The Crop Journal 11 (2023) 490–498

Contents lists available at ScienceDirect

The Crop Journal

journal homepage: www.keaipublishing.com/en/journals/the-crop-journal/

# Comparison of sequencing-based and array-based genotyping platforms for genomic prediction of maize hybrid performance

Guangning Yu [a], Yanru Cui [b], Yuxin Jiao [a], Kai Zhou [a], Xin Wang [a], Wenyan Yang [a], Yiyi Xu [a], Kun Yang [a], Xuecai Zhang [c], Pengcheng Li [a], Zefeng Yang [a], Yang Xu [a,*], Chenwu Xu [a,*]

[a] Key Laboratory of Plant Functional Genomics of the Ministry of Education/Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, College of Agriculture, Yangzhou University, Yangzhou 225009, Jiangsu, China
[b] State Key Laboratory of North China Crop Improvement and Regulation, Hebei Agricultural University, Baoding 071001, Hebei, China
[c] International Maize and Wheat Improvement Center (CIMMYT), 06600 Mexico D.F, Mexico

ARTICLE INFO

*Article history:*
Received 19 May 2022
Revised 27 June 2022
Accepted 17 September 2022
Available online 28 September 2022

*Keywords:*
Genomic selection
Maize
GBS
SNP array
Marker density

ABSTRACT

Genomic selection (GS) is a powerful tool for improving genetic gain in maize breeding. However, its routine application in large-scale breeding pipelines is limited by the high cost of genotyping platforms. Although sequencing-based and array-based genotyping platforms have been used for GS, few studies have compared prediction performance among platforms. In this study, we evaluated the predictabilities of four agronomic traits in 305 maize hybrids derived from 149 parental lines subjected to genotyping by sequencing (GBS), a 40K SNP array, and target sequence capture (TSC) using eight GS models. The GBS marker dataset yielded the highest predictabilities for all traits, followed by TSC and SNP array datasets. We investigated the effect of marker density and statistical models on predictability among genotyping platforms and found that 1K SNPs were sufficient to achieve comparable predictabilities to 10K and all SNPs, and BayesB, GBLUP, and RKHS performed well, while XGBoost performed poorly in most cases. We also selected significant SNP subsets using genome-wide association study (GWAS) analyses in three panels to predict hybrid performance. GWAS facilitated selecting effective SNP subsets for GS and thus reduced genotyping cost, but depended heavily on the GWAS panel. We conclude that there is still room for optimization of the existing SNP array, and using genotyping by target sequencing (GBTS) techniques to integrate a few functional markers identified by GWAS into the 1K SNP array holds great promise of being an effective strategy for developing desirable GS breeding arrays.



## 1. Introduction

Genomic selection (GS) holds great promise for accelerating crop breeding via early selection before phenotypes are measured [1]. Maize is an ideal crop for GS breeding technology. In hybrid breeding, GS can predict all potential hybrid combinations of a given set of genotyped parents using field evaluation of only a fraction of crosses, considerably reducing the cost of field trials [2]. The utility of GS in maize has been evidenced by simulations and empirical studies [3,4]. Moderate to high predictabilities for agronomic traits such as ear weight, flowering date, and plant height, have been reported in breeding populations [5,6]. Although the GS technology has been implemented in breeding pipelines by international research institutes and multinational breeding companies, the high cost of genotyping is the main limiting factor for large-scale breeding application of GS in developing countries and small and medium-sized enterprises [7]. A high-quality and affordable genotyping platform is needed for GS applications. Sequencing-based and array-based genotyping platforms have been developed and used for genomics-assisted breeding [8–10].

Genotyping by sequencing (GBS), one of the high-throughput genotyping technologies, uses restriction enzymes to reduce genome complexity and generates high-density genome-wide markers for large populations by multiplexing samples with unique DNA barcodes [11]. GBS has been used to sequence more than 17,000 maize materials for genetic research. Its flexibility and high throughput make it an effective genotyping platform for GS. Crossa et al. [12] demonstrated the suitability of GBS for developing GS models in maize breeding populations. Wang et al. [13] genotyped 2240 individuals from eight maize populations developed at CIMMYT using GBS and then evaluated the prediction accuracy for

* Corresponding authors.
*E-mail addresses:* xuyang_89@126.com (Y. Xu), cwxu@yzu.edu.cn (C. Xu).

https://doi.org/10.1016/j.cj.2022.09.004

grain yield using unimputed and imputed GBS data. Guo et al. [14] performed GS for Zn concentration of maize kernels in double-haploid (DH) populations and showed that the predictability assessed from the GBS markers was higher than that from repeat amplification sequencing (rAmpSeq) markers.

Despite the wide application of GBS, there are still some flaws in the form of low sequencing coverage and complex statistical analysis [15]. In contrast to GBS, SNP arrays involve fixed locus sets and simplified data analysis. In maize, several array-based genotyping platforms have been built based on Illumina and Affymetrix systems. Initial representative arrays such as the MaizeSNP50 BeadChip [16] and MaizeSNP600K [17] were developed using early sequencing data mostly from European and American materials and were more suited to genotyping temperate than tropical germplasm. Tian et al. [18] developed a high-quality Maize6H-60K array using whole-genome sequencing data for 388 globally collected inbred lines and showed that this array was suitable for germplasm resource analysis and variety identification. Recently [7], a high-resolution multiple-SNP maize array was developed using a genotyping by target sequencing (GBTS) system coupled with liquid chip technology. Compared to solid chips, GBTS is more flexible and efficient. With the GBTS system, multiple panels with various marker densities can be obtained from one 40K SNP mother panel by sequencing to different depths. Arrays with varying numbers of SNPs can be designed for diverse research objectives. Although several maize arrays have been successfully applied to predicting agronomic traits and disease resistance, GS application of the 40K SNP array has not yet been reported.

In addition to using whole-genome markers genotyped by GBS or array platforms to predict the phenotypic traits, Zhang et al. [19,20] proposed a novel strategy that uses genes significantly associated with target traits to implement genomic prediction (GP) and demonstrated its utility and efficiency in maize and cotton breeding. They found that prediction of maize grain yield from associated genes was more accurate than that from randomly selected genes. However, GP using gene-based SNPs has not yet been used to predict maize hybrid performance.

This study employed a maize training population consisting of 305 hybrids derived from 149 inbred lines. All parental lines were genotyped with three technologies. Four agronomic traits of the hybrids and their parents were evaluated. The main objectives were to (1) compare the prediction performance of several genotyping platforms for predicting agronomic traits in maize hybrids, (2) determine the appropriate marker density and prediction models under each genotyping platform, and (3) identify an effective strategy for selecting optimum SNP subsets to further reduce genotyping cost in maize GS breeding.

