the course of breeding cycles. Genetic gain may be regarded as the yield improvement achieved by artificial selection, and can be enhanced through refining the field management, enlarging the experimental scale, shortening the breeding cycle, heightening the selection intensity, and balancing costs-benefits. GS, which does not involve a procedure for QTL detection or marker selection, has advantages of accelerating the breeding cycle, reducing breeding costs, and improving breeding efficiency, and also can be combined with other strategies to augment genetic gain. Hence, any ways to improve GS can enhance genetic gain. Using six yield-related traits and three populations in this study a few suggestions for plant breeding were proposed based on evaluation of several important factors that affect prediction accuracy in GS. Increased prediction accuracy occurs with increases in marker density and population size until r_{MG} comes a plateau. Then a moderate marker density and an appropriate population size can ensure adequate predictive performance of GS and subsequently optimize costs in developing candidate lines and potential hybrids. The effects of genetic relationship on r_{MG} should be taken into account in the breeding process to determine whether targeted lines can be selected successfully and whether superior hybrids can be created. Plant breeders aiming at better predictability should carefully consider the choice of inbred lines for inclusion in the training population. Further studies on statistical models will be required to develop improved models that can efficiently capture non-additive

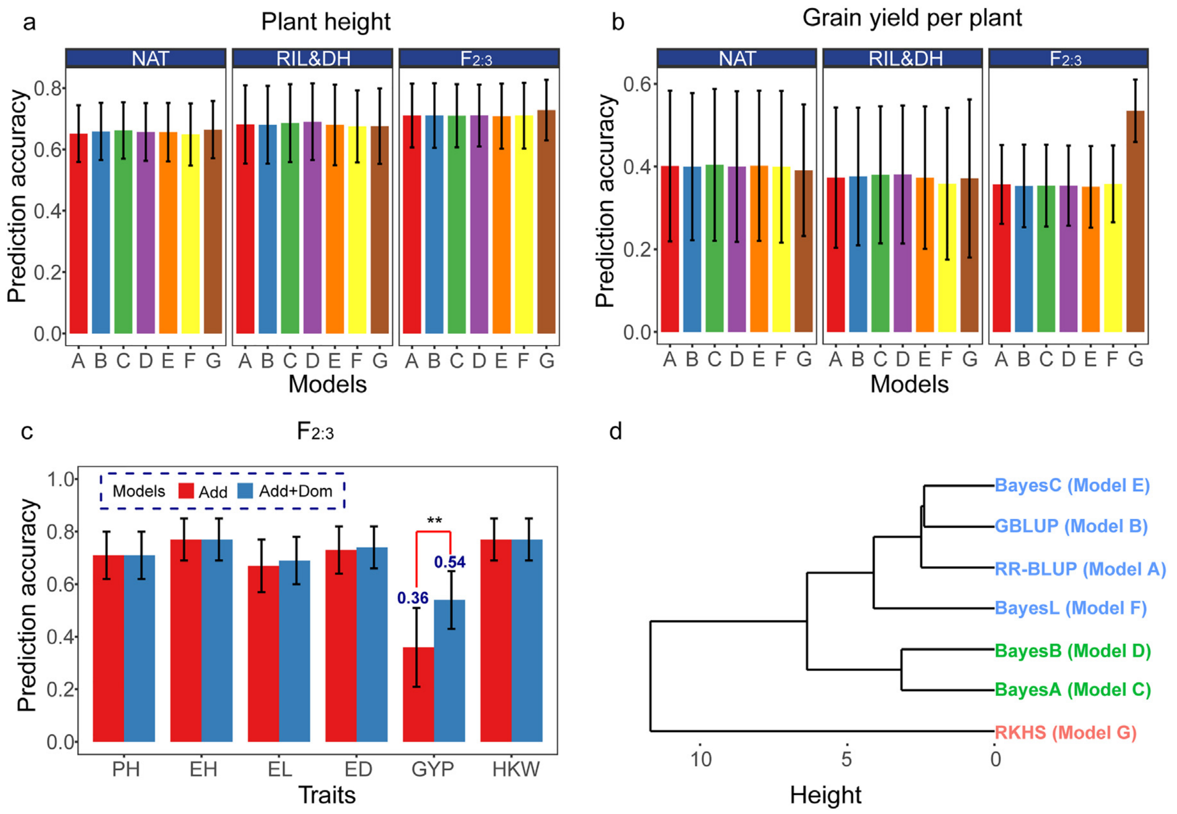


Fig. 4 – Effects of GS statistical models on prediction accuracy and statistical model dendrogram. (a) and (b) Prediction ability of different GS models. NAT, natural population (N = 435; N_m = 37,803); RIL&DH, a population containing RILs and DH lines (N = 291; N_m = 8271); $F_{2:3}$, F_3 families derived from F_2 plants (N = 304; N_m = 8271). (c) Prediction accuracies of GBLUP model with additive (Add model) and additive-dominance (Add + Dom model) effects in the $F_{2:3}$ population (N = 304) with 8271 markers. PH, plant height (cm); EH, ear height (cm); EL, ear length (cm); ED, ear diameter (cm); GYP, grain yield per plant (kg); HKW, hundred-kernel weight (g). $^{**}P < 0.001$. (d) Hierarchical clustering of prediction models; height on the x axis is defined as the distance between clusters; The GBLUP model with 10-fold cross-validation scheme was implemented.

and epistatic effects in order to predict hybrid performance in breeding hybrid crops.

Acknowledgments