## 2. Materials and methods

### 2.1. Plant materials and field trials

A set of 305 hybrids were generated based on a spatial partial diallel crossing experiment involving 149 inbred lines, which were a subset of a previously described [21,22] maize association panel of 346 lines. The 149 lines were collected from diverse geographical locations, including five heterotic groups of Reid, Lancaster, Tangsipingtou, Lvdahonggu, and mixed group. All maize materials were planted in a randomized block design with two repetitions in experimental fields at Yangzhou and Tai'an during the 2017 and 2018 maize growing seasons. For each maize material in each replication, 13 plants were grown in a row of 3.0 m with 0.5 m space between rows. For each inbred line and hybrid, five uniform plants were selected to evaluate four traits: plant height (PH), ear height (EH), ear weight (EW), and grain yield (GY). For each inbred line and hybrid, the best linear unbiased predictor (BLUP) values of all traits across two environments were calculated using the linear mixed model in the R package lme4 [23].

### 2.2. Genotyping and genotypic data analysis

Fresh young leaves were collected from a natural population of 346 inbred lines at the vegetative growth stage, and a modified CTAB method [24] was used to extract genomic DNA. The population was genotyped by GBS. The ApeK1 restriction enzyme was used for library preparation, and GBS was performed on an Illumina platform by Novogene Co., Ltd. (Beijing, China). SNP calling was performed using the TASSEL GBS discovery pipeline with B73 as the reference genome [25]. SNP filtering was performed by eliminating SNPs with a missing rate above 0.1 and minor-allele frequency (MAF) below 0.05. A SNP density plot was drawn with the R package rMVP [26]. After quality control, 102,654 high-quality markers remained. The SNP data of the 149 inbred lines used in this study were extracted from the full dataset. These SNPs were relatively evenly distributed on the 10 chromosomes (Fig. 1A).

The 149 inbred lines were also genotyped with the 40K maize liquid array based on the GBTS platform developed by Molbreeding Biotechnology Company (Shijiazhuang, Hebei, China), by use of which 40K, 10K, and 1K SNPs could be obtained simultaneously. The 1K SNP and 10K SNP panels were generated from the 40K SNP panel by sequencing to different depths, and the 1K SNP panel was a subset of the 10K SNP panel. Elimination of SNPs with a missing rate above 0.1 and MAF below 0.05 left respectively 41,855, 11,255, and 1319 SNPs from the 40K, 10K, and 1K SNP arrays. The 41,855 SNPs were distributed throughout the genome, with slightly higher density at telomeres (Fig. 1B).

A set of 163 candidate genes associated with maize yield and plant-type traits were screened, including known functional genes controlling maize yield and plant-type traits, and maize homologs of cloned functional genes in other species. The selected genes in a maize natural population of 346 inbred lines were resequenced using target sequence capture (TSC) sequencing technology in the NimbleGen platform by Beijing Genomics Institute (BGI, Shenzhen, China). The corresponding genomic sequences and positions of the selected genes in the B73 reference genome were used as references for TSC. A total of 582 Gb of resequencing data were obtained. The total length of the target region was 1193.2 kb, the average length of each region was 7.32 kb, and the average sequencing depth of the target region was greater than 100x. A total of 29,510 SNPs were obtained, with an average of 181 SNPs per gene. The genomic data of the subset of 149 inbred lines were extracted from the initial data. After filtering SNPs by the criteria MAF greater than 0.05 and missing rate less than 0.1, 16,492 markers remained for subsequent analysis. The markers were unevenly distributed on chromosomes (Fig. 1C).

### 2.3. Genotype coding of inbred lines and hybrids

Let $M = \{M_{jk}\}$ and $F = \{F_{jk}\}$ be $n \times m$ genotype matrices for the maternal and paternal parents of the corresponding hybrids. The genotype code of marker $k$ ($k$ = 1, 2, …, $m$) for individual $j$ ($j$ = 1, 2, …, $n$) was defined as $M_{jk} = F_{jk} = 1$ for the major allele homozygote $A_1A_1$, $M_{jk} = F_{jk} = 0$ for the heterozygote $A_1A_2$, and $M_{jk} = F_{jk} = -1$ for the minor allele homozygote $A_2A_2$. The genotype code of the hybrid was defined as $Z_{jk} = \frac{1}{2}(M_{jk} + F_{jk})$. For example, if the mating type for the marker was $A_2A_2 \times A_1A_1$, the hybrid was coded as $(-1 + 1)/2 = 0$.

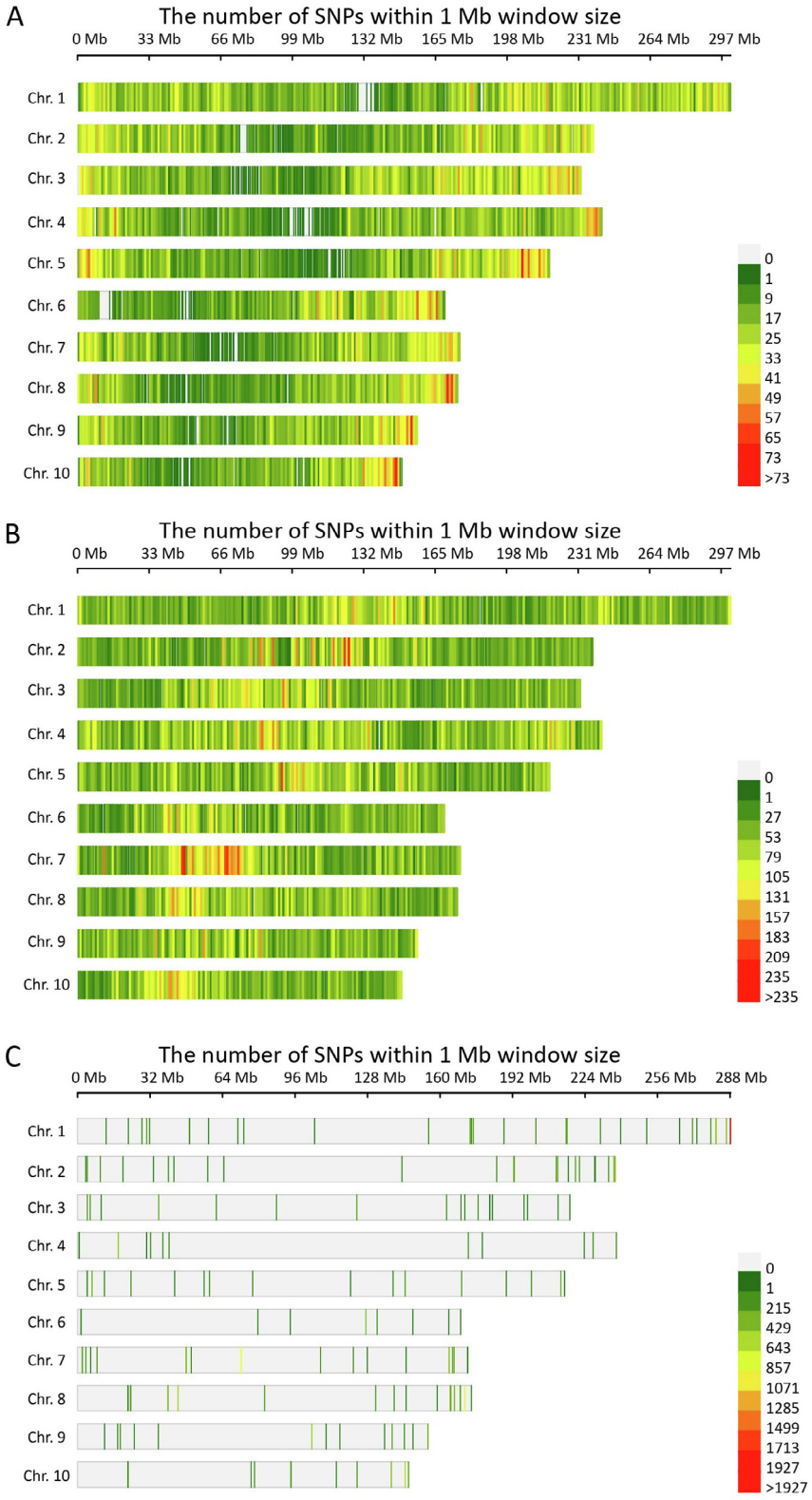


**Fig. 1.** Distribution of SNPs on 10 chromosomes under three genotyping platforms of (A) Genotyping by sequencing (GBS), (B) 40K SNP array, and (C) target sequence capture (TSC). Marker density is indicated by different bar colors, and each bar represents 1 Mb window size.

### *2.4. Models of genomic prediction*

GP of plant height, ear height, ear weight, and grain yield was performed with the genomic data obtained from the GBS, 40K maize array, and TSC platforms. Eight commonly used statistical models comprising five parametric and three nonparametric models were fitted. The parametric models were genomic best linear unbiased prediction (GBLUP), BayesB, least absolute shrinkage and selection operator (LASSO), Elastic Net (EN), and partial least squares (PLS). The three nonparametric models were reproducing

kernel Hilbert space regression (RKHS), extreme gradient boosting (XGBoost), and random forest (RF). GBLUP was fitted with our own R package, predhy. LASSO and EN were fitted with the R package glmnet [27], and the shrinkage parameter lambda was determined by tenfold cross-validation. PLS was implemented in the R package pls [28], and the optimum number of latent components minimizing the root mean squared error of prediction (RMSEP) was chosen by tenfold cross-validation. BayesB, and RKHS were implemented using the R package BGLR [29], and all parameters set to their default values. XGBoost was implemented using the R package xgboost [30]. The maximum number of boosting iterations was set to 1000, the booster parameter eta was set to 0.07, and other parameters were set to default. RF was implemented in the R package randomForest [31]. The number of decision trees was set to 500 and other parameters were set to their defaults.

Trait predictability was evaluated by fivefold cross-validation, in which the sample was randomly divided into five equal parts and each part was predicted once using parameters estimated from the remaining four parts. Predictability was defined as the correlation coefficient between the predicted and observed phenotypic values. To reduce the variation caused by sample partitioning, the fivefold cross-validations were repeated 20 times for each analysis and the average predictabilities were reported.

### 2.5. Selection of SNP subsets for prediction

A genome-wide association study (GWAS) was performed to identify the SNP subset in three maize panels. Panel I was the maize natural population consisting of 346 inbred lines, panel II was composed of the 149 inbred lines used for construction of the hybrid training population, and panel III comprised the 197 remaining lines not involved in hybrid mating. With high-quality SNPs obtained from GBS sequencing, GWAS was performed using multi-locus methods, including mrMLM, FASTmrMLM, pLARmEB, and ISIS EM–BLASSO implemented in the R package mrMLM [32]. All parameters were set at their default values. Significantly associated SNPs were identified by the threshold of LOD score $\geq 3$. The chosen SNP subsets from three panels were then separately used to evaluate the predictabilities of four traits in the maize hybrid population using the GBLUP method with 20 repeated fivefold cross-validations.

## 3. Results

### 3.1. Evaluation of predictabilities in maize hybrid population

In a maize hybrid population consisting of 305 hybrids, the predictabilities of PH, EH, EW, and GY were evaluated using respectively 102,654, 41,855, and 16,492 filtered SNPs obtained from the GBS, 40K array, and TSC platforms. The average predictabilities for the four traits drawn from 20 repeated fivefold cross-validation with eight prediction methods including GBLUP, BayesB, LASSO, EN, PLS, RKHS, XGBoost, and RF are listed in Table 1. The predictabilities of all traits were moderate to high and varied across genotyping platforms and prediction methods. Among the four, PH showed the highest predictability, followed by EH and EW, with GY being the least predictable trait. For genotyping platforms, the GBS strategy performed the best, and the TSC platform outperformed the 40K maize array. For prediction models, the largest differences in predictabilities among the eight models varied from 0.041 to 0.103 for the identical trait and SNP dataset. By comparison, the differences in predictability among the models for the 40K array dataset were larger than those for the GBS and TSC marker datasets. Across all traits and genotyping platforms, the highest average predictability was 0.589 from BayesB, and the lowest was 0.527 from XGBoost.

### 3.2. Comparison of predictabilities of different genotyping platforms and marker density

To further investigate the influence of genotyping platform on GP, we evaluated the predictability based on the equal number of SNPs under three genotyping platforms. After SNP filtering, respectively 1319 and 11,255 SNPs remained for the 1K and 10K SNP subsets. To maintain consistency in marker density across genotyping platforms, we selected the same number of SNPs (respectively 11,255 and 1319) from the GBS and TSC datasets as for the 1K and 10K subsets. For the maize array, the 10K and 1K SNP subsets were fixed, and for GBS and TSC platforms, the SNP subsets were randomly resampled 20 times. The predictabilities for the four traits from 20 repeated fivefold cross-validation using eight GS models for each marker subset are depicted in Fig. 2. The predictabilities varied greatly across genotyping platforms and prediction methods but relatively little across marker densities.

The GBS and TSC platforms yielded significantly higher predictabilities than SNP array for all traits whether 1K, 10K or all SNPs were used (Fig. 3A). The average predictabilities of PH, EH, EW, and GY evaluated from the 1K GBS marker subset were respectively 0.075, 0.095, 0.078, and 0.123 higher than those estimated from the 1K SNP array, and those evaluated from the 10K GBS marker subset were 0.071, 0.073, 0.093 and 0.121 higher than those estimated from the 10K SNP array. Although using all marker dataset yielded slightly better prediction performance, there were no significant differences between the 1K and 10K SNP subsets or between the 10K and all SNPs irrespective of genotyping platform or trait (Fig. 3B). The differences in predictabilities of PH, EH, EW, and GY between 1K and all SNPs were less than 0.02 for all genotyping platforms.