The authors thank the anonymous reviewers for valuable comments and suggestions that improved the manuscript, and Dr. Huihui Li, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS) for technical assistance. This research was supported by the National Basic Research Program of China (2014 CB138206), National Key Research and Development Program of China (2016YFD0101803), the National Natural Science Foundation of China-CGIAR International Collaborative Program (31361140364), the Agricultural Science and Technology Innovation Program (ASTIP) of CAAS, and Fundamental Research Funds for Central Non-Profit of Institute of Crop Sciences, CAAS (1610092016124). Research activities of CIMMYT staff were supported by the Bill and Melinda Gates Foundation and the CGIAR Research Program MAIZE.

Authors' contribution

Conceived and designed the experiments: Y. Xu, C. Huang, H. Wang, B.M. Prasanna, and M.S. Olsen. Performed the experiments: X. Liu, H. Wang, Z. Guo, X. Xu, and J. Liu. Analyzed the data: X. Liu and H. Wang. Contributed materials/analysis tools: Y. Xu, W. Li, C. Zou, and S. Wang. Wrote the paper: Y. Xu, X. Liu, B.M. Prasanna, and M.S. Olsen.

REFERENCES

[1] S. Michel, C. Ametz, H. Gungor, D. Epure, H. Grausgruber, F. Löschenberger, H. Buerstmayr, Genomic selection across multiple breeding cycles in applied bread wheat breeding, Theor. Appl. Genet. 129 (2016) 1179–1189.

[2] S.L. Cao, A. Loladze, Y.B. Yuan, Y.S. Wu, A. Zhang, J.F. Chen, G. Huestis, J.S. Cao, V. Chaikam, M. Olsen, B.M. Prasanna, F. San Vicente, X.C. Zhang, Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction, Plant Genome 10 (2017) 1–14.

[3] X.C. Zhang, P. Pérez-Rodríguez, J. Burgueño, M. Olsen, E. Buckler, G. Atlin, B.M. Prasanna, M. Vargas, F. San Vicente, J. Crossa, Rapid cycling genomic selection in a multiparental tropical maize population, G3-Genes Genomes Genet. 7 (2017) 2315–2326.

[4] T.H.E. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics 157 (2001) 1819–1829.

[5] E. Jonas, D.J. de Koning, Does genomic selection have a future in plant breeding? Trends Biotechnol. 31 (2013) 497–504.

[6] E.L. Heffner, M.E. Sorrells, J.L. Jannink, Genomic selection for crop improvement, Crop Sci. 49 (1) (2009) 12.

[7] Y. Xu, Y.L. Lu, C.X. Xie, S.B. Gao, J.M. Wan, B.M. Prasanna, Whole-genome strategies for marker-assisted plant breeding, Mol. Breed. 29 (2012) 833–854.

[8] R. Bernardo, J.M. Yu, Prospects for genomewide selection for quantitative traits in maize, Crop Sci. 47 (2007) 1082–1090.

[9] A. Nakaya, S.N. Isobe, Will genomic selection be a practical method for plant breeding? Ann. Bot. 110 (2012) 1303–1316.

[10] E.L. Heffner, A.J. Lorenz, J.L. Jannink, M.E. Sorrells, Plant breeding with genomic selection: gain per unit time and cost, Crop Sci. 50 (2010) 1681–1690.

[11] Y. Xu, P. Li, C. Zou, Y.L. Lu, C.X. Xie, X.C. Zhang, B.M. Prasanna, M.S. Olsen, Enhancing genetic gain in the era of molecular breeding, J. Exp. Bot. 68 (2017) 2641–2666.