**Table 1**
Average predictabilities of four traits drawn from fivefold cross-validation repeated 20 times under three genotyping platforms using eight statistical models.

| Trait | Genotypingplatform | GBLUP | BayesB | LASSO | EN | PLS | RKHS | RF | XGBoost | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| PH | GBS | 0.7285 | 0.7291 | 0.6923 | 0.6944 | 0.7185 | 0.7337 | 0.7036 | 0.6977 | 0.7122 |
| | TSC | 0.7062 | 0.7064 | 0.6867 | 0.6896 | 0.6824 | 0.7149 | 0.6659 | 0.6470 | 0.6874 |
| | Array | 0.6665 | 0.6679 | 0.6598 | 0.6629 | 0.6358 | 0.6573 | 0.6043 | 0.5654 | 0.6400 |
| EH | GBS | 0.6550 | 0.6541 | 0.5933 | 0.5892 | 0.6446 | 0.6499 | 0.6245 | 0.6150 | 0.6282 |
| | TSC | 0.6305 | 0.6293 | 0.6330 | 0.6332 | 0.6025 | 0.6396 | 0.6120 | 0.5953 | 0.6219 |
| | Array | 0.5816 | 0.5863 | 0.5626 | 0.5648 | 0.5507 | 0.5630 | 0.5387 | 0.5128 | 0.5576 |
| EW | GBS | 0.6140 | 0.6165 | 0.5878 | 0.5909 | 0.5948 | 0.6168 | 0.5756 | 0.5684 | 0.5956 |
| | TSC | 0.5947 | 0.5966 | 0.5652 | 0.5641 | 0.5710 | 0.5898 | 0.5489 | 0.5387 | 0.5711 |
| | Array | 0.5290 | 0.5403 | 0.5047 | 0.5046 | 0.5215 | 0.5247 | 0.4841 | 0.4658 | 0.5094 |
| GY | GBS | 0.4710 | 0.4778 | 0.4286 | 0.4346 | 0.4580 | 0.4825 | 0.4717 | 0.4551 | 0.4599 |
| | TSC | 0.4642 | 0.4700 | 0.3956 | 0.4068 | 0.4258 | 0.4649 | 0.4166 | 0.3740 | 0.4272 |
| | Array | 0.3511 | 0.3885 | 0.3343 | 0.3331 | 0.3487 | 0.3470 | 0.3352 | 0.2859 | 0.3405 |

PH, plant height; EH, ear height; EW, ear weight; GY, grain yield; GBLUP, genomic best linear unbiased prediction; LASSO, least absolute shrinkage and selection operator; EN, elastic net; PLS, partial least squares; RKHS, reproducing kernel Hilbert space; XGBoost, extreme gradient boosting; RF, random forest.

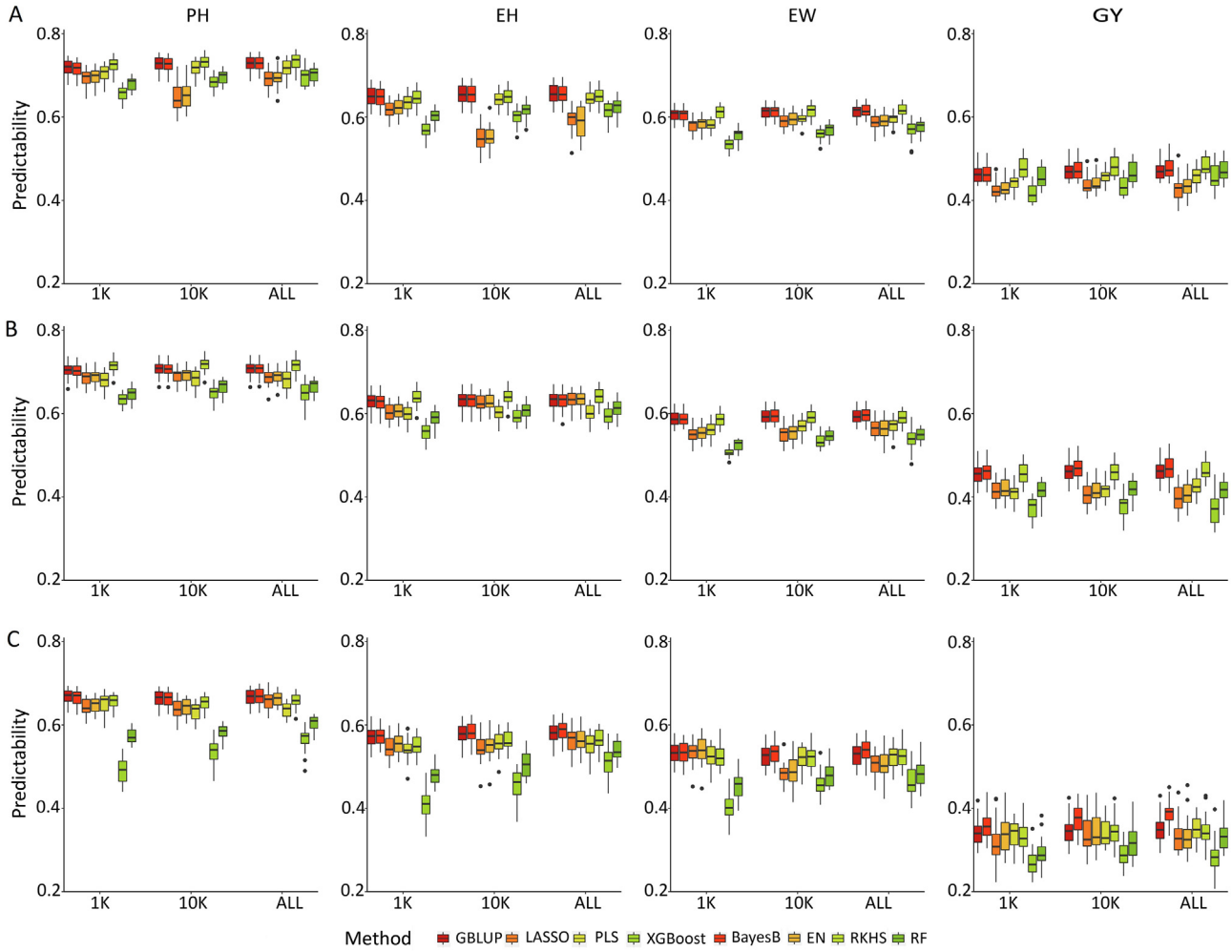


**Fig. 2.** The predictabilities of four traits from eight statistical models with three SNP datasets under three genotyping platforms: (A) GBS, (B) TSC, and (C) SNP array. The four traits are PH, EH, EW and GY, the eight models are GBLUP, BayesB, LASSO, EN, PLS, RKHS, XGBoost and RF, and the three SNP datasets include ~1K (1319), ~10K (11,255) and all SNPs. PH, plant height; EH, ear height; EW, ear weight; GY, grain yield; GBLUP, genomic best linear unbiased prediction; LASSO, least absolute shrinkage and selection operator; EN, elastic net; PLS, partial least squares; RKHS, reproducing kernel Hilbert space; XGBoost, extreme gradient boosting; RF, random forest.

### 3.3. Investigation of appropriate marker density and prediction model

The above study revealed that high marker density might not be essential for GP in the maize hybrid population. To further investigate the appropriate marker density for GP, we selected seven SNP subsets from the GBS marker dataset with the number of markers varying from ~ 0.1K to all markers (100, 300, 500, 700, 1319, 11,255, 102,654). Twenty selections were made randomly for each SNP subset, and for each marker subset, the fivefold cross-validations were repeated 20 times to obtain average predictabilities using the GBLUP method. Fig. 4 presents mean predictabilities from the randomly sampled SNP subsets and their variation measured as standard deviations. The predictabilities of the four traits increased rapidly when the marker number increased from 0.1K to 0.5K and thereafter only slightly as the marker density kept increasing. The fewer the markers, the larger was the variation in predictability. The differences in predictabilities between the 1K and 0.5K marker subsets are 0.016, 0.016, 0.009, and 0.022 for predicting PH, EH, EW, and GY, respectively, whereas the corresponding differences between using all markers and 1K marker subsets were only 0.009, 0.006, 0.008 and 0.007.

To identify the appropriate prediction model among multiple genotyping strategies, we compared the predictabilities of eight models across all traits and marker density for each genotyping platform. GBLUP, BayesB, and RKHS performed best with the GBS and TSC marker dataset, BayesB performed best with the SNP array dataset, and XGBoost performed the worst in all cases (Fig. 5). GBLUP, BayesB, and RKHS yielded the highest predictabilities for PH, EH, and EW, BayesB yielded the highest predictability for GY, and XGBoost yielded the lowest predictabilities for all traits (Fig. S1). The poor prediction performance of XGBoost may have been due to the difficulty of tuning several hyperparameters. We also compared the average computing time of the models under each genotyping platform for evaluating prediction performance. All analyses were performed in a CentOS 6.0 Linux server with 2.40 GHz Intel(R) Xeon(R) CPU E5-2680 v4. Among the platforms, 1K and 10K SNP arrays were the most time-saving platforms, followed by TSC, and the GBS platform was the most time-consuming. Among the eight methods, RF consumed the maximum computing time, followed by XGBoost, BayesB, RKHS, and GBLUP with the lowest computing time (Table S1).

### 3.4. GP with significantly associated SNPs

The above result showed that the same number of marker subsets from different genotyping platforms led to large differences in predictabilities. Accordingly, identifying an optimum marker subset is potentially useful for reducing the genotyping cost. We

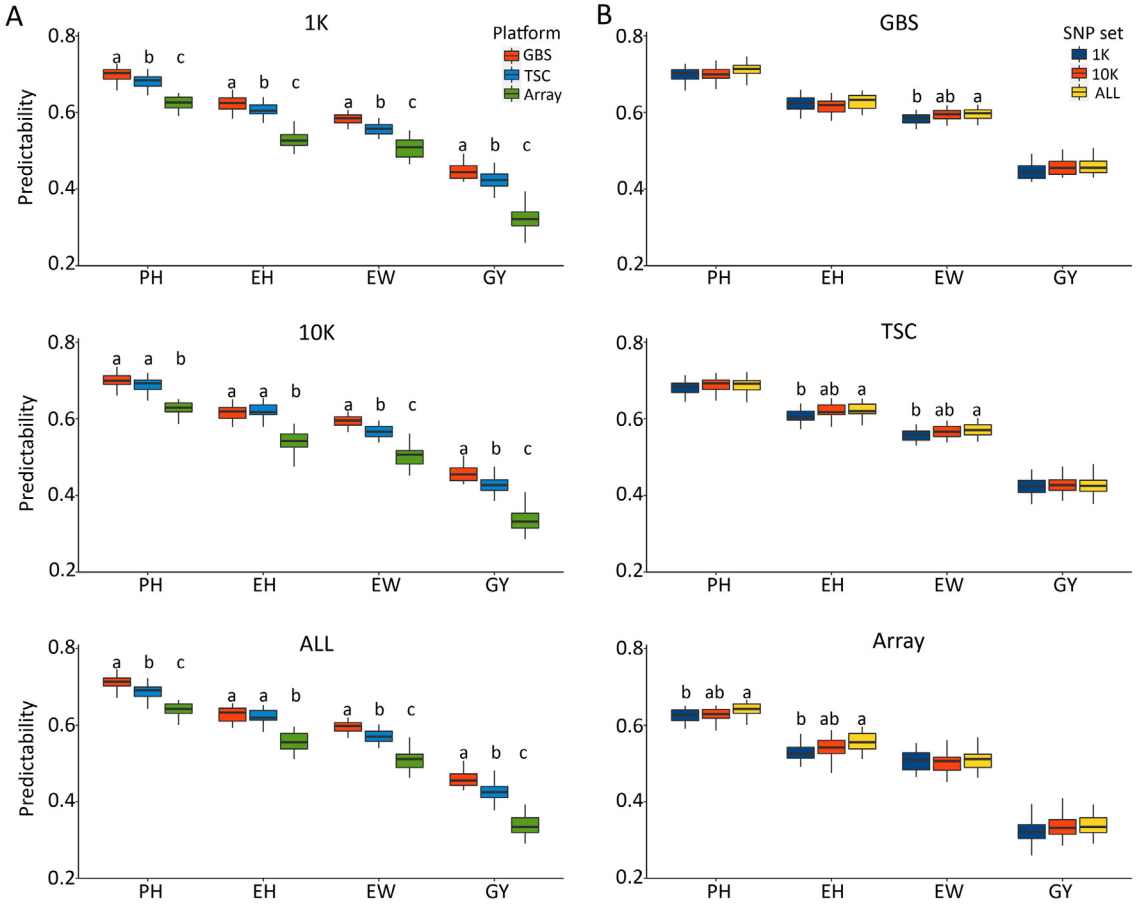


**Fig. 3.** Multiple comparisons of mean predictabilities illustrated by box plots. (A) Comparison of mean predictabilities of three genotyping platforms across eight statistical models for four traits under three SNP sets. (B) Comparison of mean predictabilities of three SNP sets across eight statistical models for four traits under three genotyping platforms. Different lowercase letters above the box plot indicate significant differences at the 0.05 probability level, while no letters indicate no significant differences. PH, plant height; EH, ear height; EW, ear weight; GY, grain yield.

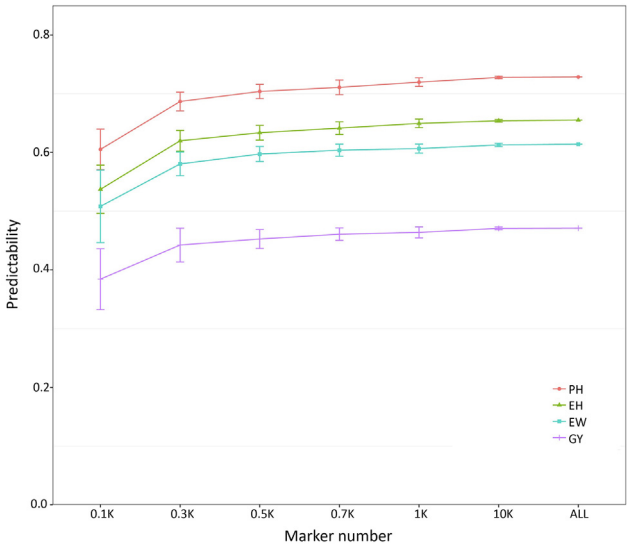


**Fig. 4.** Effect of marker number on the predictability. Seven SNP datasets were selected with marker number varying from ~0.1K to All SNPs (100, 300, 500, 700, 1319, 11,255, 102,654). Twenty random selections were made for each SNP subset, and fivefold cross-validations were repeated 20 times for each selection. Lines indicate the mean predictability from each SNP set and error bars indicate standard deviation. PH, plant height; EH, ear height; EW, ear weight; GY, grain yield.

selected significant SNP subsets using GWAS in three panels to perform GP. A total of 193, 124, and 130 significant SNPs were identified for four traits by all GWAS methods in panels I, II, and III, respectively (Table S2). The predictabilities of PH, EH, EW, and GY from the three selected SNP subsets are listed in Table 2. The corresponding predictabilities obtained from 0.3K and all GBS markers are also listed as references. SNP subsets I and II showed much higher predictabilities than SNP subset III for all traits. Comparison with the references showed that the predictabilities from SNP subsets I and II were higher than those from the 0.3K random selected SNP subset and almost as high as those from all SNPs, whereas the predictabilities from SNP subset III were lower than those from the 0.3K random selected SNP subset.