[12] L.R. Schaeffer, Strategy for applying genome-wide selection in dairy cattle, J. Anim. Breed. Genet. 123 (2006) 218–223.

[13] M.M. Farah, A.A. Swan, M.R. Fortes, R. Fonseca, S.S. Moore, M. J. Kelly, Accuracy of genomic selection for age at puberty in a multi-breed population of tropically adapted beef cattle, Anim. Genet. 47 (2016) 3–11.

[14] C.M. Kariuki, E.W. Brascamp, H. Komen, A.K. Kahi, J.A. van Arendonk, Economic evaluation of progeny-testing and genomic selection schemes for small-sized nucleus dairy cattle breeding programs in developing countries, J. Dairy Sci. 100 (2017) 2258–2268.

[15] R.E. Lorenzana, R. Bernardo, Accuracy of genotypic value predictions for marker-based selection in biparental plant populations, Theor. Appl. Genet. 120 (2009) 151–161.

[16] S.Z. Xu, D. Zhu, Q.F. Zhang, Predicting hybrid performance in rice using genomic best linear unbiased prediction, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 12456–12461.

[17] X. Wang, L. Li, Z. Yang, X. Zheng, S. Yu, C. Xu, Z. Hu, Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II, Heredity 118 (2017) 302–310.

[18] G. Charmet, E. Storlie, F.X. Oury, V. Laurent, D. Beghin, L. Chevarin, A. Lapierre, M.R. Perretant, B. Rolland, E. Heumez, L. Duchalais, E. Goudemand, J. Bordes, O. Robert, Genome-wide prediction of three important traits in bread wheat, Mol. Breed. 34 (2014) 1843–1852.

[19] F.M. Bassi, A.R. Bentley, G. Charmet, R. Ortiz, J. Crossa, Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.), Plant Sci. 242 (2016) 23–36.

[20] A.H. Sallam, J.B. Endelman, J.L. Jannink, K.P. Smith, Assessing genomic selection prediction accuracy in a dynamic barley breeding population, Plant Genome 8 (2015) 1–15.

[21] M. Schmidt, S. Kollers, A. Maasberg-Prelle, J. Großer, B. Schinkel, A. Tomerius, A. Graner, V. Korzun, Prediction of malting quality traits in barley based on genome-wide marker data to assess the potential of genomic selection, Theor. Appl. Genet. 129 (2016) 203–213.

[22] J. Crossa, P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett, K. Mathews, Genomic prediction in CIMMYT maize and wheat breeding programs, Heredity 112 (2014) 48–60.

[23] X. Zhang, P. Pérez-Rodríguez, K. Semagn, Y. Beyene, R. Babu, M.A. López-Cruz, F. San Vicente, M. Olsen, E. Buckler, J.L. Jannink, B.M. Prasanna, J. Crossa, Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs, Heredity 114 (2015) 291–299.

[24] Y. Beyene, K. Semagn, S. Mugo, A. Tarekegne, R. Babu, B. Meisel, P. Sehabiague, D. Makumbi, C. Magorokosho, S. Oikeh, J. Gakunga, M. Vargas, M. Olsen, B.M. Prasanna, M. Banziger, J. Crossa, Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress, Crop Sci. 55 (2015) 154–163.

[25] C.K. Wong, R. Bernardo, Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations, Theor. Appl. Genet. 116 (2008) 815–824.

[26] M.F.R. Resende, P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando, J.M. Davis, E.J. Jokela, T.A. Martin, G.F. Peter, M. Kirst, Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.), Genetics 190 (2012) 1503–1510.

[27] F. Isik, J. Bartholomé, A. Farjat, E. Chancerel, A. Raffin, L. Sanchez, C. Plomion, L. Bouffier, Genomic selection in maritime pine, Plant Sci. 242 (2016) 108–119.

[28] N. Heslot, H.P. Yang, M.E. Sorrells, J.L. Jannink, Genomic selection in plant breeding: a comparison of models, Crop Sci. 52 (2012) 146–160.

[29] J.O. Ogutu, T. Schulz-Streeck, H.P. Piepho, Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions, BMC Proc. 6 (2012) S10.

[30] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, M.P.L. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics 193 (2013) 327–345.

[31] P. Juliana, R.P. Singh, P.K. Singh, J. Crossa, J.E. Rutkoski, J.A. Poland, G.C. Bergstrom, M.E. Sorrells, Comparison of models and whole-genome profiling approaches for genomic-enabled prediction of *Septoria tritici* blotch, *Stagonospora nodorum* blotch, and tan spot resistance in wheat, Plant Genome 10 (2017) 1–16.