## 4. Discussion

In this study, we compared three genotyping strategies for predicting maize hybrids. In terms of prediction performance, the predictabilities obtained from the GBS and TSC marker datasets were higher than those from the SNP array dataset. Compared to the SNP array, GBS and TSC can reveal novel or population-specific polymorphisms, which may yield more accurate predictions. However, the costs of genotyping cannot be ignored in practical breeding. The genotyping cost of GBS was almost $35 per sample at the 96- plex level, while the cost per sample can be as low as $14, $10, and $5 per sample for 40K, 10K, and 1K SNP arrays with the

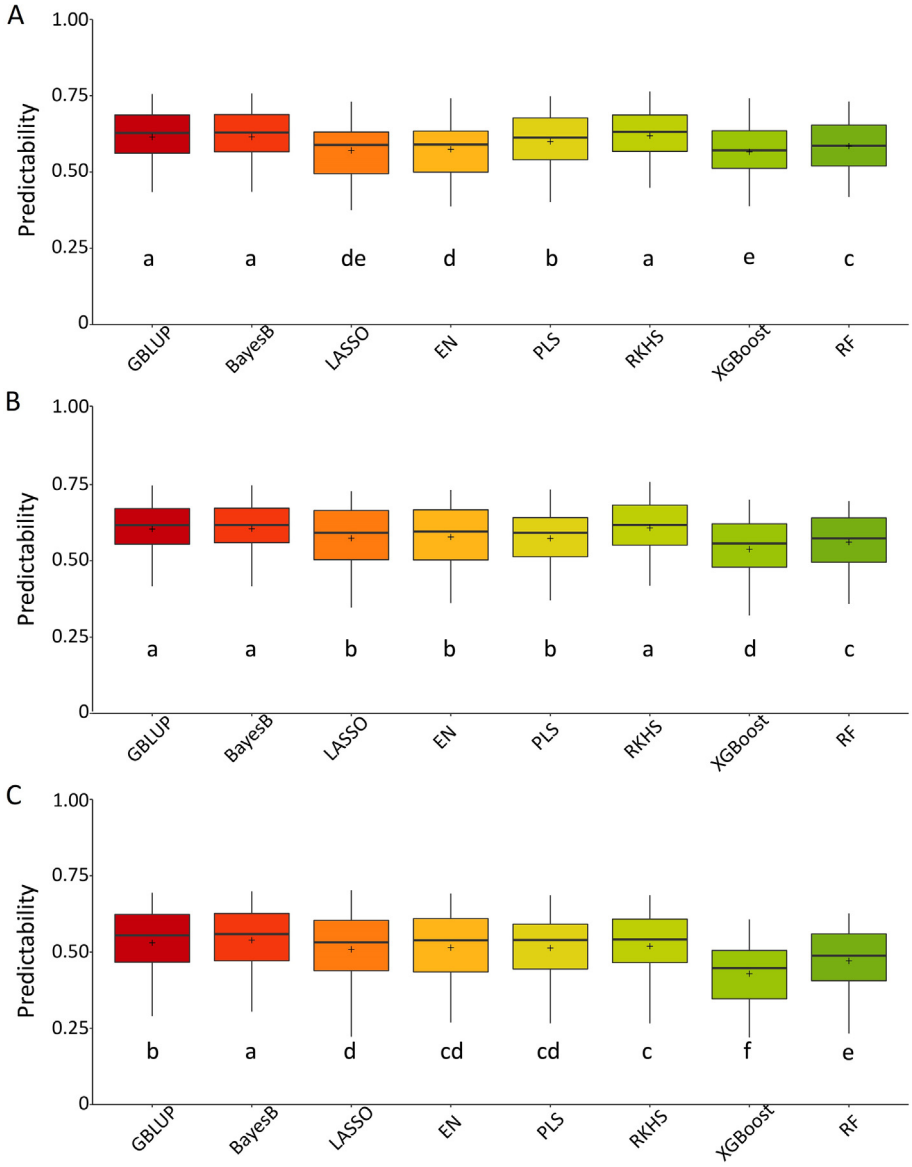


**Fig. 5.** Multiple comparisons of mean predictabilities of eight statistical models across four traits and three SNP sets under three genotyping platforms of (A) GBS, (B) TSC, and (C) SNP array. Different lower-case letters below the box plot indicate significant differences at the 0.05 probability level. GBLUP, genomic best linear unbiased prediction; BayesB, Bayesian B; LASSO, least absolute shrinkage and selection operator; EN, elastic net; PLS, partial least squares; RKHS, reproducing kernel Hilbert space; XGBoost, extreme gradient boosting; RF, random forest.

**Table 2**
The predictabilities of four traits from three SNP subsets identified in three GWAS panels.

| Marker | Trait | | | |
|---|---|---|---|---|
| | PH | EH | EW | GY |
| 0.3K SNPs | 0.6867 | 0.6198 | 0.5804 | 0.4422 |
| All SNPs | 0.7285 | 0.6550 | 0.6140 | 0.4710 |
| SNP Subset I | 0.7179 | 0.6543 | 0.6223 | 0.4675 |
| SNP Subset II | 0.7252 | 0.6467 | 0.6075 | 0.4704 |
| SNP Subset III | 0.6302 | 0.6199 | 0.4693 | 0.3577 |

SNP subsets I, II, and III comprise 193, 124, and 130 significantly associated SNPs identified by four GWAS methods in panels I, II, and III, respectively. The corresponding predictabilities obtained from 0.3K and all SNPs were listed as references. PH, plant height; EH, ear height; EW, ear weight; GY, grain yield.

GBTS strategy [15]. The array-based platform is also more stable than GBS, which randomly captures genomic sequences, allowing more convenient data sharing across research institutions and breeding companies. By taking advantage of shared genotypic datasets, researchers could improve their own GS models and predict candidates more efficiently and precisely.

Marker density directly affects the cost of genotyping. The 1K marker subset yielded predictabilities comparable to those of 10K and all SNPs irrespective of genotyping platform. The predictabilities of PH, EH, EW, and GY quickly reached a plateau as the number of markers increased. Similar results have been reported for several traits in maize. For common rust resistance, about 300 SNPs were sufficient for good predictability in a DH population [33]. For maize kernel zinc concentration, prediction accuracies reached a plateau at 500 markers with both GBS and rAmpSeq datasets [14]. For anthesis date and plant height, only 200 SNPs were found to be enough to attain good prediction in biparental maize populations [34]. These studies suggest that low-density marker panels are feasible and cost-effective for GS. The results also showed that the 1K SNP subset from GBS and TSC yielded significantly higher predictability than the 1K SNP

array, indicating that there is still room for optimization of the SNP array.

Selecting the optimum SNP set is crucial for developing SNP arrays and reducing the genotyping cost. The use of selected markers or significant markers identified by GWAS in GS has been reported as an effective strategy to improve prediction performance for yield-related traits in several crop species [35–37]. Yuan et al. [35] proposed that the prediction accuracy obtained from trait-associated SNPs was higher than those obtained from genome-wide markers for grain yield in maize. Liu et al. [36] reported that trait-relevant markers identified from the training population gave higher prediction accuracy for six agronomic traits than randomly selected markers did. Ali et al. [37] found that GWAS-derived markers improved prediction accuracy for yield-related traits in winter wheat. However, Xu et al. [38] showed that significant SNPs detected by GWAS were unable to improve GP with all markers. Kristensen et al. [39] also found that GP based on all 10,802 SNPs was superior to prediction based on a few associated SNPs for most traits in wheat. A possible reason for these contradictory results may be differing levels of genetic relatedness between training and validation populations. We evaluated the predictability of 305 hybrids using three associated SNP subsets selected from GWAS in different panels. The results showed that 193 SNPs identified from panel I and 124 SNPs identified from panel II yielded predictabilities comparable to or even higher than those from all 102,654 SNPs, but 130 SNPs identified from panel III produced much lower predictabilities than all SNPs, even lower than 0.3K randomly selected SNPs. The heat map illustrated that the level of genetic relatedness between panel II and panel III was low (Fig. S2). We further mined candidate genes in the 50-kb regions flanking the detected SNPs. Respectively 401, 265 and 270 candidate genes were detected in panels I, II, and III, but only eight overlapping candidate genes were detected by panel II and panel III simultaneously (Fig. S3; Table S3). Similar results have been reported by Liu et al. [40], who detected almost no overlapping SNPs or genes underlying ear rot disease across different maize populations with weak genetic relatedness. These results indicate that identification of SNPs associated with complex traits is limited by GWAS populations. We conclude that GWAS is conducive to selecting effective SNP subsets for GP, but that a large panel that contains some individuals closely related to those to be predicted is needed.

We suggest that incorporating a few functional markers detected via GWAS in GBS platforms into the current SNP array is apt to be an effective strategy for reducing genotyping cost as well as achieving desirable prediction performance. The GBTS strategy provides the opportunity to flexibly integrate novel markers in array without resynthesizing. In this study, we focused on a few agronomic traits of a specific population. More collaborative research by multiple institutions might identify the optimum SNP subset for developing cost-effective SNP arrays, thus making GS technology affordable and feasible in breeding pipelines.

## CRediT authorship contribution statement

**Guangning Yu:** Investigation, Formal analysis, Writing – original draft. **Yanru Cui:** Methodology. **Yuxin Jiao:** Data curation, Visualization. **Kai Zhou:** Investigation. **Xin Wang:** Methodology. **Wenyan Yang:** Formal analysis. **Yiyi Xu:** Visualization. **Kun Yang:** Data curation. **Xuecai Zhang:** Validation. **Pengcheng Li:** Validation. **Zefeng Yang:** Supervision, Project administration. **Yang Xu:** Conceptualization, Writing – review & editing, Funding acquisition. **Chenwu Xu:** Conceptualization, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (32061143030, 32170636, 32100448), the Key Research and Development Program of Jiangsu Province (BE2022343), the Seed Industry Revitalization Project of Jiangsu Province (JBGS[2021]009), Project of Hainan Yazhou Bay Seed Lab (B21HJ0223), the State Key Laboratory of North China Crop Improvement and Regulation (NCCIR2021KF-5, NCCIR2021ZZ-4), Jiangsu Province Agricultural Science and Technology Independent Innovation (CX(21)1003), the Independent Scientific Research Project of the Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding (PLR202102), the Open Funds of the Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding (PL202005), Yangzhou University High-end Talent Support Program, and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

## Appendix A. Supplementary data

Supplementary data for this article can be found online at https://doi.org/10.1016/j.cj.2022.09.004.

## References

[1] 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.
[2] Y. Xu, Y. Zhao, X. Wang, Y. Ma, P. Li, Z. Yang, X. Zhang, C. Xu, S. Xu, Incorporation of parental phenotypic data into multi-omic models improves prediction of yield-related traits in hybrid rice, Plant Biotechnol. J. 19 (2021) 261–272.
[3] Y. Xu, X. Liu, J. Fu, H. Wang, J. Wang, C. Huang, B.M. Prasanna, M.S. Olsen, G. Wang, A. Zhang, Enhancing genetic gain through genomic selection: from livestock to plants, Plant Commun. 1 (2020) 100005.
[4] B.R. Rice, A.E. Lipka, Diversifying maize genomic selection models, Mol. Breed. 41 (2021) 33.
[5] Y. Xiao, S. Jiang, Q. Cheng, X. Wang, J. Yan, R. Zhang, F. Qiao, C. Ma, J. Luo, W. Li, H. Liu, W. Yang, W. Song, Y. Meng, M.L. Warburton, J. Zhao, X. Wang, J. Yan, The genetic mechanism of heterosis utilization in maize improvement, Genome Biol. 22 (2021) 148.
[6] E.J. Millet, W. Kruijer, A. Coupel-Ledru, S. Alvarez Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk, F. Tardieu, Genomic prediction of maize yield across European environmental conditions, Nat. Genet. 51 (2019) 952–956.
[7] Z. Guo, Q. Yang, F. Huang, H. Zheng, Z. Sang, Y. Xu, C. Zhang, K. Wu, J. Tao, B.M. Prasanna, M.S. Olsen, Y. Wang, J. Zhang, Y. Xu, Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip, Plant Commun. 2 (2021) 100230.
[8] Y.S. Chung, S.C. Choi, T.H. Jun, C. Kim, Genotyping-by-sequencing: a promising tool for plant genetics research and breeding, Hortic Environ. Biote. 58 (2017) 425–431.
[9] A. Rasheed, Y. Hao, X. Xia, A. Khan, Y. Xu, R.K. Varshney, Z. He, Crop breeding chips and genotyping platforms: progress, challenges, and perspectives, Mol. Plant 10 (2017) 1047–1064.
[10] Y. Wu, F. San Vicente, K. Huang, T. Dhliwayo, D.E. Costich, K. Semagn, N. Sudha, M. Olsen, B.M. Prasanna, X. Zhang, R. Babu, Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs, Theor. Appl. Genet. 129 (2016) 753–765.
[11] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS ONE 6 (2011) e19379.
[12] J. Crossa, Y. Beyene, S. Kassa, P. Perez, J.M. Hickey, C. Chen, G. de los Campos, J. Burgueno, V.S. Windhausen, E. Buckler, J.L. Jannink, M.A. Lopez Cruz, R. Babu, Genomic prediction in maize breeding populations with genotyping-by-sequencing, G3-Genes Genomes Genet. 3 (2013) 1903–1926.