[32] J.C. Whittaker, R. Thompson, M.C. Denham, Marker-assisted selection using ridge regression, Genet. Res. 75 (2000) 249–252.

[33] L.B. Endelman, Ridge regression and other kernels for genomic selection with R package rrBLUP, Plant Genome 4 (2011) 250–255.

[34] R. Bernardo, Best linear unbiased prediction of maize single-cross performance, Crop Sci. 36 (1996) 50–56.

[35] P.M. VanRaden, Efficient methods to compute genomic predictions, J. Anim. Sci. 91 (2008) 4414–4423.

[36] D. Habier, R.L. Fernando, K. Kizilkaya, D.J. Garrick, Extension of the bayesian alphabet for genomic selection, BMC Bioinf. 12 (2011) 186.

[37] D. Gianola, Priors in whole-genome regression: the Bayesian alphabet returns, Genetics 194 (2013) 573–596.

[38] G. de los Campos, D. Gianola, G.J.M. Rosa, Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation, J. Anim. Sci. 87 (2009) 1883–1887.

[39] G. de los Campos, D. Gianola, G.J. Rosa, K.A. Weigel, J. Crossa, Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods, Genet. Res. 92 (2010) 295–308.

[40] J.M. González-Camacho, G. de los Campos, P. Pérez, D. Gianola, J.E. Cairns, G. Mahuku, R. Babu, J. Cross, Genome-enabled prediction of genetic values using radial basis function neural networks, Theor. Appl. Genet. 125 (2012) 759–771.

[41] J.M. González-Camacho, J. Crossa, P. Pérez-Rodríguez, L. Ornella, D. Gianola, Genome-enabled prediction using probabilistic neural network classifiers, BMC Genomics 17 (2016) 208.

[42] L. Ornella, P. Pérez, E. Tapia, J.M. González-Camacho, J. Burgueño, X. Zhang, S. Singh, F.S. Vicente, D. Bonnett, S.

Dreisigacker, R. Singh, N. Long, J. Crossa, Genomic-enabled prediction with classification algorithms, Heredity 112 (2014) 616–626.

[43] J. Burgueño, G. de los Campos, K. Weigel, J. Crossa, Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers, Crop Sci. 52 (2012) 707–719.

[44] D. Jarquín, J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt, J. Lorgeou, F. Piraux, L. Guerreiro, P. Pérez, M. Calus, J. Burgueño, G. de los Campos, A reaction norm model for genomic selection using high-dimensional genomic and environmental data, Theor. Appl. Genet. 127 (2014) 595–607.

[45] J. Cuevas, J. Crossa, O.A. Montesinos-López, J. Burgueño, P. Pérez-Rodríguez, G. de los Campos, Bayesian genomic prediction with genotype × environment interaction kernel models, G3-Genes Genomes Genet. 7 (2017) 41–53.

[46] J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, J.M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, S. Dreisigacker, R. Singh, X. Zhang, M. Gowda, M. Roorkiwal, J. Rutkoski, R.K. Varshney, Genomic selection in plant breeding: methods, models, and perspectives, Trends Plant Sci. 22 (2017) 961–975.

[47] T. Albrecht, V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, C.C. Schön, Genome-based prediction of testcross values in maize, Theor. Appl. Genet. 123 (2011) 339–350.

[48] M. Pszczola, T. Strabel, H.A. Mulder, M.P. Calus, Reliability of direct genomic values for animals with different relationships within and to the reference population, J. Dairy Sci. 95 (2012) 389–400.

[49] J.M. Elsen, Approximated prediction of genomic selection accuracy when reference and candidate populations are related, Genet. Sel. Evol. 48 (2016) 1–19.

[50] E.L. Heffner, J.L. Jannink, H. Iwata, E. Souza, M.E. Sorrells, Genomic selection accuracy for grain quality traits in biparental wheat populations, Crop Sci. 51 (2011) 2597–2606.

[51] E. Combs, R. Bernardo, Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers, Plant Genome 6 (2013) 1–7.

[52] Y.S. Zhao, M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, J.C. Reif, Accuracy of genomic selection in European maize elite breeding populations, Theor. Appl. Genet. 124 (2012) 769–776.

[53] Z.G. Guo, D.M. Tucker, C.J. Basten, H. Gandhi, E. Ersoz, B.H. Guo, Z.Y. Xu, D.L. Wang, G. Gay, The impact of population structure on genomic prediction in stratified populations, Theor. Appl. Genet. 127 (2014) 749–762.