[13] N. Wang, Y. Yuan, H. Wang, D. Yu, Y. Liu, A. Zhang, M. Gowda, S.K. Nair, Z. Hao, Y. Lu, F. San Vicente, B.M. Prasanna, X. Li, X. Zhang, Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding, Sci. Rep. 10 (2020) 16308.
[14] R. Guo, T. Dhliwayo, E.K. Mageto, N. Palacios-Rojas, M. Lee, D. Yu, Y. Ruan, A. Zhang, F. San Vicente, M. Olsen, J. Crossa, B.M. Prasanna, L. Zhang, X. Zhang, Genomic prediction of kernel zinc concentration in multiple maize populations using genotyping-by-sequencing and repeat amplification sequencing markers, Front. Plant Sci. 11 (2020) 534.
[15] Z. Guo, H. Wang, J. Tao, Y. Ren, C. Xu, K. Wu, C. Zou, J. Zhang, Y. Xu, Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize, Mol. Breed. 39 (2019) 37.
[16] M.W. Ganal, G. Durstewitz, A. Polley, A. Bérard, E.S. Buckler, A. Charcosset, J.D. Clarke, E.M. Graner, M. Hansen, J. Joets, M.C. Le Paslier, M.D. McMullen, P. Montalent, M. Rose, C.C. Schön, Q. Sun, H. Walter, O.C. Martin, M. Falque, A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome, PLoS ONE 6 (2011) e28334.
[17] S. Unterseer, E. Bauer, G. Haberer, M. Seidel, C. Knaak, M. Ouzunova, T. Meitinger, T.M. Strom, R. Fries, H. Pausch, C. Bertani, A. Davassi, K.F.X. Mayer, C.C. Schön, A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array, BMC Genomics 15 (2014) 823.
[18] H. Tian, Y. Yang, H. Yi, L. Xu, H. He, Y. Fan, L. Wang, J. Ge, Y. Liu, F. Wang, J. Zhao, New resources for genetic studies in maize (*Zea mays* L.): a genome-wide Maize6H-60K single nucleotide polymorphism array and its application, Plant J.105 (2021) 1113–1122.
[19] Y.H. Liu, Y. Xu, M. Zhang, Y. Cui, S.H. Sze, C.W. Smith, S. Xu, H.B. Zhang, Accurate prediction of a quantitative trait using the genes controlling the trait for gene-based breeding in cotton, Front. Plant Sci. 11 (2020) 583277.
[20] M. Zhang, Y. Cui, Y.H. Liu, W. Xu, S.H. Sze, S.C. Murray, S. Xu, H.B. Zhang, Accurate prediction of maize grain yield using its contributing genes for gene-based breeding, Genomics 112 (2020) 225–236.
[21] X. Wang, Z. Zhang, Y. Xu, P. Li, X. Zhang, C. Xu, Using genomic data to improve the estimation of general combining ability based on sparse partial diallel cross designs in maize, Crop J. 8 (2020) 819–829.
[22] P. Li, J. Wei, H. Wang, Y. Fang, S. Yin, Y. Xu, J. Liu, Z. Yang, C. Xu, Natural variation and domestication selection of *ZmPGP1* affects plant architecture and yield-related traits in maize, Genes-Basel 10 (2019) 664.
[23] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, J. Stat. Softw. 67 (2015) 1–48.
[24] G.C. Allen, M.A. Flores-Vergara, S. Krasynanski, S. Kumar, W.F. Thompson, A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide, Nat. Protoc. 1 (2006) 2320–2325.
[25] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, E.S. Buckler, TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, PLoS ONE 9 (2014) e90346.
[26] L. Yin, H. Zhang, Z. Tang, J. Xu, D. Yin, Z. Zhang, X. Yuan, M. Zhu, S. Zhao, X. Li, X. Liu, rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study, Genom. Proteom. Bioinf. 19 (2021) 619–628.
[27] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.
[28] B.H. Mevik, R. Wehrens, The pls package: principal component and partial least squares regression in R, J. Stat. Softw. 18 (2007) 1–23.
[29] P. Pérez, G. de los Campos, Genome-wide regression and prediction with the BGLR statistical package, Genetics 198 (2014) 483–495.
[30] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, Association for Computing Machinery: San Francisco, CA, USA. pp. 785–794.
[31] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (2002) 18–22.
[32] Y.W. Zhang, C.L. Tamba, Y.J. Wen, P. Li, W.L. Ren, Y.L. Ni, J. Gao, Y.M. Zhang, mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies, Genom. Proteom. Bioinf. 18 (2020) 481–487.
[33] J. Ren, Z. Li, P. Wu, A. Zhang, Y. Liu, G. Hu, S. Cao, J. Qu, T. Dhliwayo, H. Zheng, M. Olsen, B.M. Prasanna, F. San Vicente, X. Zhang, Genetic dissection of quantitative resistance to common rust (*Puccinia sorghi*) in tropical maize (*Zea mays* L.) by combined genome-wide association study, linkage mapping, and genomic prediction, Front. Plant Sci. 12 (2021) 692205.
[34] 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.
[35] Y. Yuan, J.E. Cairns, R. Babu, M. Gowda, D. Makumbi, C. Magorokosho, A. Zhang, Y. Liu, N. Wang, Z. Hao, F. San Vicente, M.S. Olsen, B.M. Prasanna, Y. Lu, X. Zhang, Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize, Front. Plant Sci. 9 (2019) 1919.
[36] X. Liu, H. Wang, X. Hu, K. Li, Z. Liu, Y. Wu, C. Huang, Improving genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize, Front. Plant Sci. 10 (2019) 1129.
[37] M. Ali, Y. Zhang, A. Rasheed, J.K. Wang, L.Y. Zhang, Genomic prediction for grain yield and yield-related traits in chinese winter wheat, Int. J. Mol. Sci. 21 (2020) 1342.
[38] Y. Xu, C. Xu, S. Xu, Prediction and association mapping of agronomic traits in maize using multiple omic data, Heredity 119 (2017) 174–184.
[39] P.S. Kristensen, A. Jahoor, J.R. Andersen, F. Cericola, J. Orabi, L.L. Janss, J. Jensen, Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines, Front. Plant Sci. 9 (2018) 69.
[40] Y. Liu, G. Hu, A.O. Zhang, A. Loladze, Y. Hu, H. Wang, J. Qu, X. Zhang, M. Olsen, F. San Vicente, J. Crossa, F. Lin, B.M. Prasanna, Genome-wide association study and genomic prediction of Fusarium ear rot resistance in tropical maize germplasm, Crop J. 9 (2021) 325–341.