[54] J. Isidro, Training set optimization under population structure in genomic selection, Theor. Appl. Genet. 128 (2015) 145–158.

[55] J. Spindel, H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redoña, G. Atlin, J.L. Jannink, S.R. McCouch, Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines, PLoS Genet. 11 (2015), e1004982.

[56] C. Xu, Y.H. Ren, Y.Q. Jian, Z.F. Guo, Y. Zhang, C.X. Xie, J.J. Fu, H.W. Wang, G.Y. Wang, Y. Xu, P. Li, C. Zou, Development of a maize 55 K SNP array with improved genome coverage for molecular breeding, Mol. Breed. 37 (2017) 20.

[57] A. Raj, M. Stephens, J.K. Pritchard, fastSTRUCTURE: variational inference of population structure in large SNP data sets, Genetics 197 (2014) 573–589.

[58] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, 1000 genomes project analysis group. The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158.

[59] P. Pérez, G. de los Campos, Genome-wide regression & prediction with the BGLR statistical package, Genetics 198 (2014) 483–495.

[60] G.S. Su, O.F. Christensen, T. Ostersen, M. Henryon, M.S. Lund, Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers, PLoS One 7 (2012), e45293.

[61] G. Covarrubias-Pazaran, Genome-assisted prediction of quantitative traits using the R package *sommer*, PLoS One 11 (2016), e0156744.

[62] A. Gianola, R.L. Fernando, A. Stella, Genomic-assisted prediction of genetic value with semiparametric procedures, Genetics 173 (2006) 1761–1776.

[63] D. Gianola, J.B.C.H.M. van Kaam, Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits, Genetics 178 (2008) 2289–2303.

[64] Y. Jiang, J.C. Reif, Modeling epistasis in genomic selection, Genetics 201 (2015) 759–768.

[65] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna2017.

[66] D. Habier, R.L. Fernando, J.C.M. Dekkers, The impact of genetic relationship information on genome-assisted breeding values, Genetics 177 (2007) 2389–2397.

[67] E.L. Heffner, J.L. Jannink, M.E. Sorrells, Genomic selection accuracy using multifamily prediction models in a wheat breeding program, Plant Genome 4 (2011) 65–75.

[68] H. Wang, C. Xu, X.G. Liu, Z.F. Guo, X.J. Xu, S.H. Wang, C.X. Xie, W.X. Li, C. Zou, Y. Xu, Development of a multiple-hybrid population for genome-wide association studies: theoretical consideration and genetic mapping of flowering traits in maize, Sci. Rep. 7 (2017) 40239.

[69] J.S.C. Smith, T. Hussain, E.S. Jones, G. Graham, D. Podlich, S. Wall, M. Williams, Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops, Mol. Breed. 22 (2008) 51–59.

[70] Z.G. Guo, D.M. Tucker, J.W. Lu, V. Kishore, G. Gay, Evaluation of genome-wide selection efficiency in maize nested association mapping populations, Theor. Appl. Genet. 124 (2012) 261–275.

[71] K.T. Muleta, P. Bulli, Z.W. Zhang, X.M. Chen, M. Pumphrey, Unlocking diversity in germplasm collections via genomic selection: a case study based on quantitative adult plant resistance to stripe rust in spring wheat, Plant Genome 10 (2017) 1–15.

[72] F. Technow, A. Bürger, A.E. Melchinger, Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups, G3-Genes Genomes Genet. 3 (2013) 197–203.

[73] T. Schulz-Streeck, J.O. Ogutu, Z. Karaman, C. Knaak, H.P. Piepho, Genomic selection using multiple populations, Crop Sci. 52 (2012) 2453–2461.

[74] A. Zhang, H.W. Wang, Y. Beyene, K. Semagn, Y.B. Liu, S.L. Cao, Z.H. Cui, Y.Y. Ruan, J. Burgueño, F. San Vicente, M. Olsen, B.M. Prasanna, H.Q. Yu, X.C. Zhang, Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations, Front. Plant Sci. 8 (2017) 1916.

[75] A. Lehermeier, N. Krämer, E. Bauer, C. Bauland, C. Camisan, L. Campo, P. Flament, A.E. Melchinger, M. Menz, N. Meyer, L. Moreau, J. Moreno-González, M. Ouzunova, H. Pausch, N. Ranc, W. Schipprack, M. Schönleben, H. Walter, A. Charcosset, C.C. Schön, Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction, Genetics 198 (2014) 3–16.

[76] R. Howard, A.L. Carriquiry, W.D. Beavis, Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures, G3-Genes Genomes Genet. 4 (2014) 1027–1046.